# AN INVESTIGATION OF HUMAN PROTEIN INTERACTIONS USING THE COMPARATIVE METHOD 

Saif Ur-Rehman

A Thesis Submitted for the Degree of PhD at the University of St Andrews


2012

Full metadata for this item is available in Research@StAndrews:FulIText
at:
http://research-repository.st-andrews.ac.uk/

Please use this identifier to cite or link to this item: http://hdl.handle.net/10023/3119

This item is protected by original copyright

This item is licensed under a Creative Commons Licence


# School of Biology 

PhD Thesis
An investigation of human protein interactions using the comparative method
by

Saif Ur-Rehman

20th Jan 2012

## 1. Candidate's declarations:

I, Saif Ur-Rehman hereby certify that this thesis, which is approximately 52,000 words in length, has been written by me, that it is the record of work carried out by me and that it has not been submitted in any previous application for a higher degree.

I was admitted as a research student in [09, 2006] and as a candidate for the degree of PhD in [09, 2007]; the higher study for which this is a record was carried out in the University of St Andrews between [2006] and [2010].

Date signature of candidate $\qquad$

## 2. Supervisor's declaration:

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of PhD in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree.

Date $\qquad$ signature of supervisor $\qquad$
3. Permission for electronic publication: (to be signed by both candidate and supervisor)

In submitting this thesis to the University of St Andrews I understand that I am giving permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work not being affected thereby. I also understand that the title and the abstract will be published, and that a copy of the work may be made and supplied to any bona fide library or research worker, that my thesis will be electronically accessible for personal or research use unless exempt by award of an embargo as requested below, and that the library has the right to migrate my thesis into new electronic forms as required to ensure continued access to the thesis. I have obtained any third-party copyright permissions that may be required in order to allow such access and migration, or have requested the appropriate embargo below.

The following is an agreed request by candidate and supervisor regarding the electronic publication of this thesis:
Add one of the following options:
(ii) Access to [all or part] of printed copy but embargo of [all or part] of electronic publication of thesis for a period of 2 years (maximum five) on the following ground(s):
publication would preclude future publication;

Date $\qquad$ signature of candidate $\qquad$ signature of supervisor

A supporting statement for a request for an embargo must be included with the submission of the draft copy of the thesis. Where part of a thesis is to be embargoed, please specify the part and the reasons.


#### Abstract

There is currently a large increase in the speed of production of DNA sequence data as next generation sequencing technologies become more widespread. As such there is a need for rapid computational techniques to functionally annotate data as it is generated. One computational method for the functional annotation of protein-coding genes is via detection of interaction partners. If the putative partner has a functional annotation then this annotation can be extended to the initial protein via the established principle of "guilt by association".

This work presents a method for rapid detection of functional interaction partners for proteins through the use of the comparative method. Functional links are sought between proteins through analysis of their patterns of presence and absence amongst a set of 54 eukaryotic organisms. These links can be either direct or indirect protein interactions. These patterns are analysed in the context of a phylogenetic tree.

The method used is a heuristic combination of an established accurate methodology involving comparison of models of evolution the parameters of which are estimated using maximum likelihood, with a novel technique involving the reconstruction of ancestral states using Dollo parsimony and analysis of these reconstructions through the use of logistic regression. The methodology achieves comparable specificity to the use of gene coexpression as a means to predict functional linkage between proteins.

The application of this method permitted a genome-wide analysis of the human genome, which would have otherwise demanded a potentially prohibitive amount of computational resource.

Proteins within the human genome were clustered into orthologous groups. 10 of these proteins, which were ubiquitous across all 54 eukaryotes, were used to reconstruct a phylogeny. An application of the heuristic predicted a set of functional protein interactions in human cells. 1,142 functional interactions were predicted. Of these predictions 1,131 were not present in current protein-protein interaction databases.


## Acknowledgements

I thank the BBSRC for funding my work and making this project possible. I would also like to thank my supervisor, Dr Daniel Barker for his tireless support and good advice. I would like to thank Professor Mike Ritchie, Dr Jeff Graves and Dr Anne Smith for acting as my committee. I would also like to thank Dr Rona Ramsay, Ji-Hiyun Lim, Wim Verleyen, Dr Christoph Echtermeyer and Maria Keays for helpful discussion. I would also like to thank Dr Neil Symington and Dr Herbert Fruchtl for their helpful advice in using cluster-computing resources.

On a personal note I would like to thank my wife Cathryn for her constant unconditional support without which I could not have brought this project to fruition. I would also like to thank my friends Ken Armstrong and Jack Levell who played a large part in keeping me balanced over the course of my work. Finally I would like to thank my parents for their love and support, which has been a source of strength.
Abstract ..... 2
Acknowledgements ..... 3
Chapter 1 ..... 8
1.1 History ..... 8
$1.2 \mathrm{DNA} / \mathrm{RNA}$ ..... 9
1.2.1 RNA ..... 10
1.3 Proteins ..... 11
1.3.1 Protein secondary structure ..... 11
1.3.2 Protein tertiary structure ..... 12
1.3.3 Protein quaternary structure ..... 12
1.3.4 Protein domains ..... 12
1.3.5 Protein motifs ..... 12
1.4 Genes ..... 13
1.4.1 Structure of a gene ..... 13
1.4.1.1 Regulatory region of a gene ..... 14
1.4.2 Transcription ..... 15
1.4.2.1 Post Transcriptional processing ..... 15
1.4.2.1.1 Genetic Code ..... 15
1.4.2.1.2 Open reading frames ..... 16
1.4.2.1.3 Exons/Introns ..... 17
1.4.2.2 Post Transcriptional processing (cont) ..... 18
1.4.2.2.1 RNA splicing ..... 18
1.4.2.2.2 Capping ..... 19
1.4.2.2.3 Polyadenylation ..... 20
1.4.3 Translation ..... 20
1.5 Genomics ..... 22
1.5.1 Genome annotation ..... 22
1.5.2 Genome and cDNA assembly ..... 23
1.5.3 Gene detection ..... 23
1.5.4 Functional annotation of genes ..... 25
1.5.4.1 Laboratory based techniques ..... 25
1.5.4.2 Computational methods for functional annotation of genes ..... 26
1.5.4.2.1 Alignment based methods ..... 26
1.5.4.2.2 Genome context methods ..... 28
1.5.4.2.2.1 Rosetta stone ..... 29
1.5.4.2.2.2 Gene neighbour ..... 29
1.5.4.2.2.3 Interolog detection. ..... 29
1.5.4.2.2.4 Phylogenetic profiling ..... 29
1.5.4.2.2.5 Comparative methods ..... 31
1.5.4.2.2.6 Mirror trees ..... 32
1.5.5 Storage of functional information ..... 33
1.5.5.1 GO ..... 33
1.5.5.1 KEGG ..... 33
1.6 Transcriptomics ..... 34
1.6.1 Microarrays ..... 34
1.6.2 Other methods for transcriptome examination ..... 34
1.7 Proteomics ..... 34
1.7.1 Protein Structure ..... 35
1.7.2 Protein interactions ..... 35
1.7.2.1 Experimental detection of protein interactions ..... 36
1.8 Description of project ..... 37
Chapter 2 ..... 39
2.1 Introduction ..... 39
2.1.1 Homology ..... 39
2.1.2 Molecular evolution ..... 40
2.1.2.1 Synonymous and non-synonymous mutations ..... 41
2.1.3 Phylogenetic trees ..... 41
2.1.3.1 Species trees and gene trees ..... 42
2.1.3.2 Topologies and branch lengths ..... 44
2.1.3.3 Bootstrap support values ..... 46
2.1.3.4 Evolutionary models in tree estimation ..... 46
2.1.4 Detection of homology in molecular data ..... 47
2.1.5 Multiple sequence alignment ..... 49
2.1.5.1 Multiple sequence alignment quality filtration ..... 50
2.1.6 Methods to estimate phylogenetic trees ..... 50
2.1.6.1 Distance methods ..... 51
2.1.6.2 Discrete character state methods ..... 53
2.1.6.2.1 Maximum Parsimony ..... 54
2.1.6.2.2 Maximum likelihood ..... 55
2.1.6.2.3 Bayesian Methods ..... 56
2.1.6.3 Heuristic search methods ..... 57
2.1.8 Model selection in phylogenetic tree estimation ..... 60
2.1.10 Phylogenetic analysis using gene presence ..... 63
2.2 Methods ..... 64
2.2.1 Data Selection ..... 65
2.2.2 Data Acquisition ..... 65
2.2.3 Pairwise Alignment ..... 69
2.2.4 Orthology Determination ..... 69
2.2.4.1 Inparanoid ..... 69
2.2.4.2 Implementation ..... 70
2.2.4.2.1 Design ..... 71
2.2.4.3 Application ..... 72
2.2.5 Phylogenetic profiles ..... 72
2.2.5.1 Single copy proteins. ..... 78
2.2.5.2 Proteome content data/tree ..... 79
2.2.6 Multiple alignment ..... 79
2.2.7 Model selection ..... 80
2.2.8 Phylogeny reconstruction ..... 81
2.2.9 Comparison of protein content tree with super matrix tree ..... 82
2.3 Results ..... 82
2.4 Discussion ..... 93
2.4.1.ML tree ..... 93
2.4.2 Proteome content phylogeny ..... 94
2.4.3 Conclusion ..... 95
Chapter 3 ..... 96
3.1 Introduction ..... 96
3.1.1 Hamming distance ..... 96
3.1.2 Comparative method ..... 97
3.1.2.1 Phylogenetic profile analysis using the comparative method. ..... 97
3.1.3 Co-expression as measured by microarray ..... 101
3.1.4 Bayesian classifier ..... 101
3.2 Methods ..... 103
3.2.1 Assessing quality ..... 103
3.2.2 Training and test data. ..... 104
3.2.3 Hamming distance ..... 105
3.2.4 Constrained ML ..... 105
3.2.5 Co-expression of mRNA ..... 106
3.2.6 Bayesian classifier ..... 108
3.3 Results ..... 109
3.3.1 Hamming distance ..... 109
3.3.2 Constrained ML ..... 110
3.3.2.1 Likelihood ratio statistic ..... 115
3.3.3 Co-expression of mRNA ..... 121
3.3.4 Bayesian classifier ..... 124
3.3.5 Method Comparison ..... 127
3.4 Discussion ..... 128
3.4.1 Low Sensitivities ..... 129
Chapter 4 ..... 130
4.1. Introduction ..... 130
4.1.1 Ancestral state reconstruction ..... 132
4.1.1.1 Parsimony ..... 132
4.1.1.2 Likelihood ..... 133
4.2 Filters ..... 134
4.2.1 Hamming distance filter ..... 134
4.2.2 Ancestral state reconstruction filter ..... 134
4.2.2.1 Dollo parsimony ..... 134
4.2.2.2 Maddison Test for correlated evolution ..... 136
4.3 Methods ..... 138
4.3.1 Maddison test for correlated evolution ..... 141
4.3.1.1 Algorithm ..... 141
4.3.1.2 Calculation of total number of ways of having $x$ gains and $y$ losses over the tree. ..... 142
4.3.1.3 Calculation of total number of ways of having $p$ gains and $q$ losses in subset $k$ given $x$ gains and$y$ losses over the entire tree143
4.3.1.4 Permutation effects ..... 144
4.3.1.5 Evaluation of Maddison test as heuristic for constrained ML ..... 148
4.3.2 Modification of test to match Dollo constraints ..... 148
4.3.3 Differential parsimony ..... 150
4.3.4 Dollo-pos/ Dollo-overall ..... 150
4.3.5 Test based on logistic regression ..... 150
4.3.5.1 Evaluation of logistic regression as a heuristic for constrained ML ..... 156
4.4 Results ..... 157
4.4.1 Maddison test for correlated evolution ..... 158
4.4.2 Differential parsimony ..... 160
4.4.3 Dollo-pos ..... 161
4.4.4 Dollo-overall. ..... 162
4.4.5 Logistic regression ..... 162
4.4.6 Hamming distance ..... 168
4.5 Discussion ..... 168
Chapter 5 ..... 170
5.1 Introduction ..... 170
5.1.1 PPI databases ..... 170
5.1.1.1 MIPS ..... 170
5.1.1.2 BIND ..... 171
5.1.1.3 MINT ..... 171
5.1.1.4 INTACT ..... 171
5.1.1.5 HPRD ..... 171
5.1.1.6 DIP ..... 171
5.1.1.7 REACTOME ..... 171
5.1.1.8 STRING ..... 172
5.1.1.9 I2D ..... 172
5.1.1.10 KEGG ..... 172
5.1.1.11 BIOGRID ..... 172
5.1.1.12 Discussion ..... 172
5.1.2 Power law ..... 172
5.2 Methods ..... 173
5.2.1 Short Branch filtration ..... 174
5.2.2 GO term enrichment ..... 178
5.2.3 Intersection with other data sources ..... 181
5.3 Results ..... 182
5.3.1 GO Enrichment ..... 182
5.3.2 Intersection with known data ..... 187
5.3.3 Network statistics ..... 188
5.4 Discussion ..... 189
5.4.1 GO enrichment ..... 189
5.4.2 Intersection with known data ..... 190
5.4.3 Weaknesses. ..... 190
5.4.3.1 Scaling. ..... 191
5.4.4 Conclusions ..... 193
Chapter 6 ..... 194
6.1 Summary of Project ..... 194
6.1.1 Repeat Analysis ..... 199
6.2 Conclusion ..... 200
6.3 Future directions ..... 202
6.3.1 Computational extensions ..... 202
6.3.2 Consensus profiles ..... 202
6.3.3 Correlated evolution of proteins with the presence or absence of phenotypes ..... 203
6.3.4 Drug Targets ..... 204
References ..... 205
Appendix A Description of divergence of Java implementation of Inparanoid algorithm from Perl implementation. ..... 223
Appendix B Individual Gene trees for genes in super matrix utilised in construction of Phylogeny ..... 232
Appendix C: Predictions made by constrained ML ..... 242
Appendix D Concatenated Filtered Alignment ..... 316

## Chapter 1

## Introduction to computational annotation of protein coding genes

### 1.1 History

The discovery in the 1940s (Avery et al. 1944) and confirmation in the 1950s (Hershey and Chase 1952) of DNA (deoxyribonucleic acid) as the physical basis for inheritance was a milestone in biological research. It provided for a means to examine the materials and processes underlying phenotypic traits and provided a conceptual link to the other natural sciences. This was rapidly followed by the elucidation of the three dimensional structure of B-DNA (Watson and Crick 1953) which is the form of DNA prevalent in living cells as it is conducive to nucleosome formation (Richmond and Davey 2003). This structure was the now famous double helix. It had been previously established (Beadle and Tatum 1941) that genes exist as discrete regions within the genome whose sequence codes for the sequence of a corresponding chain of amino acids. The genome of an organism is the full set of hereditary material it possesses (Alberts 2010). This is RNA in the case of some viruses and DNA in the case of all other types of cellular organism (Brown 2006). The discovery of the genetic code (Crick et al. 1961) provided information on the mechanism for this production which operates via initial intermediary transcription into RNA (ribonucleic acid) and then translation into proteins. (Some genes also code for RNA products such as tRNAs and other non-coding RNAs (Brown 2006)).

The first feasible method for determining the sequence of DNA was the MaxamGilbert chemical degradation method (Maxam and Gilbert 1977). This method was however supplanted by the near simultaneous invention of the chain termination reaction method by Frederick Sanger (Sanger et al. 1977) of DNA sequencing which was both safer and more efficient (Brown 2006). This led to the first full genome of an organism to be sequenced, which was bacteriophage fX174 (Sanger et al. 1978). Another contribution by Sanger was that of shotgun sequencing. This entails the shattering of a piece of DNA into random fragments and the sequencing of those fragments. The sequences of the fragments are then assembled through searching for overlaps between them. This method facilitated the sequencing of number of relatively larger viral and prokaryotic genomes such as Bacteriophage MS2 (Fiers et al. 1976).

In 1996 Saccharomyces cerevisiae was the first eukaryotic genome to be sequenced (Goffeau et al. 1996) via a large collaborative effort. This was followed by the publication of the first multi cellular eukaryotic genome Caenorhabditis elegans in 1998 (C. elegans Sequencing Consortium 1998) and the draft genomes of the vertebrate Homo sapiens soon followed in 2001 (Venter et al. 2001). The application of industrial streamlining and automation to sequencing efforts over the last 20 years as well as more recently with the onset of next generation sequencing technologies there has been almost exponential growth to sequence databases such as NCBI GenBank (Benson et al. 2009). Sequence data without further processing and annotation cannot shed any light on either biological function or evolutionary relationships between organisms. This means that there has been a focus on the development of highly accurate high throughput methods for functional annotation of genes and other functional genomic elements in recent years as the parity between rates of data generation and rates of accurate and verifiable annotation becomes more divergent (Zhu et al. 2007).

### 1.2 DNA/RNA

DNA itself is made up of a linear backbone of alternating deoxyribose sugar and phosphate residues (Strachan and Read 2004). There is a nitrogenous base attached to the 1' (one prime) carbon of each individual sugar residue. There are two forms of nitrogenous base present within DNA. One form possesses a single interlocked heterocyclic ring of carbon and nitrogen atoms. Bases that exist in this conformation are known as pyrimidines (Strachan and Read 2004). The second form of base consists of two interlocked heterocyclic rings of carbon and nitrogen atoms. These bases are known as purines (Strachan and Read 2004). There are two pyrimidines represented within DNA (Strachan and Read 2004). These are cytosine and thymine commonly represented by the abbreviations C and T respectively (Brown 2006). There are also two purines present, adenine and guanine represented as A and G (Brown 2006). The stability of the double helix structure of DNA is maintained through hydrogen bond formation between the pyrimidine-purine pair C and G and hydrogen bond formation between the remaining pyrimidine-purine pair T and A as well as base stacking interactions between adjacent bases (Yakovchuk et al. 2006). Due to structural constraints base pairing can only occur between a pyrimidine and a purine (Brown 2006).

The linear backbone of DNA/RNA is maintained by a phosphodiester bond formed between the 3 ' ( 3 prime) carbon atom of the sugar and the $5^{\prime}$ ( 5 prime) carbon of the succeeding sugar (Strachan and Read 2004). The backbone is terminated by a sugar where the $5^{\prime}$ carbon is not linked to a succeeding sugar residue. This point is known as the $5^{\prime}$ ' end.

Similarly the other end of the molecule lacks a phosphodiester bond on the 3' carbon and is known as the 3 ' end (Strachan and Read 2004). The sequence of DNA is usually described in the $5^{\prime} \rightarrow 3^{\prime}$ direction, as this is the direction of DNA replication as well as transcription of RNA using DNA as a template (Strachan and Read 2004). Thus a feature along a DNA molecule is referred to as being upstream of another feature if it is closer to the $5^{\prime}$ end. The length of a DNA molecule is measured in units of individual base pairs (bp).

DNA is a biopolymer and as such can be fully represented by the sequence of its constituent nucleotide bases. Determination of this sequence for a complete organism effectively represents the DNA blueprints for the construction of that organism, i.e. the amino acid sequences of its constituent proteins and RNA molecules, as well as the regulatory sequences that regulate production of these molecules both spatially and temporally.

### 1.2.1 RNA

RNA is constructed of similar residues, however the sugar is a ribose as opposed to deoxyribose and the pyrimidine base thymine is replaced with the base uracil commonly represented by the abbreviation U (Strachan and Read 2004). There is a diverse population of RNA molecules produced by the eukaryotic genome. These molecules are involved with a number of processes essential to life, including protein synthesis and regulation of gene expression. A breakdown of general RNA types and their functions is presented in Table 1.1.

| Abbreviated Name Full name |  | Primary Function |
| :--- | :---: | :---: |
| mRNA Messenger RNA Provides a template for protein <br> synthesis. <br> tRNA Transfer RNA Connection of mRNA to relevant <br> amino acid during protein synthesis. <br> rRNA Ribosomal RNA Component of protein synthesising <br> organelles known as ribosomes. <br> snRNA Small nuclear RNA Component of RNA-protein machine <br> (involved in post transcriptional <br> modification of mRNA) known as the <br> spliceosome. <br> snoRNA Small nucleolar RNA Involved in the modification of rRNA <br> and snRNA <br> miRNA Micro RNA Involved in the regulation of RNA <br> stability and translation. <br> siRNA Short interfering RNA Involved in the targeted degradation of <br> RNA. |  |  |

Table 1.1: General types of RNA molecules with function (Blow 2004).

### 1.3 Proteins

Protein molecules are polymers comprised of one or more chains of amino acids. A chain of amino acids can also be referred to as a polypeptide chain. Amino acids are molecules that consist of an amino group, a carboxylic group, an R group and a hydrogen atom (Berg et al. 2001). These components are all linked to a central carbon atom known as the $\alpha$ carbon (Berg et al. 2001). A polypeptide chain is formed when a peptide bond is formed between the amino group of one amino acid and the carboxyl group of another. All polypeptide chains have a free amino group at one end and a free carboxyl group at the other. These are known as the N -terminus and C-terminus respectively (Alberts 2002). The sequence of a polypeptide chain is presented as moving from the N -terminus to the C -terminus (Alberts 2002). A linear polypeptide chain is also considered the primary structure of a protein (Brown 2006).

It is the R group that distinguishes amino acids (Berg et al. 2001). R groups vary in factors such as "size, shape, charge, hydrogen-bonding capacity, hydrophobic character, and chemical reactivity" (Berg et al. 2001). There are 20 naturally occurring amino acids that are typically utilised by living cells (Alberts 2002).

### 1.3.1 Protein secondary structure

The interactions of the R, carboxyl, and amine groups of individual amino acids in a polypeptide chain with each other cause polypeptide chains to fold into characteristic conformations. These conformations are known as the secondary structure of a protein. There are two main types of secondary structure (Brown 2006).

- The $\alpha$ helix: This is a structure formed by interactions between the carboxyl groups and amine groups of amino acids which are separated by a number intermediate amino acids (Berg et al. 2001).
- The $\beta$ sheet: This is a structure formed by the interactions between two polypeptide chains running either parallel or anti parallel to each other (Brown 2006).
- Random coils: In the absence of particular structural imperatives polypeptide chains can take on any number of shapes that are sterically possible. These shapes are referred to as random coils (Shortle and Ackerman 2001).


### 1.3.2 Protein tertiary structure

The tertiary structure of a protein is formed by the folding up of the secondary structural constructs formed by the polypeptide chain into a three dimensional configuration (Brown 2006). This configuration is held together a number of chemical forces including hydrogen bonding between individual amino acids and the interactions of hydrophobic amino acids with water (Brown 2006).

### 1.3.3 Protein quaternary structure

The quaternary structure of a protein is formed by the interactions of multiple polypeptide chains. Quaternary structure is a hallmark of proteins with a complex function (Brown 2006).

### 1.3.4 Protein domains

A protein domain can be defined as "a substructure produced by any part of a polypeptide chain that can fold independently into a compact, stable structure" (Alberts 2002). There are a number of recurrent protein domains that are functionally important within the eukaryotic cell. These include:

- Helix turn helix: This is a domain comprised of two $\alpha$ helices separated by a short strand of amino acids. It is functionally important due to its ability to bind DNA (Brennan and Matthews 1989).
- Transmembrane domain: This is a domain consisting of $\alpha$ helical structures capable of passing through the lipid bilayer (cell membrane) that surrounds the cell. These are crucially important in facilitating cell-cell communication and relaying information about the external environment into a cell (Brown 2006).


### 1.3.5 Protein motifs

Protein motifs are conceptually similar to protein domains in that they are distinct substructures within a protein molecule (Brown 2006). In contrast with domains they are not able to form outside of the context of the overall protein. Functionally important protein motifs include:

- Leucine zipper: This motif is important in that it facilitates the formation of protein quartenary structure by the dimerisation of two leucine rich regions of separate polypeptides (Brown 2006). It is a motif that is found in a number of proteins that bind DNA (Brown 2006).
- Zinc finger: The zinc finger motif is a set of polypeptide chains whose interactions is stabilised by the presence of zinc ions. It is also present in DNA binding proteins (Brown 2006).


### 1.4 Genes

As mentioned above the blueprints for the production of given protein and RNA molecules within an organism are contained in subsections of its genome known as genes. A current more specific definition of a gene presented by Pesole (Pesole 2008) defines them as " $a$ discrete genomic region whose transcription is regulated by one or more promoters and distal regulatory elements and which contains the information for the synthesis of functional proteins or non-coding RNAs, related by the sharing of a portion of genetic information at the level of the ultimate products (proteins or RNAs)".

### 1.4.1 Structure of a gene

As implied by that definition a gene is made up of two distinct parts. These are firstly a transcribed area, which is the portion of DNA that is actually converted into RNA and secondly regulatory regions, which can occur either upstream or down stream of the transcribed region. Regulatory regions within the vicinity of a gene provide recognition signals for proteins known as transcription factors. These proteins regulate the transcription rate of a gene by either carrying out the actual transcription, or by binding to DNA and either promoting or silencing transcription (Maston et al. 2006). As the binding of the proteins to these regions provides this functionality, the regions are known as transcription factor binding sites.


Figure 1.1: General structure of a gene. Adapted from (Maston et al. 2006).

### 1.4.1.1 Regulatory region of a gene

A typical regulatory region associated with a gene consists of a promoter element and distal regulatory elements (Maston et al. 2006). The promoter element consists of a core promoter and proximal promoter elements and typically spans less than 1 kb (kilobase) pairs (Maston et al. 2006). The core promoter of a gene is the region of DNA at which the proteins primarily responsible for transcription bind and initiate the process of transcription. Wellstudied elements of the eukaryotic core promoter include the TATA box and the initiator or Inr sequence (Brown 2006;Strachan and Read 2004). The TATA box generally has a consensus sequence of $5^{\prime}$-TATAWAW- $3^{\prime}$ where W is A or T (Brown 2006). The INR sequence has a consensus $5^{\prime}-$ YYCARR-3', where $Y$ is $C$ or T, and $R$ is $A$ or $G$ (Brown 2006). The TATA box and Inr sequence are generally present upstream of a large number of eukaryotic genes. Generally most of the elements of the core promoter are generally comprised of near identical DNA sequences.

The proximal promoter is generally located a few hundred base pairs upstream of the core promoter element (Maston et al. 2006). This region of DNA typically contains binding sites for other proteins, which contribute to the transcription of the gene but are not the primary mechanism (Maston et al. 2006).

Distal regulatory tend to be further away from the transcribed portion of the gene and contains elements that either activate or repress the transcription of the gene. Elements that activate transcription are known as enhancers and conversely elements that repress it are known as silencers (Raab and Kamakaka 2010).

### 1.4.2 Transcription

A family of enzymes known as RNA polymerases carry out the process of transcription of DNA into RNA in eukaryotic cells (Brown 2006). This process is known as transcription as the fundamental chemical language is not changed (Alberts 2002). There are three RNA polymerases typically encoded by the eukaryotic genome (Strachan and Read 2004). RNA polymerase I and RNA polymerase III tend to transcribe genes which code for functional RNA molecules, while RNA polymerase II is generally utilised for the production of RNA which is further translated into a protein (Alberts 1998). Transcription proceeds via the following general steps (Brown 2006):

- A protein known as TATA binding protein (TBP) binds to the TATA box sequence. This causes a bend in the DNA molecule.
- This bend provides a recognition signal for other transcription factors to bind to the DNA creating a structure known as the preinitiation complex (PIC) (Brown 2006). The formation of the PIC also disrupts base pairing thus creating a single stranded DNA template from which the RNA molecule is synthesised.
- RNA polymerase binds to the PIC and them moves along the single strand on DNA creating a complementary RNA molecule that conforms to base pairing rules. This RNA molecule is known as the primary transcript.


### 1.4.2.1 Post Transcriptional processing

After the primary transcript has been produced it is subjected to further modifications. In the case of primary transcripts associated with protein coding gene the primary transcript is also known as pre-mRNA (messenger RNA). In order to explain why these modifications occur it is necessary to understand how RNA molecules specify corresponding polypeptide molecules.

### 1.4.2.1.1 Genetic Code

It was established in work by Francis Crick (Crick et al. 1961) that polypeptide chains are specified by RNA molecules via triplets of nucleotides known as codons. As there are only twenty naturally occurring amino acids in eukaryotic proteins, and $4^{3}=64$ possible triplets from the 4 nucleotide types, the genetic code is redundant. Three of the codons specify the termination of the polypeptide chain and the remaining 61 specify amino acids.

The table below presents the genetic code

SECOND BASE


Table 1.2: The genetic code (Brown 2006).
The process by which these codons are translated into these amino acids will be presented in the next section. This code is widely utilised though there are a number of exceptions where a different code is utilised, e.g. in translation of mitochondrial genes (Knight et al. 2001).

### 1.4.2.1.2 Open reading frames

Given this code a sequence of triplets that specify a chain of amino acids commencing with a start codon and ending with a stop codon can be defined as an open reading frame (ORF) (Brown 2006). An open reading frame can exist in 6 possible orientations as there are two strands to a DNA molecule and an ORF can start from the first, second or third nucleotide within either strand as illustrated below.


Figure 1.2: Starting positions for possible ORFs within a double stranded DNA molecule.

### 1.4.2.1.3 Exons/Introns

ORFs as discussed above are subsections of the primary transcript or pre-mRNA molecule. ORFs are interrupted within pre-mRNA by sections known as introns (Brown 2006). The sections of the ORF thus separated by the introns are known as exons (Brown 2006). Thus in order to produce a molecule containing the full-uninterrupted ORF it is necessary to excise the introns and splice the exons together as shown in Figure 1.4.


Figure 1.3: Exons and introns within a pre-mRNA molecule.

## Exon 1 Exon $2 \quad$ Exon 3

Figure 1.4: Exons post splicing.


Figure 1.5: Exons post splicing in an alternate configuration.

It is not necessary however for all the exons within a given ORF to be utilised (Brown 2006) as shown in Figure 1.5. Different permutations of exons can be created to produce different protein molecules. This process is known as alternative splicing and is responsible for the disparity between the number of genes within a eukaryotic genome and the number of proteins it is capable of producing (Strachan and Read 2004). Alternate splicing is a feature of higher eukaryotes and contributes to overall protein diversity (Black 2003). Estimates of how many human gene products are alternately spliced include $60 \%$ (Black 2003) and $74 \%$ (Johnson et al. 2003).

### 1.4.2.2 Post Transcriptional processing (cont)

Having now discussed the necessity of posttranscriptional modification it is now possible to move on to the mechanisms by which splicing is carried out as well as covering other elements of posttranscriptional processing.

### 1.4.2.2.1 RNA splicing

As mentioned above the primary transcript or pre-mRNA is treated so as to excise intronic sequences and splice together exonic sequences. In order for this process to occur a necessary first step is the recognition of the borders between exons and introns. These areas are known as splice junctions (Strachan and Read 2004). It has been observed in a large number of cases
that introns in pre-mRNA commence with the sequence GU and end with the sequence AG (Strachan and Read 2004). These dinucleotides are not in themselves sufficient to signal a splice junction (Strachan and Read 2004) as splice junctions have been observed to show a greater degree of conservation (Breathnach et al. 1978). In vertebrates the following motifs have been observed at splice junctions (Brown 2006).

- 5 ' splice site $5^{\prime}-\mathrm{AG} \downarrow$ GUAAGU-3'
- 3' splice site 5'-PyPyPyPyPyPyNCAG $\downarrow-3^{\prime}$

In these consensus sequences the $\downarrow$ symbol indicates the border between an exon and intron or vice versa (Brown 2006). Py indicates that the nucleotide is a pyrimidine and N indicates that any nucleotide could be present at this position (Brown 2006). In addition to the conserved sequences at splice junctions introns also contain a conserved sequence around 40bp away from the end on the intron known as the branch sequence (Strachan and Read 2004). A large RNA-protein complex known as the spliceosome actually carries out the actual process of RNA splicing (Strachan and Read 2004). The spliceosome is one of the largest molecular machines in the human cell containing $\sim 170$ distinct proteins (Valadkhan and Jaladat 2010).

The process of RNA splicing typically involves the following sequence (Brown 2006; Strachan and Read 2004):

- Cleavage of the 5 ' splice junction detaching the exon from the intron at one end.
- The attachment of the cleaved $5^{\prime}$ end to the branch sequence forming a lariat like structure.
- Removal of the intronic lariat like RNA structure and the ligation of the two exons.


### 1.4.2.2.2 Capping

Another step in posttranscriptional modification of protein-coding genes is capping. This process is the first step in posttranscriptional processing of eukaryotic pre-mRNAs (Alberts 2002). This entails the addition of a methylated nucleoside (a nucleoside is a molecule consisting of a deoxyribose or ribose sugar bound to a nitrogenous base (Brown 2006)) to the
first 5' prime end of the transcript (Strachan and Read 2004). This process protects the transcript from rapid degradation via ribonuclease digestion (Strachan and Read 2004).

### 1.4.2.2.3 Polyadenylation

Post the termination of transcription the primary transcript is also modified via the addition of about 200 adenosine nucleotides to the 3 ' end of the transcript (Alberts 2002). This structure is known as a poly-A tail. The process is thought to facilitate the transport of the mature mRNA molecule into the cytoplasm (Strachan and Read 2004).

### 1.4.3 Translation

After a transcript associated with a protein-coding gene has been transcribed and processed, it then migrates to the cytoplasm, where a process known as translation occurs. This process entails the production a polypeptide chain that is specified by the transcript via the genetic code. The mature mRNA molecule is not synonymous with an ORF (Strachan and Read 2004). Generally an ORF is a subsection within the mature transcript. The ORF is flanked by sequences known as the 5' UTR and 3'UTR (UTR=untranslated regions) (Brown 2006) as illustrated in Figure 1.6.

ORF/ Translated
Region


Figure 1.6: Schematic of mature mRNA.

The process of translation occurs at cytoplasmic structures known as ribosomes. Ribosomes are large RNA-protein complexes, which consist of two subunits (Strachan and Read 2004).

The larger subunit is known as the 60S subunit and consists of three different types of ribosomal RNA (rRNA) molecule and up to 50 ribosomal proteins (Strachan and Read 2004). The smaller subunit is known as the 40S subunit and contains a single rRNA molecule and over 30 ribosomal proteins (Strachan and Read 2004). The two subunits of the ribosome exist as separate entities and attach for the process of translation.

The other molecule that provides the physical basis for the implementation of the genetic code is transfer RNA (tRNA). tRNA has a secondary structure consisting of four double helical structures as illustrated in Figure 1.7. tRNA attaches to an amino acid at its 3' end. The anticodon arm of the tRNA molecule has a triplet sequence, which is complementary to the codon of the amino acid to which it is bound. Thus tRNA attaches codons to their corresponding amino acids.


Figure 1.7: Structure of a tRNA molecule. Adapted from (Alberts 2008).

The process of translation typically proceeds via the following steps (Strachan and Read 2004):

- The two subunits of the ribosome attach to each other and also to a mature mRNA molecule at the methylated cap at the 5 ' end.
- The mRNA molecule is then pulled through the ribosome.
- When a start codon is encountered a tRNA molecule with an anticodon arm complementary to the start codon enters the ribosome. This tRNA molecule will have the relevant amino acid pre-bound to it.
- The next tRNA corresponding to next codon will then enter the ribosome.
- The amino acid attached to the first tRNA will detach from the tRNA and attach to the amino acid attached to the 3 ' end of the second tRNA.
- This process is iterated constructing a polypeptide chain or protein molecule.
- When a stop codon is encountered an enzyme known as a release factor causes the ribosome to disassociate and release the protein molecule.
- In order to prevent premature folding of proteins during translation the emerging polypeptide chain is stabilised by proteins known as chaperones (Alberts 2008).


### 1.5 Genomics

The term genome can de defined as the "entire genetic complement of a living organism" (Brown 2006). The field of study around ascertaining information about the genome of a living organism is thus known as genomics. The primary step of any full genomic study is the determination of the DNA sequence of the genome of the organism in question. Once this has been determined the next step is annotating the sequence.

### 1.5.1 Genome annotation

The full genome of an organism is generally a mosaic of functional and non-functional elements. The percentage of an organism's genome that is functional is variable. In the case of the human genome it has been calculated that potentially between $2.56 \%$ and $3.25 \%$ is functional (Lunter et al. 2006).

Functional elements in a genome include:

- Genes.
- DNA binding sites.
- CpG Islands: These are stretches of the dinucleotide repeat CG. These areas of DNA are subject to methylation, which is a form of epigenetic control over gene transcription (Kawaji and Hayashizaki 2008).

Genome annotation can be described as the systematic location of these functional elements within a genome sequence (structural annotation) and the ascertainment of that function (functional annotation). The location of functional elements is based on the principle of sequence specifying function. Thus the sequence of a functional element will vary in some detectable way from the remainder of the background sequence.

### 1.5.2 Genome and cDNA assembly

The initial challenge post the generation of sequencing data is the fact that the output of DNA sequencing is generally reads of short stretches of DNA. These reads range in length from > 700 bp long for Sanger sequencing (Hert et al. 2008) and ~200bp for pyro sequencing (Sundquist et al. 2007) and down to $\sim 50 \mathrm{bp}$ for ligation based sequencing methods (McKernan et al. 2009).

These short reads have to be assembled into a full sequence for the whole genome. This process is known as contig assembly. Contig assembly is carried out through scanning a set of short reads for overlaps. The discovery of an overlap indicates that two fragments are contiguous and should be connected. This process is necessary both at the level of the full genome as well at the level of the individual gene (Wang et al. 2005a).

### 1.5.3 Gene detection

Given a fully sequenced and assembled genome lacking annotation there are a number of computational techniques available to delineate coding sequence. These can be divided into two main subtypes: extrinsic and intrinsic (Borodovsky et al. 1994). Extrinsic methods utilise comparisons of sequence data to an external reference point while intrinsic methods evaluate sequences based on properties that are internal to the sequence (Borodovsky et al. 1994).

Construction of a cDNA library is one of the standard methods of extrinsic gene detection. cDNA stands for complementary DNA and is created through application of an enzyme known as reverse transcriptase to mature mRNA. Reverse transcriptase as the name implies reverses the process of transcription and creates a DNA strand complementary to the single stranded mRNA. Further steps are then taken in order to create a double stranded DNA molecule (Strachan and Read 2004).

A library of cDNA sequences is compiled through the collection of mRNA molecules from cells under various experimental conditions. This RNA is then converted to cDNA using the enzyme reverse transcriptase. The resultant cDNA is then amplified using the
polymerase chain reaction (PCR) (Mount 2004) and then sequenced. The library of sequences thus generated corresponds to the sequence of protein coding genes within the genome minus the introns. These sequences are then systematically mapped onto the genomic sequence using local alignment algorithms. This technique is known as cis-alignment. There are a number of local sequence programs that can be used to carry out these alignments. Exonerate is one such program. It utilises a bounded dynamic programming approach (Slater and Birney 2005) to generate local alignments. Dynamic programming is discussed in more detail later in this chapter. Another program, which can be utilised, is Spidey (hosted by the NCBI). This program employs the Blast heuristic algorithm (Altschul et al. 1990) to generate its alignments. SIM4 is another program that utilises an algorithm based on Blast but tailored to the specific problem of mapping cDNA to genomic DNA by factoring in introns and potential sequencing errors (Florea et al. 1998).

The Ensembl automatic genome annotation system (Curwen et al. 2004;Potter et al. 2004) uses the algorithm GeneWise (Birney et al. 2004) to map cDNA to full genomic data and the algorithm GenomeWise (Birney et al. 2004) to create a final putative structure for the gene in question post the initial alignment. cis alignment can be considered to be one of the most reliable methods for protein coding gene detection/prediction (Brent 2008).

In cases where cDNA libraries are not available or incomplete for the organism under consideration it is also possible to use cDNA sequences of homologous genes from either the same species or a different species in order to detect coding sequence. This technique is also referred to as trans-alignment and is central to various gene prediction tools (Brent 2008). The GeneWise (Birney et al. 2004), algorithm is also used in this context by the Ensembl pipeline (Potter et al. 2004). Extrinsic methods for genome annotation are far more cost and labour intensive as opposed to the strictly in-silico intrinsic approach.

Intrinsic approaches to gene detection are predominantly computational and as such require an explicit definition/description in order to delineate between coding and non-coding sequence (Picardi and Pesole 2010). Picardi (Picardi and Pesole 2010) gives a good working definition of a gene for detection purposes, which defines a gene as a transcribed region of DNA whose expression is regulated by cis acting elements such as upstream promoters. Examples of tasks undertaken as a part of intrinsic gene detection include:

- ORF (Open Reading Frame) detection: Detection of a potential ORF in genomic DNA is an indicator of a potential gene (Mount 2004). As prokaryotes in most cases (exceptions are pointed out in (Edgell et al. 2000)) lack exons and introns ORF
detection drastically reduces the search space for potential genes in the case of prokaryotes.
- Promoter regions detection: Genes are typically associated with one to several promoter regions. In prokaryotes these include the upstream Pribnow box with the consensus sequence TATAAT. This sequence is homologous to the eukaryotic TATA box (Berg et al. 2007). Detection of these motifs within a sequence upstream of an ORF strengthens the case for a potential gene.
- Internal splice junction detection: As the sequence of exon intron borders is broadly conserved discovery of splice junctions can also contribute to the case for a prospective gene.

These features can be can be detected within a stretch of sequence using various techniques to model sequence motifs ranging from simple regular expressions to hidden Markov models and position weight matrices (Picardi and Pesole 2010). Examples of specific applications of the intrinsic approach to gene prediction include SNAP (Korf 2004) and Genscan (Burge and Karlin 1997) both of which utilise Markov models in order to detect delineating features of genes. The primary weaknesses of the intrinsic approach lie in the fact that that it requires a representative sample of protein coding genes specifically from the organism under consideration in order to operate (Aubourg and Rouze 2001).

### 1.5.4 Functional annotation of genes

After a putative gene has been identified the next stage is determination of the exact biological role of the product coded for. This process can be carried out computationally or by entirely laboratory based techniques.

### 1.5.4.1 Laboratory based techniques

Laboratory based techniques for determination of biological function involve alteration of the gene in question either in the organism of study (in the case of prokaryotes, unicellular eukaryotes as well as higher eukaryotes which are deemed suitable) or in the case of organisms where modification would be impractical or unethical such as Homo sapiens alteration of the homologous gene in a model organism. The main model organism of choice for study of mammalian gene function is Mus musculus (Kim et al. 2010). The main alterations that are possible include:

- Knockouts: This entails the removal of the gene in order to observe the effects of its absence. This technique is only effective if the gene in question is not essential to organism survival and has a visible/measurable effect on phenotype (Moore 1999).
- Alteration in expression: In cases where the gene in question is essential to the survival of the organism, alterations can be made to the cis-regulatory regions of the gene in question in order to affect levels of expression (Capecchi 2005).
- In order to physically pinpoint specific tissues (in the case of multi-cellular organisms) or areas within a cell that a protein is active it is possible to place a reporter gene such as GFP (green fluorescent protein) upstream of the promoter region of the gene of interest (Chalfie et al. 1994).
- Detection of genetic interactions: The interaction of two non-essential genes (and hence their associated proteins) can be detected if the mutation of both genes leads to lethality (von Mering et al. 2002). This method has been applied to a large-scale study in Saccharomyces cerevisiae in order to characterise its set of genetic interactions (Ooi et al. 2006). The detection of an interaction partner of known function can aid in the determination of the function of an unknown gene.


### 1.5.4.2 Computational methods for functional annotation of genes

Computational methods to determine gene function have only become applicable relatively recently as most computational methods depend on comparison of novel sequence data with sequence of known function. Computational methods of functional annotation of genes can be split into a number of broad categories (Pellegrini 2001).

- Alignment based methods.
- Genome Context methods.


### 1.5.4.2.1 Alignment based methods

Sequence alignment is a problem that has been at the heart of bioinformatics since the inception of the field. The basic sequence alignment problem is searching two strings for areas of similarity (Mount 2004). The products of genes with similar/identical sequences are extremely likely to carry out the same function. Genes that share a significant degree of sequence similarity are potentially homologous (descended from a recent common ancestral gene) to each other. Using these methods the results of laboratory-based annotations only need to be carried out on one representative of a given set of identical sequences and the
derived functional annotation can be applied to all members. Alignment methods can be applied at either the gene or the protein level.

There are three primary ways of carrying out pairwise sequence alignments.

- Dot matrix analysis: This method entails arranging one sequence horizontally and the other sequence vertically perpendicular, starting from the left end of the horizontal sequence. Matches between the two sequences are then marked with a dot. Areas of similarity can then be viewed as diagonal lines between the two sequences (Mount 2004).
- Dynamic programming: Dynamic programming is a programming paradigm which entails the reduction of a large problem to a series of sub-problems whose solutions are constructed incrementally and summed to provide the overall solution (Russell et al. 2003). In terms of sequence alignment it entails the construction of a matrix similar to the dot matrix and calculating a path through it, where the next step in the path is determined only by the state of the current cell and its neighbouring cells. Two popular dynamic programming algorithms utilised in pairwise alignment of sequences are the Needleman-Wunsch algorithm (Needleman and Wunsch 1970) which returns an optimal global alignment of two sequences and the previously mentioned Smith-Waterman algorithm (Smith and Waterman 1981) which provides an optimised local alignment. Both of these algorithms are proven to return the optimal alignment between two sequences (Mount 2004).
- Heuristic Algorithms: Both the dynamic programming algorithms mentioned above are $O\left(n^{2}\right)$ in terms of both memory utilisation as well as time taken to run (Mount 2004). As such heuristic algorithms such as Blast (Altschul et al. 1990) and Fasta (Pearson and Lipman 1988) were developed as usable alternatives. The Fasta algorithm constructs a sequence alignment by searching for matching sequence patterns called $k$-tuples. These patterns are $k$ consecutive matches between the two sequences. These matches are then extended to provide the alignment (Krane and Raymer 2003). Blast constructs an alignment in a similar manner by locating short matches and then building an alignment around it. The difference between Blast and Fasta is that while Fasta examines all possible $k$-tuples the Blast algorithm is restricted to only examining matches that are significant and score over a given threshold (Mount 2004). These matches have to be of a length to achieve significance. This is 3 for proteins and 11 for DNA. Significance for proteins is
judged through use of the BLOSUM62 substitution matrix (Mount 2004). Given the rapid expansion of most of the large sequence databases it is typical to use heuristic algorithms as a search tool.
- Profile Hidden Markov models have been used by Eddy (Eddy 1998) to create a scoring system, which allows detection of remotely homologous sequences. Hidden Markov models score the probability of a discrete chain of events based on model parameters whose values are unknown (Durbin 1998).

Alignment methods can also be applied to the three dimensional structures of protein molecules as well as sequence (Hasegawa and Holm 2009). This method is potentially useful in cases where sequence divergence reaches a point where two proteins can no longer be identified as homologous. However as the rate of structure generation lags behind sequence generation by a considerable degree this method can only be applied in a small subset of cases.

Detection of a significant alignment with a gene of known function can be used to attach the same function to a gene of known function. Martin (Martin et al. 2004) used GO terms (Ashburner et al. 2000) in conjunction with Blast (Altschul et al. 1990) to achieve this with some success. There is however a danger with alignment based methods of a "Chinese whispers" effect where if for example a gene $p$ with known function $a$ displayed $90 \%$ identity using some form of pairwise alignment algorithm with gene $q$ of unknown function. Assigning function $a$ to gene $q$ would seem to be intuitively legitimate. However if gene $q$ was assigned function $a$ and the process was iterated a number of times a situation could arise where a gene $x$ would be assigned function $a$ with little or no sequence similarity to the original protein $p$. Examples of incorrect annotation by automated methods of homology detection occur in the case of genes where translations of the antisense strand of the coding region are entered into databases such as GenBank (Linial 2003).

### 1.5.4.2.2 Genome context methods

The recent proliferation of genome data has made it possible to detect and assign function to proteins through examination of their genomic context. Genome context methods compare and contrast the context of a gene between genomes (i.e. the arrangement of its homologues) in other genomes. Context methods are based on the principle of "guilt by association" which is the hypothesis that genes, which show proximity or association by some measure, e.g. phyletic distribution or chromosomal ordering are functionally associated (Aravind 2000).

Thus through demonstration of functional association or interaction between one gene/protein of known function with one of unknown function, the latter entity may be annotated with the function of the former.

### 1.5.4.2.2.1 Rosetta stone

The Rosetta stone method or detection of domain fusion was recognised through work by Marcotte (Marcotte et al. 1999) and Enright (Enright et al. 1999) which showed that sets of separate proteins in one organism which exist in a unified (fused) homologous form in another organism are likely to be interaction partners. As fusion events are comparatively rare and generally affect genes that are tightly functionally coupled this method is effective at detection of interaction partners (Kensche et al. 2008). However the rareness of these events lowers the overall coverage of this method.

### 1.5.4.2.2.2 Gene neighbour

Examination of the genomes of nine bacterial and archaeal genomes by Dandekar (Dandekar et al. 1998) showed that the proteins encoded by genes which showed conserved physical order along a chromosome tended to interact physically.

### 1.5.4.2.2.3 Interolog detection

A term introduced by Walhout (Walhout et al. 2000) an interolog is a pair of proteins that interact in a given organism. If both proteins involved in the interaction are conserved in another organism a similar interaction can be inferred in the second organism. This method has shown comparable accuracy with large-scale experimental data (Yu et al. 2004b).

### 1.5.4.2.2.4 Phylogenetic profiling

Phylogenetic profiling is a method that operates on the "hypothesis that functionally linked proteins evolve in a correlated manner" (Pellegrini et al. 1999). Consider for example a group of genes/proteins, which exist as a self-contained modular group and are associated with a particular cellular function. If this associated function was no longer needed by a given set of organisms the selective pressure to maintain all the genes/proteins within that group would be lowered thus leading to an eventual correlated cascade of losses for the genes in question. Genes are primarily lost through psdeudogenisation, which is the conversion of a functional gene to a non-functional copy. This can be caused by mutations that cause the premature truncation of a transcript through the creation of a premature stop codon or a
mutation in upstream cis-regulatory sequences thus removing the potential for transcription (Brown 2006). Pseudogenes can also be formed through retrotransposition of mature mRNA (Graur et al. 1989).

Thus through examination of multiple genomes for correlations in the presence and absence of proteins potential functional linkages can be detected. A phylogenetic profile is typically a binary string representing the presence or absence of a homolog of a given gene/protein. Predictions are made through examination of levels of similarity between these strings. These suggestions are suggestive in their nature rather than specific as it is unclear what the nature of a functional linkage between two proteins with similar profiles might be. The relationship could be a direct physical interaction such as subunits involved in heterodimerisation or more indirect such as the link between a transcription factor and the product of its associate gene.

The first use of phylogenetic profiles to predict functional linkages used Hamming distance as a metric in order to cluster similar profiles (Pellegrini et al. 1999). The Hamming distance of two strings can be defined as the number of points at which they differ (Hamming 1950). There have been various extensions and reinterpretations of the method since then (Ranea et al. 2007). Some of these involved examination of profiles using higher logical operations to carry out more complex comparisons of profiles (Bowers et al. 2004; Antonov and Mewes 2008). The method was also applied to protein domains rather then whole sequences (Pagel et al. 2004b). Work by Ranea utilised domain information from the Gene3D database to create phylogenetic profiles of the presence and absence of structural domains within genomes (Ranea et al. 2007). This method thus bypasses the problem of identification of genes that are functionally homologous by focussing on the presence and absence of predefined domains within proteins. Chen and Vitkup used examination of correlation coefficients to measure similarity in phylogenetic profiles (Chen and Vitkup 2006). They observed that the method was successful in identifying genes that were members of the same metabolic pathways (Chen and Vitkup 2006).

As a tool phylogenetic profiling could be used to detect errors in genome annotation through the detection and displays of gene absences, which are not plausible in closely related species. A similar approach has in fact been used by Pinney to detect and annotate enzyme-coding genes in the protist $E$. tenella (Pinney et al. 2005).

Other extensions to the method involved the utilisation of the phylogenetic relationships of the organisms include work by Barker and Pagel (Barker and Pagel 2005).

This method made use of an explicit phylogeny and ancestral reconstruction over the phylogeny based on a continuous-time Markov model. The likelihood of a model of dependent or contingent evolution was compared with the likelihood of a model of independent evolution over the phylogeny. This method was then further extended by investigating the effects of constraining the rate at which genes could be acquired over the phylogeny (Barker et al. 2007).

Other methods of incorporating phylogenetic information included the work by Vert (Vert 2002), which utilised support vector machines, as well as the work by Cokus (Cokus et al. 2007), which utilised phylogeny as a heuristic by ordering profiles by the phylogenetic closeness of the organisms involved.

### 1.5.4.2.2.5 Comparative methods

Comparing phylogenetic profiles over a phylogenetic tree can be considered to be an application of the comparative method to traits at the molecular level. The comparative method is a well-established method in biology (Harvey and Pagel 1991). The fundamental idea of underpinning the comparative method is how the state of one factor (which can be a trait or environmental condition) influences the state of another over the context of a topology of a phylogenetic tree (Maddison 1990). Testing for correlations without considering the phylogeny will detect correlations in gene content based on phylogenetic relationships rather than functional linkage. For example the set of all genes that are intrinsic to the class Mammalia will share similar phylogenetic profiles. This does not however suggest that they are all functionally linked.

There are a number of tests that have been developed in order to test the correlations in the states of traits over a phylogeny. Ridley (Ridley 1983) developed one of the earliest of these tests. This test involved the construction of a $2 \times 2$ contingency table where the state of each trait was considered as a categorical variable defined at each node in the tree. The method assumed that the construction of an accurate phylogeny and accurate reconstruction of ancestral context for each node within the phylogeny. Ridley's method did not however differentiate between dependant and independent variables in measuring the significance of a given set of changes (Maddison 1990). The method did not take into account the sequence of changes in the states of traits (i.e. was a change in state $A$ followed by a change in state $B$ or vice versa). This makes the results of the method difficult to interpret (Maddison 1990).

Joe Felsenstein (Felsenstein 1985b) developed another test for correlations in traits over a phylogeny. This test was developed to measure continuous data and modelled changes
over a tree as a Brownian process. Another test for detection of correlations in traits and/or external environmental conditions was devised by Grafen. This test was a phylogenetically corrected regression, which did not rely on any form of ancestral reconstruction (Grafen 1989).

Maddison developed a similar test to Ridleys in 1990 (Maddison 1990). It however did distinguish between dependant and independent variable by defining areas of a phylogeny to be in state $A$ or state $B$ depending on the state of one of the traits under consideration. The test then measured how many of the changes in the other trait occurred in the area of the tree that was in state $A$ compared to how many changes were possible over the whole tree.

One of the issues with the tests described above was the fact that none of them integrated information on branch lengths of the phylogeny. This meant that the probability of a change in the state of a given trait was equally likely over a branch of a phylogenetic tree regardless of its length. However clearly a change on a short branch is less likely than a longer branch. Work by Pagel took this into account by integrating branch lengths into a test for correlated evolution (Pagel 1994). The parameters defined by this work were utilised by Barker and Pagel in their approach to phylogenetic profile analysis (Barker and Pagel 2005).

### 1.5.4.2.2.6 Mirror trees

Another method of detection potential protein interactions is known as mirror trees. This method involves the detection of protein interactions through the construction and comparison of phylogenetic trees of proteins with a single genome (Pazos and Valencia 2001). The rationale behind this method is similar to that of phylogenetic profiling. However correlation is sought not in the presence and absence of homologous genes but in the pattern of sequence evolution of interacting proteins. Trees are examined by examining distance matrices of homologous sequences for correlations. These matrices are the inputs used in the formation of the trees in question. The phylogenetic tree of any given protein in a genome will however carry signal from the speciation events, which shaped the genome of the organism in question. An upgrade of the method has been developed to take into account this background similarity (Pazos et al. 2005). Hakes and others have however pointed out that the evolutionary pressures as well as the functional constraints on duplicated genes differ depending whether the mechanism of duplication was whole genome duplication or smallscale duplication (Hakes et al. 2007). This indicates that sequence divergence and functional evolution are not necessarily correlated (Robertson and Lovell 2009). Thus any similarity in
the phylogenetic trees of functionally linked genes is more likely to be due to chance or as mentioned above due to background similarity.

### 1.5.5 Storage of functional information

With the exponential increase in sequence data that has been generated through the 2000s there have been a number of attempts with which to organise and contextualise function information surrounding genomic entities.

### 1.5.5.1 GO

A notable attempt to do this has been the establishment of a controlled vocabulary with which to describe the functional role of a gene as well as its physical location within the cell. The vocabulary is known as the Gene Ontology (GO) (Ashburner et al. 2000). GO associates a set of terms with gene products. These terms are known as GO terms and fall into three general domains. These are

- Cellular component: This is the physical location within the cell where the gene product is generally to be found.
- Biological process: This is the biological pathway or process that the gene product has been localised in.
- Molecular function: This is a lower level to the biological process domain and includes the specific molecular capabilities of the molecule in question. An example of molecular function could be the ability to bind a particular metal.

Terms are organised as a network starting from the root terms defined above. As the network is traversed starting from a root term, terms become more specific, i.e. if term $B$ is directly below term $A$ in the ontology then term $B$ is a subclass of term $A$.

### 1.5.5.1 KEGG

Another database that localises gene products within functional pathways is KEGG (Kyoto Encylopedia of Genes and Genomics) (Kanehisa 1997; Kanehisa et al. 2006). KEGG maintains a list of functional pathways of processes that occur within the cell. These processes are arranged in a similar manner to GO in that they start from general categories and become more specific.

### 1.6 Transcriptomics

The transcriptome of a cell can be considered to be the sum total of its genome that is transcribed into RNA. Studying the transcriptome can also yield insights into the functionality of gene products.

### 1.6.1 Microarrays

At the transcriptomic level the putative function of a gene can be at least partially determined through establishing the association of the expression of a particular gene with a particular external condition or treatment. This can be achieved through the use of glass slides known as microarrays (Mount 2004). These slides have oligonucleotides, which are subsections of a set of genes attached to them. Cells of the organism under study are subjected to variable experimental conditions. mRNA is then extracted from these cells, converted to cDNA and fused with a unique florescent dye. By examining the relative degrees of florescence for the colours associated with the two versions of the cDNA of the gene of interest it is possible to measure levels of gene expression in response to a given experimental condition. A variant of this involves using full cDNA molecules as the contents of the chip.

### 1.6.2 Other methods for transcriptome examination

Expression levels for a given environmental condition can also be measured through direct sequencing and counting through use of the SAGE (Serial analysis of gene expression). In this method mRNA is extracted from the cells of interest. A small section is excised from each mRNA molecule. A tag is then connected to each separate subsection. These subsections are then amplified and the tags counted thus providing a measure of gene expression levels (Velculescu et al. 1997). Another protocol for sequencing mRNA to detect gene expression levels has also been developed. This protocol is known as RNA-Seq and is made feasible through the utilisation of the high throughput nature of next generation sequencing (Wang et al. 2009b).

### 1.7 Proteomics

Proteomics in a similar way to genomics and transcriptomics is the study of the full protein complement produced by a cell. The proteomic level is the point where the connection between macromolecules and measurable phenotypes is first bridged. Proteins can be considered as making up close to the totality of both structural (e.g. microtubules) and active (e.g. enzymes) components of a cell. The function of a protein can be determined by the determination of its structure and/or the determination of its interaction partners.

### 1.7.1 Protein Structure

There are two main methods utilised to determine the three dimensional structure of a protein molecule (Brown 2006). These are:

- X-Ray crystallography: This procedure involves the production of a crystal from the protein of interest. X-rays are then fired through this crystal to acquire a backscatter diffraction pattern. This diffraction pattern can then be used to reconstruct the structure of the protein. X-ray crystallography is limited by the fact that it requires the protein to be able to crystallise (Brown 2006).
- NMR spectroscopy: NMR or nuclear magnetic resonance is electro-magnetic radiation produced by the absorption and re-emition of electro-magnetic radiation by the nuclei of atoms. By bombarding a protein with electro-magnetic radiation, these patterns of resonance can be used to work out the structure of the protein (Brown 2006).


### 1.7.2 Protein interactions

In terms of protein interactions there are two primary modes of protein interaction. The first is a direct physical interaction. Direct physical interactions between distinct proteins can occur in two contexts (Orengo et al. 2003). These are:

- Formation of a stable complex: A protein complex is a stable structure formed by two or more proteins to carry out a specific function. In order to maintain the structural integrity of a complex proteins within the complex have to maintain relatively long term direct physical interactions. The subunits of the ribosome are an example of a stable protein complex as well as the histone octamer and RNA polymerases (Orengo et al. 2003). Not all interactions within a protein complex are direct as members of a complex with more then two interacting partners do not necessarily have to be physically connected to every other protein within that complex.
- Transient interaction: These are functional interactions where proteins physically interact but also exist independently in their own right (Orengo et al. 2003). An example of a transient interaction is the interaction between the human proteins Rho and RhoGap, which triggers a signalling cascade, involved in cytoskeleton formation and cell proliferation (Nooren and Thornton 2003).

The other form of interaction between proteins is indirect interactions. Examples of these could be two proteins that have a role in a given metabolic pathway but whose production is temporally and spatially separated. Examples of indirect interactions include the interaction between SHC-transforming protein and mitogen-activated protein kinase 1 over several steps of the insulin-signalling pathway (Sasaoka and Kobayashi 2000).
The full collection of all protein interactions within a cell has been labelled the interactome.

### 1.7.2.1 Experimental detection of protein interactions

Protein interactions can be detected using a variety of techniques. The main techniques include:

- Yeast two-hybrid: In order to detect protein interactions one widely used (Marcotte et al. 1999) method is the yeast two-hybrid technique. This technique exploits the $S$. cerevisiae GAL4 transcription factor. This transcription factor has two domains that require physical proximity in order to operate. One of these domains binds DNA and the other domain is an activator for the transcription factor. A protein interaction can be detected by fusing two genes of interest to both of these domains respectively on separate plasmids and insertion of these plasmids into a yeast cell with a reporter gene upstream of the GAL4 transcription factor-binding site. Reporter gene transcription is only possible if the protein products of the two genes of interest were able to maintain a physical interaction (Griffiths 2002). The primary drawbacks to this method are the facts that all interactions must take place in the nucleus removing a large number of proteins from their native cell compartment and that only binary protein interactions can be tested for (von Mering et al. 2002). The yeast two-hybrid method does have a high rate of false positives. One reason for this is that pairs of proteins that stick together are not necessarily ever expressed at the same time or in the same tissue (Vidalain et al. 2004). Also some proteins such as heat shock proteins are inherently promiscuous in their binding affinities (Vidalain et al. 2004).
- Proteome chips: In a manner similar to the use of microarrays described above for the measurement of gene expression levels microarrays can also be used with proteins. By printing translations of 5800 ORFs from S. cerevisiae on to a microarray chip Zhu and others (Zhu et al. 2001) were able to detect 33 novel interactions for the multi functional calcium binding protein calmodulin. The drawbacks to this method are that it is low throughput and again is restricted to binary interactions.
- Mass spectrometry of purified complexes: In order to detect interactions that are not binary, complexes of proteins can be isolated using techniques such as tandem affinity purification. This technique entails the tagging of a protein of interest with a tag that allows the purification of the main protein and any complex partners that it might have. These complexes can be characterised through the use of mass spectrometry (von Mering et al. 2002).


### 1.8 Description of project.

This work details the development and application of a novel heuristic which combines application of the Barker and Pagel approach to phylogenetic profiling (Barker et al. 2007; Barker and Pagel 2005) in conjunction with a novel data filter. The Barker and Pagel approach to phylogenetic profiling will subsequently be referred to as constrained ML (maximum likelihood). It was observed over the course of this project that this method could be useful in elucidating novel protein interactions. Novel protein interactions will allow further elucidation and annotation of protein function through the principle of guilt by association as articulated above. The proteome of Homo sapiens is still filled with known unknowns in terms of protein-protein interactions. The HPRD (Prasad et al. 2009) currently contains 38,788 binary protein interactions and data on 998 protein complexes. Current estimates of the interactome size such as work by Stumpf (Stumpf et al. 2008), which estimates the size of the interactome as 650,000 , intimate that the majority of protein-protein interactions have not yet been elucidated. The potential of phylogenetic profiling to detect novel interactions has been demonstrated in work by Ramazzina (Ramazzina et al. 2006) where two novel genes involved in the degradation of uric acid were detected. The phylogenetic profiling method has also been successful in identifying enzymes of the MEP/DOXP pathway (Cunningham et al. 2000).

Chapter 2 details the construction of a eukaryotic phylogeny over 54 taxa as well as the phylogenetic profiles of known proteins within the human genome relative to the other 53 species which was one of the essential precursor steps to this study.

Chapter 3 contains the results of the application of the method in context and compares it to a comparable high throughput experimental technique. Specifically the method is compared to detection of protein-protein interactions as well as indirect functional linkages through co-expression of genes as measured by microarrays. The method is also compared to PIPs which is the protein interaction prediction system maintained by Barton (McDowall et
al. 2009; Scott and Barton 2007). This system makes novel predictions through the combination of different informative features.

Chapter 4 describes the construction of the data filter, which is based on Dollo parsimony. The filter reduces the size of the overall search space facilitating the use of the method for whole genome comparisons. This is achieved through the elimination of pairs of proteins, whose function cannot be detected via examination of patterns of presence and absence.

Chapter 5 presents a network of predictions generated as a putative human interactome of proteins, which are susceptible to this line of enquiry. This network is analysed for consistency with known data. A set of novel predictions is presented.

Finally Chapter 6 will sum up this work and present details on potential future directions.

## Chapter 2

## Reconstruction of eukaryotic phylogeny as precursor to comparative analysis

### 2.1 Introduction

Examination of the evolutionary histories of organisms is a fundamental step for any form of study of biological function as adaptation can only be examined within an historical context (Harvey and Pagel 1991). As a phylogeny is by definition an evolutionary history of species (Harrison and Langdale 2006) it is a necessary step within the process of a comparative study. In terms of examination of changes in gene content within a probabilistic framework it provides the necessary topology over which such changes occur. This is a fundamental parameter in any such model.

### 2.1.1 Homology

The fundamental object of any phylogenetic study, whether molecular or morphological, is the comparison of homologous structures within the organisms under consideration. When genomic data is under consideration homologous structures within organisms correspond to those genomic elements, which were present in the last common ancestor of the set of organisms under consideration. These elements can provide a measure of divergence (Fitch 1970). These elements if functional (which is implied by conservation) can either maintain their ancestral function or if sufficiently diverged have a new (or no) function. In discussions of elements of genomes (genes) there are a number of subclasses of homologous relationships. These are:

- Orthology: Genetic elements are orthologous if they are the direct product of divergence from a common ancestral species (speciation) (Fitch 1970).
- Paralogy: Genetic elements are paralogous if they are the product of a duplication event within a given species. Mechanisms of duplication include retrotransposition (insertion of reverse transcribed RNA back into a genome) and unequal crossover leading to tandem duplication of a portion of a chromosome (Hurles 2004). It is thought that these duplication events are a major force in creating and broadening genetic repertoires (Zhang 2003).
- Xenology: Genetic elements are xenologous if they are the product of a direct exchange of DNA between organisms (Fitch 2000). These exchanges are known to be far more prevalent in prokaryotes given their lack of a true nucleus and the existence
of plasmids (free floating segments of DNA) in some prokaryotes. Genes have also been observed as xenologous in eukaryotes. Xenologous genes in eukaryotes can be acquired via organelles, which are the product of endosymbiosis such as the mitochondrion and chloroplasts (Blanchard and Lynch 2000).

It is important for purposes of phylogenetic reconstruction to be able to draw a distinction between genes which are paralogous and which are orthologous. If paralogous genes are compared between species the distance between them does not necessarily reflect the overall genetic divergence between the species under consideration. Genetic elements that are orthologous provide information on levels of divergence between speciation events whereas those that are paralogous provide data on duplication events.

A converse relationship to homology is that of analogy where through convergent evolution genes that share no common ancestry develop and maintain sequence similarity due to similar demands placed on the organisms in question by their environment. A classic example of this at the molecular level is that of the convergent evolution of the enzyme lysozyme in both the langur monkeys of the Indian subcontinent (Semnopithecus entellus) and ruminants due to the similar requirements imposed by a herbivorous diet (Swanson et al. 1991).

### 2.1.2 Molecular evolution

The fundamental idea at the heart of modern biology is that of random mutations guided by natural selection producing adaptation, which allow an organism to thrive in a given ecological niche. The large-scale study of evolution at a molecular level has only recently become possible due to advance in DNA sequencing technologies. This has been extremely useful as random mutations occur at the molecular level and also DNA/ amino acids are the fundamental comparable common denominator across morphologically and physiological diverse species (Nei and Kumar 2000).

At the DNA level there are four basic types of mutation (Nei and Kumar 2000). These are:

- Insertions: These mutations are insertions of additional nucleotides into a sequence of nucleotides. These can be caused by replication errors (Brown 2006). If an insertion occurs within an open reading frame it can cause the frame to be shifted hence insertions in coding regions can also be referred to as frameshift mutations.
- Deletions: Deletions are the opposite of insertions. Deletions within an ORF can also cause a frame shift (Brown 2006).
- Substitutions: These mutations are also referred to as point mutations and involve the substitution of a nucleotide with any other nucleotide. Substitutions do not necessarily have to involve a single nucleotide (Brown 2006). There are two types of substitutions transitions which entail the replacement of a purine with another purine, e.g. A to G or a pyrimidine with another pyrimidine, e.g. C to T. The other form of substitution is a transversion, which involves the replacement of a purine with a pyrimidine or vice versa (Nei and Kumar 2000).
- Inversions: An inversion mutation involves the reversing of the sequence of a strand of DNA, e.g. the sequence TGA being replaced with AGT (Nei and Kumar 2000).


### 2.1.2.1 Synonymous and non-synonymous mutations

Recalling the genetic code where nucleotide triplets known as codons specify amino acids, mutations within coding regions can also be classified by the effect that they have on the potential protein product. Thus mutations where the amino acid specified is altered are known as non-synonymous mutations, while mutations where there is no effect on the amino acid specified are referred to as synonymous mutations (Nei and Kumar 2000). Most synonymous mutations occur in the third position of codons. Measuring the relative rate of synonymous vs. non-synonymous mutations is technique for the detecting of positive Darwinian selection (Nei and Kumar 2000).

### 2.1.3 Phylogenetic trees

The evolutionary relationship between organisms has traditionally been presented as a tree like structure starting first presented in 1801 by French botanist Augustin Augier (Stevens and Augier 1983). Intuitively it is fairly clear what an evolutionary tree represents. Mathematically a tree can be defined as an acyclic graph. A graph is an abstraction, which can be used to model binary relations between objects (Parida 2008). A graph $G$ can be defined as $G(V, E)$ where $V$ is a set of vertices or nodes and $E$ is a subset defined as $E \subset(V \times$ $V)$ (Parida 2008). $E$ is thus a subset of the set of all ordered pairs that can be created from elements of $V$. The elements of $E$ are referred to as the edges of the graph (Parida 2008). A tree is a graph where all vertices are connected possessing the property that any two vertices $v_{l}, v_{2} \in V$ are connected by a unique path (Parida 2008). A vertex in a tree that has one incoming edge is known as a leaf node (Parida 2008). All other vertices by contrast are known as internal nodes (Parida 2008).

In the case of a phylogenetic tree leaf nodes are extant taxonomic units or taxa and internal nodes are proposed hypothetical common ancestors as illustrated in Figure 2.1. A subsection of a phylogenetic tree can be referred to as a clade (Nei and Kumar 2000).


Figure 2.1: Sample phylogenetic tree. In this tree the extant taxa are nodes A, B and C while node E is an ancestral node for A and C .

### 2.1.3.1 Species trees and gene trees

There are two main types of phylogenetic tree that are commonly investigated. These are:

- Species trees: The topology of these phylogenetic trees represents the branching order of species. Thus internal nodes are hypothetical common ancestors for the nodes that succeed them. The split at these ancestral nodes represent speciation events. A
speciation event is considered to be the moment in time when two species were reproductively isolated from each other (Nei and Kumar 2000).
- Gene trees: Gene trees measure the degree of divergence between homologous genes within and/or across species. Thus internal nodes in a gene tree represent a hypothetical gene that existed prior to a mutation event that created its two immediate descendants (Nei and Kumar 2000).
Figures 2.2 and 2.3 illustrate the differences between gene trees and species trees.


Figure 2.2: A species tree. Adapted from (Brown 2006).


Figure 2.3: A gene tree. Adapted from (Brown 2006).

### 2.1.3.2 Topologies and branch lengths

The branching pattern of a phylogenetic tree is known as its topology. The topologies of phylogenetic trees can be rooted as above in Figure 2.1 or unrooted as present below in Figure 2.4.


Figure 2.4: Sample of an unrooted phylogenetic tree representing four taxa.

Theoretically the topologies of most phylogenetic trees are bifurcating, as ancestral nodes will split into two descendant nodes at a given point in time. Multifurcation is possible in phylogenetic trees. A node with more than two descendents is referred to as a polytomy. There are two types of polytomy. Soft polytomies where the multifurcation is attributable to a lack of information and hard polytomies where species genuinely split into multiple descendants simultaneously (Page and Holmes 1998). Most polytomies are treated as soft as simultaneous speciation is considered unlikely (Page and Holmes 1998).

The number of possible unrooted bifurcating tree topologies $B(t)$ can be calculated using the formula given below (Salemi and Vandamme 2003) where $t$ is the number of taxa under consideration.

$$
\begin{equation*}
B(t)=\prod_{i=3}^{t}(2 i-5) \tag{1}
\end{equation*}
$$

The number of possible rooted bifurcating topologies $B(t)$ can be counted using the following formula:

$$
\begin{equation*}
B(t)=\frac{(2 t-3)!}{2^{t-2}(t-2)!} \tag{2}
\end{equation*}
$$

Thus estimation of a phylogenetic tree is a problem that quickly becomes computationally intractable as the number of taxa rises.

Another attribute that can be added to a phylogenetic tree is the length of its individual branches. As the nodes within a tree represent taxa, the lengths of the braches between them represent the degree of evolutionary change between the taxa over time. A phylogenetic tree with branch length information is also known as an additive tree, a metric tree or a phylogram (Page and Holmes 1998).

### 2.1.3.3 Bootstrap support values

Another attribute commonly associated with internal nodes in phylogenetic tree is the bootstrap support value. This value reflects the amount of times a particular internal node or split is selected if a phylogenetic analysis is repeated on a random set of re-samples (with replacement) from the original dataset (Page and Holmes 1998).

### 2.1.3.4 Evolutionary models in tree estimation

In order to estimate the amount of evolutionary change between taxa, methods considered to be effective tree estimators, utilise models of evolution that specify information on the evolutionary rate of substitution between homologous stretches of nucleotide or amino acid data. These models are framed as $m \times m$ matrices where $m$ is the number of entities in the data type, i.e. 4 in the case of nucleotides and 20 in the case of amino acids. To illustrate, an example of the simplest substitution model possible for DNA is the Jukes-Cantor model, which assumes that nucleotide substitution occurs with equal frequency (Nei and Kumar 2000). Thus the substitution matrix for the Jukes-Cantor model is presented below where $\alpha$ represents this uniform rate of substitution.

|  | A | T | C | G |
| :---: | :---: | :---: | :---: | :---: |
| A | - | $\alpha$ | $\alpha$ | $\alpha$ |
| T | $\alpha$ | - | $\alpha$ | $\alpha$ |
| C | $\alpha$ | $\alpha$ | - | $\alpha$ |
| G | $\alpha$ | $\alpha$ | $\alpha$ | - |

Table 2.1: Rates of nucleotide substitution for the Jukes-Cantor model (Nei and Kumar 2000).

The methods of tree estimation that utilise these models of evolution include the distance method, tree estimation by Bayesian methods, and tree estimation by maximum likelihood (Felsenstein 2004).

In distance methods an evolutionary model provides a measure of evolutionary distance between taxa, whereas in probabilistic methodologies such as maximum likelihood and Bayesian methods they provide a measure of probability for a given set of substitutions between taxa. Evolutionary models can be calculated via a priori assumptions about the evolutionary process or can be constructed empirically by examining the rate of observed substitutions in homologous sequences. Examples of empirically calculated substitution matrices for amino acids include the PAM matrices created in the seminal work by Margaret Dayhoff (Dayhoff et al. 1978) and more recently the WAG (Whelan and Goldman 2001) and LG matrices (Le and Gascuel 2008).

### 2.1.4 Detection of homology in molecular data

In order to construct a phylogenetic tree, which represents the evolutionary history of a set of taxa using molecular data, it is necessary to compare homologous sequences. More specifically it is necessary to detect orthologous genes/proteins. These genes/proteins are the most appropriate measure of genetic divergence between species, as an equal level of genetic divergence will have occurred since the speciation event causing the split.

There are a number of algorithms, which are utilised in the selection of homologous genes/proteins and their subsequent classification as orthologous or paralogous. These include:

- Reciprocal Best Hits (RBH): This procedure is implemented by the COGs (Tatusov et al. 2003) database hosted by the NCBI. The underlying rationale of the algorithm is that orthologous genes between two species will possess more similarity with each other then with any other gene. This similarity is generally established using pairwise sequence alignment algorithms such as BLAST (Altschul et al. 1990) or the SmithWaterman algorithm (Smith and Waterman 1981).
- InParanoid: This algorithm extends the idea behind RBHs by using them to seed orthologous clusters, and then by an application of an iterative inclusion process constructs a set of gene/protein families (Remm et al. 2001).
- OrthoMCL: This process also utilises RBHs as seed pairs for clusters. Similarity relations between gene/proteins are then established as a graph and additional paralogous sequences are determined through a process of graph clustering (Li et al. 2003).
- Reciprocal smallest distance (RSD): This procedure does not utilise RBHs and instead, for a set of hits for a given query protein, over a given E-value (Expect value), conducts pairwise alignments between each of the hits and the original query. Hits that are alignable to a given threshold are then subjected to further analysis to calculate the number of amino acid substitutions or distance between them and the original query. The hit with the shortest distance is then used to reverse the process. If the reversal yields the original query then the two sequences are declared orthologous (Wall et al. 2003).
- EnsemblCompara GeneTrees: This is an algorithm utilised by the Ensembl Compara database (Vilella et al. 2009). The process involves RBHs. Two species are subjected to an all against all pairwise alignment. Like OrthoMCL the resulting data is then converted into a graph. This graph is then clustered. Gene trees are then constructed using these clusters and reconciled against a gold-standard species tree.

In comparative studies the Inparanoid algorithm (Remm et al. 2001) was shown to perform better than its rivals (Hulsen et al. 2006). This work showed the Inparanoid algorithm tied as the best performer with simple reciprocal best hits at identification of orthologs. However reciprocal best hits in practise only yield one to one orthologous relationships (Hulsen et al. 2006). This reduces the coverage of the method (Hulsen et al.
2006). OrthoMCL was shown to perform a close second to the Inparanoid algorithm in benchmarking tests (Hulsen et al. 2006). Subsequent benchmarking work (Altenhoff and Dessimoz 2009) showed that OrthoMCL outperformed Inparanoid to an extent at lower levels of specificity but higher coverage. However at points benchmarking was applied to data and organisms common to both reviews the results were seen as broadly congruent (Altenhoff and Dessimoz 2009).

### 2.1.5 Multiple sequence alignment

Given a set of orthologous sequences further processing is required in order to convert them into a suitable input for a phylogenetic tree estimation procedure. This input is known as a multiple sequence alignment (MSA) (Edgar and Batzoglou 2006). The process involves creating an optimal alignment between three or more protein sequences. Insertions and deletions between orthologous proteins are represented by introducing gaps into the alignment. Alignments are scored through the use of substitution matrices. The process converts orthologous sequences into a rectangular array where each column of the array corresponds to a homologous attribute between the taxa under consideration (Edgar and Batzoglou 2006).

Forms of multiple sequence alignment include.

- Progressive: This form of alignment involves the construction of initial pairwise alignments between all the sequences under consideration. The distances thus established between the sequences are used to create a guide tree. The multiple sequence alignment is then built up progressively in the order suggested by the guide tree (Mount 2004). The main flaw with this method is that errors made any stage of constructing the MSA remain in the final alignment (Wheeler and Kececioglu 2007). A very prominent example of a progressive MSA tool is the Clustal suite (Higgins and Sharp 1988).
- Iterative: In order to reduce the errors introduced by the progressive approach to MSA the iterative approach realigns sub-groups of the sequences repeatedly (Mount 2004). Examples of iterative MSA programs include MUSCLE (Edgar 2004) and DIALIGN (Morgenstern et al. 1998). The performance of the iterative approach can be improved by the inclusion of consistency information between the growing MSA and the pre-
computed pairwise alignments used by some of algorithms within the MAFFT (Multiple alignment by fast Fourier transform) program (Katoh et al. 2002).

The quality of a multiple alignment is crucial to the accuracy of the phylogenetic tree created via its analysis (Blair and Murphy 2011). This is especially true when there are gaps in the alignment (Talavera and Castresana 2007). Thus benchmarking tests have been carried out to examine the performance of various algorithms currently available. The results of these have found that MAFFT (running in its iterative, consistency enhanced mode) using the Smith-Waterman algorithm (Smith and Waterman 1981) for its initial pairwise alignment outperformed its nearest rivals (Ahola et al. 2006; Nuin et al. 2006). This mode of MAFFT is known as MAFFT-L-INS-i.

### 2.1.5.1 Multiple sequence alignment quality filtration

Given the effects of MSA quality on phylogenetic analysis it is argued that filtration of areas, which are problematic to align, will improve the outcome of subsequent phylogenetic analyses (Talavera and Castresana 2007). It is common practise to edit MSAs by hand before analysing them further though it is considered that this makes all results thus gained irreproducible through the subjectivity of the overall process (Blair and Murphy 2011). Thus this process has been semi automated by programs such as Gblocks (Talavera and Castresana 2007) and Trimal (Capella-Gutierrez et al. 2009). These programs will retain sections of MSAs, which are highly conserved and remove gaps in the alignment.

Gblocks will either remove all gaps in its stringent mode or only remove gaps if they are present in more than half the sequences in the alignment in its relaxed mode (Talavera and Castresana 2007). Trimal will remove columns from an alignment based on a conservation threshold defined by the user, i.e. how much of the original alignment does the user wish to conserve (Capella-Gutierrez et al. 2009). In benchmarking tests optimum performance for Gblocks in enhancing tree estimation was observed using its relaxed mode (Capella-Gutierrez et al. 2009).

### 2.1.6 Methods to estimate phylogenetic trees

The focus of this section as mentioned above shall be on the analysis of molecular data though the methods described are applicable to any form of measurable polymorphic trait. These data provide a measure of distance between the species under consideration.

The first subdivision in types of methods of phylogenetic analyses is between discrete character state and distance matrix methods (Salemi and Vandamme 2003). Discrete character state methods examine the differences in state of a set of discrete characters or traits. Distance matrix methods utilise the distance between sets of data through the creation of a matrix of pairwise distances and application of clustering techniques. Subtypes of the character state method include the maximum parsimony method that does not utilise an explicit model of evolution and maximum likelihood, which conversely does (Salemi and Vandamme 2003).

### 2.1.6.1 Distance methods

Distance methods were originally developed to construct phenograms, i.e. (diagrams which reflect the similarity between a given group of taxa without consideration of ancestor/descendant relationships (Salemi and Vandamme 2003; Sneath and Sokal 1973) as opposed to phylogenies. Distance methods however can also be applied to elucidating phylogeny under the assumption of equal rates of mutation in cases where a quick initial result is required.

Distance methods of phylogeny depend on the construction of a matrix of pairwise distances for the trait data of the organisms under consideration. This data is generally nucleotide and or amino acid sequence data though the method is also applicable to any other form of discrete descriptive data. In the case of amino acid or nucleotide data distances are estimated according to evolutionary models, which allow a meaningful calculation of the evolutionary distance between two species.

The simplest form of evolutionary distance measure is the proportion of differing sites between two sequences $p$. This is calculated through a simple count of differing sites $n_{d}$ and division by the total number of sites $n$ as shown in Equation 3 (Nei and Kumar 2000).

$$
\begin{equation*}
p=\frac{n_{d}}{n} \tag{3}
\end{equation*}
$$

$p$ is an underestimate of evolutionary distance over extended periods of time as multiple substitutions accumulate per site. Thus in order to represent this information substitutions can be modelled as a Poisson process over time and then the probability of $k$ mutations over $t$ time can be can be calculated by the standard Poisson distribution function where $\lambda=$ the rate of mutations / unit time and $e=$ the base of the natural logarithm (Nei and Kumar 2000).

$$
\begin{equation*}
p(k ; t)=\frac{e^{-\lambda} \lambda^{k}}{k!} \tag{4}
\end{equation*}
$$

This probability can then be used to calculate a distance between two sequences. This distance is referred to as the Poisson corrected distance (Nei and Kumar 2000).

The Poisson corrected distance assumes a homogenous rate of mutations / substitutions over a molecular sequence. This assumption however is not true as different areas of a sequence (coding or not coding in the case of nucleotides, for example) will be subject to differing selective pressure hence differing mutation rates (Nei and Kumar 2000).

This information is integrated into calculations of distance via the observation that variation in rates of substitution over a sequence follows a gamma distribution (Nei and Kumar 2000).

Having created a matrix of pairwise distances between the sequences under comparison this matrix can then used to generate a phylogenetic tree via clustering.

A commonly used form of clustering in the generation of distance-based trees is neighbour joining. This algorithm follows the following steps (Brown 2006):

- Construction of a fully multifurcating star shaped tree including all taxa under consideration.
- The selection of a random pair of taxa and removing them from the star to form a tree consisting of a clade containing that pair and a clade containing the rest of the star.
- Evaluation of the total branch lengths of the new tree.
- Iteration of this process for all possible pairs storing the results of the branch length calculation.
- Identification of the pair, which yields the first interim tree with the shortest branch length.
- This pair is now placed on their own branch and the process is iterated until a fully bifurcating tree is retrieved.

Another method of tree estimation involving distance matrices is least squares fitting in which for each tree the residual sum of squares is calculated between pairs of taxa. This method is known as the Fitch-Margoliash method. This involves applying the following equation (Nei and Kumar 2000).

$$
\begin{equation*}
R_{s}=\sum\left(d_{i j}-e_{i j}\right)^{2} \tag{5}
\end{equation*}
$$

Where $d_{i j}$ is the observed distance in the matrix between taxa $i$ and taxa $j$ and $e_{i j}$ is the patristic distance between the taxa. The patristic distance between two taxa is the sum of the branch lengths that make up the shortest path between the two taxa. The tree with the lowest $R_{s}$ is selected by the method. Generally tree space is searched using a heuristic search method as described below in Section 2.1.6.3.

Other standard techniques for this process are clustering methods such as UPGMA (unweighted pair group methods with arithmetic means), which group organisms by degree of closeness in the matrix. The underlying assumption of UPGMA is that the evolutionary process occurs at a consistent pace, i.e. follows a molecular clock (Felsenstein 2004). Thus in cases where data does not follow a molecular clock, UPGMA will deliver misleading results as it will cluster species on short branches with each other (Felsenstein 2004).

Another commonly applied method is minimum evolution, which creates a tree where the overall amount of evolution (measured by the total branch lengths of the tree from root to tip) is minimised (Salemi and Vandamme 2003). Again tree space is traversed by heuristic search as described below.

Distance methods are comparatively fast compared to character based methods and given a dataset with relatively constant rates of evolution and closely related taxonomic units fairly accurate (Felsenstein 2004). However they suffer from a systemic issue where if the taxa under consideration display variability of rates of evolution along a sequence at different points in a tree this cannot be detected as all distances between the sequence are calculated locally, i.e. between adjacent species (Felsenstein 2004)

### 2.1.6.2 Discrete character state methods

Discrete state character methods operate on matrices populated with assigned attributes or characters to each taxon under consideration. Possible trees are then evaluated against this matrix in an attempt to satisfy an optimality criterion (Salemi and Vandamme 2003). One of the two most popular optimality criterions is parsimony, which entails minimisation of the amount of change required over a given tree to produce the data observed in the matrix. The other widely used criterion for selection of trees is likelihood. This method frames the tree as a hypothesis for the matrix of observed data and evaluates its likelihood given the matrix of
observed data (Felsenstein 2004). Maximisation of the likelihood function yields the optimum tree.

### 2.1.6.2.1 Maximum Parsimony

Using parsimony, as a criterion for judging potential trees was first introduced by Camin and Sokal in 1965 (Camin and Sokal 1965). The rationale behind considering a tree that is more parsimonious is based on the principle of Ockham's razor, which can be stated, as a simpler explanation for an observed phenomenon is to be preferred to a more complex ad hoc explanation (Steel and Penny 2000).

Specific variants of parsimony that can be utilised are (Felsenstein 2004):

- Fitch parsimony: this form of parsimony is also known as Wagner parsimony (Felsenstein 2004). This form of parsimony allows all possible changes in any direction and counts them all equally.
- Camin-Sokal parsimony: this form of parsimony only allows evolutionary change in one direction (Camin and Sokal 1965). For example if a two state character which can take on states 0 and 1 is considered Camin-Sokal parsimony will only allow changes in the 0 to 1 direction (assuming that is the direction selected as permissible) (Felsenstein 2004).
- Dollo parsimony: this form of parsimony is based on the principle of phylogenetic irreversibility (Lequesne 1974). The acquisition of a complex character is allowed once and then all subsequent changes over the tree can only be reversions (Felsenstein 2004).
- Parsimony on an ordinal scale: this deals with the case where changes in a multi state character are considered on an ordinal scale. Thus only changes that are adjacent are allowed (Felsenstein 2004).
- Polymorphism parsimony: this form of parsimony allows an intermediate state of polymorphism to be acquired at a point within the tree. All changes subsequent to the polymorphic areas in the tree are counted as a loss of one of the composite characters that make up the polymorphic state (Farris 1978; Felsenstein 1979; Felsenstein 2004).

Evaluating the number of character changes required over a particular tree for a given character matrix is computationally easy and can be calculated rapidly through applications of dynamic programming algorithms such as:

- The Fitch algorithm (Fitch 1971): operates by carrying out a post order traversal of the phylogenetic tree (Felsenstein 2004). At each internal node the set of potential ancestral states is set to either the intersection of the states of its immediate descendant nodes if such an intersection exists. If no such intersection exists then the state of the nodes is set to the union of the states of the two descendant nodes.
- The Sankoff algorithm (Sankoff 1975): is similar but not identical to the Fitch algorithm (Felsenstein 2004). A cost matrix is created which stores the cost of all possible changes of state within the context of the data under consideration. Ancestral node states are then assigned by selecting the state with the minimal cost.
The complexity of both these algorithms is linear as the number of operations required increases linearly with the number of taxa, the number of characters under consideration and the number of states that those characters can take on.

Parsimony based methods have been popular due to their computational as well as conceptual simplicity. Parsimony methods were recently utilised to construct the largest phylogenetic tree ever reconstructed consisting 73,060 eukaryotic taxa with a combination of morphological and molecular data (Goloboff et al. 2009). However in general over the last two decades there has been a swing toward the use of likelihood-based methods (Steel and Penny 2000). This is due in part to the demonstration (Felsenstein 1978) that under given conditions maximum parsimony will converge on the wrong tree. These conditions have come to be known as the Felsenstein zone.

### 2.1.6.2.2 Maximum likelihood

Maximising the likelihood of a given tree as a hypothesis to explain observed data was first applied to phylogenetic inference by Edwards and Cavali-Sforza (Cavalli-Sforza 1964). The likelihood function assigns a value to the ability of a hypothesis to explain an observed set of results. Assume a statistical model $M$ that associates a probability with a set of possible outcomes and a set of observed outcomes (or results) $R$. Thus $P(R \mid M)$ is the probability of observing $R$ assuming $M$ is a correct description of the process under study (Edwards 1992). The likelihood $L$ of $M$ is defined as:
$L(M)=P(R \mid M) \times k$

Where $k$ is an arbitrary constant. Use of this constant allows relative comparison of likelihoods (Edwards 1992). To paraphrase an example from (Durbin 1998) in the case of a die if our hypothesis is that the die is fair then the probability of any outcome is equal to 0.16 . If we go on to roll 5 sixes then this forms our observed data. The likelihood of the hypothesis is then proportional to $0.16^{5}$ or 0.000104 . Hypotheses can thus be judged on their relative abilities to explain observed results. A hypothesis with a higher estimate of the probability of rolling a 6 would be better fit to the observed data in the case of the die. Thus likelihood provides a framework with which to select a hypothesis or model appropriate to the observed data.

In the case of phylogeny each tree is a hypothesis explaining the distribution of the traits under consideration. The phylogeny with the maximum likelihood is selected as the optimal tree. The likelihood of a tree can be measured through the application of a substitution model of evolution, which models the probability of individual evolutionary events over the tree. Empirically calculated substitution models can be used as a substitute for the calculation of a set of probabilities, which permits the application of more generalised rules of evolution to each individual phylogenetic study. Empirical models of evolutionary events can be created through the examination of homologous sequences in different species. Models currently in use for amino acid based phylogenies include the WAG (Whelan and Goldman 2001) and LG (Le and Gascuel 2008) substitution models.

If a model is badly specified and a poor fit for the data then likelihood methods can return an inaccurate tree with high statistical support (Keane et al. 2006). There are a limited number of cases where parsimony methods can outperform likelihood-based methods, which has been called the inverse Felsenstein zone, or Farris zone (Siddall 1998). It has been shown however that these cases are extremely rare in real data (Swofford et al. 2001) and in cases where it is computationally feasible maximum likelihood has become the one of the dominant paradigms in phylogeny reconstruction.

### 2.1.6.2.3 Bayesian Methods

Another criterion related to likelihood is the posterior probability of a tree given a matrix of observations and a prior probability for the tree. The posterior probability of a hypothesis is the probability of the hypothesis being true given some observed data. The posterior probability of a tree given a multiple alignment is calculated through the application of Bayes theorem, which is defined as:

$$
\begin{equation*}
P(X \mid Y)=\frac{P(Y \mid X) P(X)}{P(Y)} \tag{7}
\end{equation*}
$$

Where $X$ and $Y$ are separate events and $P(Y \mid X)$ is the conditional probability of event $Y$ given event $X$ has occurred. $P(X)$ is known as the prior probability of event $X . P(X \mid Y)$ is the posterior probability of event $X$ given event $Y$ has occurred. $P(X)$ represents a subjective prior belief in the probability of $X$ occurring.

In the case of phylogenetic analysis $X$ is a phylogenetic tree and $Y$ is a given multiple alignment. It is however non-trivial to evaluate the posterior probabilities over all possible tree topologies exhaustively (Huelsenbeck et al. 2001). This process had been made feasible by sampling the distribution of posterior probabilities of trees. The posterior probability of a particular tree is measured as the amount of times it is visited over traversal of tree space. Tree space traversal is facilitated by the use of Metropolis-coupled MCMC (Markov chain Monte Carlo) methods first introduced by the doctoral work of Li and Mau (Pickett and Randle 2005).

The algorithm returns a set of trees sampled from the posterior distribution. An individual phylogeny is then generally assembled from the returned sample through using majority rule consensus methods (Cranston and Rannala 2007).

Bayesian methods suffer from the potential source of bias of prior probabilities (Holder and Lewis 2003). This issue can be ameliorated through the use of flat or uninformative priors. Flat priors can however still bias a Bayesian phylogenetic study towards trees with particular configurations of clades (Pickett and Randle 2005).

### 2.1.6.3 Heuristic search methods

Given the large number of possible topologies possible for even a small number of taxa the estimation of phylogenetic trees is a problem that is intractable by brute force searching. Thus the space of all possible trees is usually searched heuristically (Felsenstein 2004). What this entails is the selection of a random first tree. This tree is then evaluated on the basis for whatever measure that has been defined to evaluate the "quality" of the tree. Examples of possible quality measures for a phylogeny include as previously discussed parsimony, likelihood and distance. The tree is then altered thus moving to a new point in tree space. This new tree is then evaluated. This process is then iterated until a local optimum point has
been reached within the space. This point is not guaranteed to be a global optimum within the space (Felsenstein 2004).

Examples of alterations/moves that are used to traverse tree space include (Felsenstein 2004):

- Nearest neighbour interchange (NNI): This process involves the swapping of adjacent branches within a tree. This is a local rearrangement of the tree.
- Subtree pruning and regrafting (SPR): This process involves the removal or pruning of a subtree from an overall tree and reattaching it at another point. As opposed to NNI this is a global rearrangement of the tree.
- Tree bisection and reconnection (TBR): This involves the deletion of an interior branch to split a tree into separate trees and then all possible connections are made between the branch set of the first tree and the second. This is also a global rearrangement of the tree.

Global rearrangements are more radical moves within the tree space and thus are less likely to stabilise in local optima. Modern phylogeny estimation programs generally provide the options to carry out either form of rearrangement. The advantage of using local rearrangements is greater speed in arriving at the optimum tree. Examples of programs, which offer this choice, are a number of programs within the PHYLIP suite (Felsenstein 1989) and PhyML (Guindon and Gascuel 2003). PhyML is generally as accurate as other phylogeny estimation programs while being considerably faster (Dereeper et al. 2008). Programs within PHYLIP can carry out multiple searches through the space jumbling the order of the taxonomic data to widen space coverage. The programs within PHYLIP that offer heuristic search are:

- PROTPARS
- DNAPARS
- DNACOMP
- DNAML
- DNAMLK
- PROML
- PROMLK
- RESTML
- FITCH
- KITSCH
- NEIGHBOR
- CONTML
- PARS
- MIX
- DOLLOP

Short descriptions of these programs can be found on the PHYLIP webpage (http://evolution.genetics.washington.edu/phylip/).

### 2.1.7 Bootstrapping

Bootstrapping was first proposed as a method for evaluating confidence limits in phylogenetic trees by Felsenstein in 1985 (Felsenstein 1985a). The procedure evaluates how well supported a particular tree topology is by a given dataset. This entails constructing a dataset created by random resampling from the original dataset. This new dataset should be of the same size as the original dataset. This procedure is repeated to produce the appropriate number of replicates. The original tree estimation procedure is then repeated on this subset producing a set of trees.

The original tree is then evaluated in the light of these new trees. Each interior branch of the original tree is compared to the bootstrap trees, and for every bootstrap tree, which contains a branch, which creates an identical partition of the data the branch is marked as present. Thus each internal branch gets a score or bootstrap confidence value calculated by dividing the number of times it was found to be present in one of the bootstrap trees with the total number of bootstraps (Nei and Kumar 2000).

In the PHYLIP package presented by Felsenstein bootstrapping is carried out through the use of two of its internal programs SEQBOOT and CONSENSE (Felsenstein 1989). This procedure calculates confidence values for a consensus tree created from the bootstrap trees as opposed to the original tree (Nei and Kumar 2000). The procedure for obtaining these bootstrap support values is:

- Run SEQBOOT on original dataset. This produces the number of required resamples/replicates of the dataset. SEQBOOT requires a random odd number as a seed.
- Repeat original estimation procedure on each replicate. This produces a set of trees all of which represent the original data.
- CONSENSE is then used to merge these trees together. CONSENSE is a program, which is designed to produce a consensus tree from a set of trees. CONSENSE will add internal values /branch which represent the confidence values in those branches.


### 2.1.8 Model selection in phylogenetic tree estimation

As mentioned above, three of the main methods of phylogenetic tree estimation utilise evolutionary models as described in section 2.1.3.4 with which to convert homologous data matrices into trees. The use of an inappropriate model has been shown to adversely affect all aspects of tree reconstruction including branch lengths and topology (Bruno and Halpern 1999). A poorly specified model will return an incorrect tree with high statistical confidence (Posada and Buckley 2004).

Thus procedures have been developed with which to estimate how well a given evolutionary model fits a dataset. A standard probabilistic measure that is used to measure the fit of a model to a given dataset is likelihood. As a model can be fitted (over fit) perfectly to a dataset by adding parameters, the likelihood of any given model is evaluated in the context of how many parameters the model has. Two standard measures for integrating this information (the likelihood of the model with respect to the dataset and the number of parameters) for doing this are the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). Both of these measures penalise parameter rich models over simpler models thus selecting the simplest model that explains the observed data adequately (Felsenstein 2004). The AIC is calculated by the following equation (Felsenstein 2004).
$A I C_{i}=-2 \ln L_{i}+2 p_{i}$

Where $L_{i}$ is the likelihood of model $i$ and $p_{i}$ is the number of parameters in model $i$. The BIC has a higher penalty for parameter richness than the AIC and is calculated via the following equation (Felsenstein 2004).
$B I C_{i}=-2 \ln L_{i}+p_{i} \ln (n)$

Where $L_{i}$ is the likelihood of model $i$ and $p_{i}$ is the number of parameters in model $i$ and $n$ is the number of data points in the dataset.

An example of a model selection procedure involving likelihood, utilised by a model selection tool ModelGenerator is as follows (Keane et al. 2006).

- The construction of a simple guide tree using neighbour joining on the dataset.
- Each model to be evaluated is then examined over this guide tree and the dataset to calculate the likelihood of that model.
- The AIC/BIC are calculated for the model.
- The model with lowest AIC/BIC is presented as the best model for the given dataset.


### 2.1.9 Comparing phylogenetic trees

Given sets of two or more phylogenetic trees produced by disparate estimation procedures it can sometimes be necessary to compare trees for congruence with each other as well as with the datasets underlying them.

There are a number of procedures that can be followed to compare trees. The simplest of these is attempting to visualise the parts of trees that are topologically similar. A method for doing do is presented in work by Nye (Nye et al. 2006).

Other measures have been designed to define the difference between two trees. These include:

- The Robinson-Foulds distance: This is a distance that counts how many branches differ between two trees. This is done by ignoring branch lengths and considering each tree as a set of branches. Each branch splits the tree into two partitions. The Robinson-Foulds distance is a count of the partitions present in one tree and not in the other (Felsenstein 2004).
- The NNI distance: This distance can be considered an edit distance, analogous to the Levenshtein distance used to compare strings of text. It is the number of NNI operations it would take to transform one of the trees into the other (Felsenstein 2004).
- The Branch Score distance: This measure uses branch lengths as well as topology to calculate the distance between two trees (Felsenstein 2004).

Another way to measure the quality of a pair of estimated tree relative to a given dataset is use of the Kishino-Hasegawa ( KH ) test. This is a test of how well individual homologous sites within a dataset support a given tree in contrast to another tree (Goldman et al. 2000). Both trees are selected a priori as the possible best hypotheses for an observed dataset. The test was first introduced in the work (Hasegawa and Kishino 1989). The underlying rationale is that if the trees are equally well supported by the dataset. Thus using the notation provided in (Goldman et al. 2000) the test can be carried out via the following procedure.

- Given two trees $T_{1}$ and $T_{2}$ calculated by a given quality criterion, e.g. parsimony or likelihood.
- Assuming for the purposes of explanation that the quality criterion is likelihood, the likelihoods of $T_{1}$ and $T_{2}$ (with respect to a given dataset $D$ ) are $L_{I}$ and $L_{2}$ respectively.
- Calculate $\delta$ as the difference between the two likelihoods, i.e. $\delta=L_{1}-L_{2}$.
- The underlying hypothesis of the test is that $T_{l}$ and $T_{2}$ do not explain $D$ equally well or $\delta \neq 0$.
- Thus the hypotheses for the test are:
$H_{0}: E[\delta]=0$
$H_{1}: E[\delta] \neq 0$
Where $E[x]$ corresponds to the expected value of the random variable $x$.
- In order to test these hypotheses it is necessary to calculate how extreme the observed value of $\delta$ is with respect to the distribution of $\delta$.
- In order to calculate this distribution a bootstrapping procedure is followed to calculate multiple replicated datasets from $D$.
- The likelihoods of the trees $T_{1}$ and $T_{2}$ are then recalculated for each bootstrapped dataset.
- For each of these likelihoods a corresponding $\delta$ is calculated.
- This provides a distribution for $\delta$ against which the position of the initial $\delta$ can be compared via a two-tailed test. A two tailed test is used as there is no a priori expectation of which tree is to be preferred (Goldman et al. 2000).

The test assumes that all columns within the dataset are independent and identically distributed according the evolutionary history of the taxa under consideration.

The KH test is used to compare trees, which are selected a priori as possible best explanations for a given dataset. To examine how well a given best estimated tree matches an underlying dataset relative to another tree the SOWH (Swofford-Olsen-Wadell-Hillis) test can be used in conjunction with the KH test. This test is an example of parametric bootstrapping (Goldman et al. 2000). Essentially the test involves the construction of a tree with a given quality criterion over a given dataset. Then this initial tree is used to create multiple datasets, simulated using the parameters that define the tree. Each of these datasets is then used to create a new tree. The likelihoods of these new trees can then be compared to the likelihood of the initial tree relative to the simulated datasets. This creates a set of likelihood differences, the distribution of which can be compared to the difference between the initial tree and other trees of interest. This test is not widely used due the computational demands of the construction of multiple trees and the construction of simulated datasets.

### 2.1.10 Phylogenetic analysis using gene presence

Leaving aside sequence data there are other aspects to the genomes of a set of organisms that also provide signal indicating their evolutionary divergence. One of these aspects is proteome content, i.e. which genomic features are present in a genome and which are absent (Snel et al. 1999). This aspect is essentially the phylogenetic profile of the genomic feature. The presence and absence of the genomic feature can be treated as a binary trait and then used as an input to apply standard phylogenetic tree estimation procedures. Methods for the estimation of trees from discrete traits precede the methods described above for the analysis of molecular sequence data. The issues and criteria for estimation procedures surrounding the tree estimation from this form of data remain the same, as the underlying process for the generation of the data is the same.

As an illustrative example consider 5 species with six genes. Any of these genes can either be present or absent. An absence of a gene is coded as 0 and the presence of a gene is coded as 1.

Thus given the following distribution of the genes over the 5 species named [A-E].
A 111010
B 111111
G 000001
D 111100
E 000000
A phylogenetic tree can be estimated from this data clustering these species. An example tree reconstructed using Dollo parsimony as implemented in the PHYLIP package in the program DOLLOP (Felsenstein 1989) is shown in the figure below.


Figure 2.5: Dollo parsimonious estimation of a phylogenetic tree from example data.
As species E is devoid of all six genes it is placed as an outgroup relative to the other 5 .

### 2.2 Methods

In order to carry out a comparative study on human protein function a phylogeny was constructed to provide a framework on which to evaluate the distribution of human proteins over the eukaryotic kingdom. Sets of phylogenetic profiles, which detail this distribution, were also generated.

### 2.2.1 Data Selection

Given the computational nature of this project as well as the abundance of molecular data it was clear that its use was to be preferred over morphological data. Also given the wide morphological divergence of eukaryotes isolating individual features to be compared was not considered a plausible option.

The next point of consideration was whether to utilise nucleotide or amino acid molecular data. It is known that over long periods of evolution it is more likely for nucleotide data to become saturated with multiple back substitutions as nucleotide data has four potential changes of state at any given site as opposed to 20 potential changes in amino acid data. This can lead to an underestimate of genetic distance (Harrison and Langdale 2006; Salemi and Vandamme 2003). Thus despite nucleotide data outperforming amino acid data over smaller time frames such as the period of time covering the divergence of the division Angiospermae (Simmons et al. 2002) and of the sub phylum Vertebrata (Townsend et al. 2008) it was felt that amino acid data was a more appropriate choice as a measure of genetic distance over all eukaryotes.

### 2.2.2 Data Acquisition

Having decided to utilise amino acid data the next step was to acquire usable data. The protein sets of 54 eukaryotic genomes were downloaded on the 16th and 17th of August 2007 of which 41 organisms were accessed from the NCBI RefSeq database (using the Entrez data retrieval interface at http://www.ncbi.nlm.nih.gov/sites/gquery) (Pruitt et al. 2005) and the remainder from the Sanger Centre (ftp://ftp.sanger.ac.uk), Genoscope (http://www.genoscope.cns.fr/spip/spip.php?lang=en), TIGR ((now the JCVI )
http://www.jcvi.org/), the Broad Institute (http://www.broadinstitute.org), Ensembl (using BioMart at http://www.ensembl.org/biomart) and lastly SilkDb (http://silkworm.genomics.org.cn) (Wang et al. 2005b). These databases were employed, as they were (at the time of access) the sources utilised by the KEGG database (Kanehisa et al. 2006). An additional archeon Methanosarcina acetivorans was downloaded from the NCBI RefSeq database (Pruitt et al. 2005) in order to root the phylogeny using the outgroup criterion. This method entails using an organism that falls outside the known taxonomy of the group under consideration to provide a point of reference for the overall topology of the tree (Felsenstein 2004). It is widely accepted after work by Carl Woese (Woese et al. 1990) that the archea form a sister group to the Eucarya. Thus it was felt that an archeon was an
appropriate choice for an outgroup species. Full details of all data sources and species can be seen in the table below.

| Organism | Database | Common Name |
| :---: | :---: | :---: |
| Methanosarcina acetivorans (Outgroup) | RefSeq | NA |
| Anopheles gambiae | Ensembl | Mosquito |
| Arabidopsis thaliana | RefSeq | Thale Cress |
| Ashbya gossypii | RefSeq | NA |
| Aspergillus fumigatus | RefSeq | NA |
| Aspergillus niger | RefSeq | NA |
| Bombyx mori | SilkDB | Silkworm |
| Bos Taurus | RefSeq | Cow |
| Caenorhabditis briggsae | Ensembl | NA |
| Caenorhabditis elegans | RefSeq | NA |
| Candida albicans | RefSeq | NA |
| Candida glabrata | RefSeq | NA |
| Canis familiaris | RefSeq | Dog |
| Ciona intestinalis | RefSeq | Sea squirt |
| Cryptococcus neoformans | RefSeq | NA |
| Cryptosporidium hominis | RefSeq | NA |
| Cryptosporidium parvum | RefSeq | NA |
| Danio rerio | RefSeq | Zebrafish |
| Debaryomyces hansenii | RefSeq | NA |
| Dictyostelium discoideum | RefSeq | NA |
| Drosophila melanogaster | RefSeq | Fruitfly |
| Drosophila pseudoobscura | RefSeq | NA |
| Encephalitozoon cuniculi | RefSeq | NA |
| Entamoeba histolytica | RefSeq | NA |
| Gallus gallus | RefSeq | Chicken |
| Homo sapiens | RefSeq | Human |
| Kluyveromyces lactis | RefSeq | NA |
| Leishmania major | Sanger | NA |
| Macaca mulatta | RefSeq | Rhesus macaque |
| Magnaporthe grisea | RefSeq | NA |

Table 2.1: Organisms in study.

| Monodelphis domestica | RefSeq | Grey short-tailed oppossum |
| :---: | :---: | :---: |
| Mus musculus | RefSeq | Mouse |
| Neurospora crassa | Broad Institute | NA |
| Oryza sativa | RefSeq | Rice |
| Ostreococcus lucimarinus | RefSeq | NA |
| Pan troglodytes | RefSeq | Chimpanzee |
| Paramecium tetraurelia | Genoscope | NA |
| Pichia stipitis | RefSeq | NA |
| Plasmodium falciparum | RefSeq | NA |
| Plasmodium knowlesi | Sanger | NA |
| Plasmodium yoelii | TIGR | NA |
| Populus trichocarpa | JGI | Black cottonwood tree |
| Rattus norvegicus | RefSeq | Rat |
| Saccharomyces cerevisiae | RefSeq | Brewers yeast |
| Schizosaccharomyces pombe | RefSeq | Fission yeast |
| Strongylocentrotus purpuratus | RefSeq | NA |
| Takifugu rubripes | Ensembl | Pufferfish |
| Tetrahymena thermophila | RefSeq | NA |
| Theileria annulata | RefSeq | NA |
| Theileria parva | RefSeq | NA |
| Trichomonas vaginalis | RefSeq | NA |
| Trypanosoma brucei | RefSeq | NA |
| Trypanosoma cruzi | RefSeq | NA |
| Ustilago maydis | RefSeq | NA |
| Yarrowia lipolytica | RefSeq | NA |

Table 2.1: Organisms in study (cont).

### 2.2.3 Pairwise Alignment

In order to gauge the relatedness of individual proteins in the organisms it was necessary to use a pairwise alignment algorithm, which would deliver a measure of similarity between any two given sequences. The Smith-Waterman algorithm (Smith and Waterman 1981) was selected as it is guaranteed to locate optimal regions of local similarity. Speed is an issue with use of the Smith-Waterman algorithm however an accelerated implementation developed by Michael Farrar provided within the Fasta package made its use feasible (Farrar 2007; Pearson and Lipman 1988).

A necessary pre-processing step was to subject all sequences to low complexity filtering to remove regions of the sequence which are non random but not biologically significant such as regions of compositional bias. Thus all sequences were fed into the SEG program (Wootton and Federhen 1993) with the parameter - x , which masks out regions of low complexity sequence and replaces them with the character lower case $x$.

Each protein set was then split into its individual proteins and each protein compared against every other organism in the dataset in order to locate sequences that were significantly similar. Each comparison was run with a gap-opening penalty of -12 and a gap extension penalty of -2 . The substitution matrix BLOSUM62 (Henikoff and Henikoff 1992) was used to score the alignments. The results of these searches were then parsed and pertinent data, i.e. raw Smith-Waterman score, E value, bit score and the coordinates of the alignments along the sequences were stored in a relational database structure in MySQL to facilitate further analysis.

### 2.2.4 Orthology Determination

In order to select data which would allow the measurement of evolutionary divergence over the species it was necessary to cluster the proteins into orthologous clusters. As the Inparanoid procedure (Remm et al. 2001) had been observed to perform well in this function it was decided to utilise this procedure. The Inparanoid procedure as described in work published by Remm (Remm et al. 2001) is detailed below.

### 2.2.4.1 Inparanoid

Given a set of $n$ pairwise alignments between organism $A$ and organism $B$ the Inparanoid algorithm returns a set of clusters $s$ using sequence similarity as an inverse distance. A pairwise alignment of two proteins protein $a$ and protein $b$ in this case is a composite input consisting of

- Bit score $a S b$ : This is the result of the normalisation of a raw pairwise alignment with respect to the scoring system (Karlin and Altschul 1990). Normalisation places all scores on the same scale, which is a fundamental prerequisite for use as a distance metric.
- Sequence Lengths $a x$ and $b x$.
- Alignment lengths: The length of the alignments along both proteins alength and blength.

The latter two length inputs are used to eliminate short hits using a minimum length cut-off. These short hits may reflect functionally homologous (potentially orthologous) domains as opposed to whole proteins, which are inherited intact as a discrete unit from the last common ancestor of the two species under consideration. The bit score is employed as a cut-off to limit the radius of the clustering step. The Inparanoid algorithm runs the following steps in a pairwise comparison (Remm et al. 2001).

- Sort all bit scores in ascending order.
- Read in all hits excluding those that fall below score and length cutoffs.
- Select the best scores between organism $A$ and $B$.
- For each best-hit protein from $A-B$ examine the reciprocal relationship $B-A$.
- All reciprocal best hits are stored as a set of seed pairs for orthologous clusters.
- For each seed pair paralogous genes are grouped around them if the largest score that the putative paralog has in the set of all scores is against the putative ortholog in the seed cluster of the organism under consideration.
- Overlapping clusters are then resolved through either deletion or subsumption depending on the degree and topology of the overlap.


### 2.2.4.2 Implementation

An implementation of the Inparanoid algorithm in the Perl language (Remm et al. 2001) was acquired from the Inparanoid website (http://inparanoid.sbc.su.se). However as this implementation provided by proved to not be amenable to the analysis of bespoke output, it was decided to re-implement the procedure described above.

The Inparanoid algorithm was implemented through the application of object orientated (OO) software design principles. OO principles involve the characterisation of a problem domain as a collection of interacting objects where functionality inherent to each
object is implemented as internal to that particular object (Pressman 2001). Objects within a problem domain are generally identified through the identification of nouns within a problem statement (Pressman 2001). Perusal of the algorithm specification provided in (Remm et al. 2001) led to the above design.

### 2.2.4.2.1 Design

The main object that the procedure required in order to operate was identified as a Cluster object, which was implemented with the following attributes and operations.

| Cluster |
| :--- |
| theAGenes : ArrayList |
| theBGenes : ArrayList |
| equals( ) |
| merge(Cluster) |
| Intersects(Cluster) |
| addGroupAMember0 |
| addGroupBMember0 |
| removeGroupAMember0 |
| removeGroupBMember0 |

Figure 2.6: Class diagram of the main entity.

The box that represents the object above is split into three segments. The first segment contains the name of the object. The second segment contains the attributes that the object contains. In the case of the Cluster object it has two attributes. One of these attributes theAGenes is a list of genes/proteins clustered together from species A. Conversely theBGenes is a list of genes/proteins clustered together from species B.

The third segment of the boxes represents the operations that it is possible for the object to carry out. In the case of the Cluster object the main operations necessary are the ability to add and delete members as well as to detect whether two Clusters overlap. Finally the ability to merge two Clusters is an essential operation.

It was decided to use the programming language Java to implement the above design as the language provides functionality, which facilitates the OO paradigm. The implementation was then tested against the author provided Perl implementation to ensure correctness.

The Java implementation deviated slightly from the author provided Perl implementation in two respects. The first deviation from the author provided Perl implementation was the use of higher precision double values to represent the bit scores of alignments as opposed to the use of integers by Perl. This led to cases where scores that had been rounded in the Perl implementation and (thus marked as reciprocal best hits) were not marked as reciprocal best hits (examples in Appendix A).

The second deviation was to cluster orthologous genes with two equal reciprocal best hits between the species at the stage of sorting. This change in the order of steps has no effect overall on the groups produced by the implementation. The implementation was run using a bit score cut-off of 50 and an alignment length cut-off, which was $50 \%$ of the length of the longer protein.

### 2.2.4.3 Application

The data generated from the similarity searches was then clustered to identify orthologous genes using the constructed Inparanoid implementation. The study was carried out using $H$. sapiens as a reference species. Orthologous groups are sought in each organism for every protein within the human proteome. This study can be considered unbalanced as no information was collected on proteins that are absent in H. sapiens (Davey et al. 2007).

As the dataset under consideration was amino acid sequence data there was a choice as to how to deal with alternatively spliced isoforms of the same protein. As the goal of this project was to examine the presence and absence of the proteins under consideration it was decided that the retention of all isoforms in the dataset and clustering them as inparalogs was appropriate. This would allow an examination of correlations in gain and loss of the protein as an independent phenotypic entity.

### 2.2.5 Phylogenetic profiles

The output from the clustering step was then used to generate phylogenetic profiles, which as mentioned previously are binary strings of presence and absence of an orthologous group for a given gene in the reference species.

In order to establish create phylogenetic profiles of each protein within the human proteome the following steps were undertaken.

- A list of each GI identifier for the set of human proteins was generated.
- For each entry in this list the relevant identifier was scanned against all files containing orthology predictions in alphabetical order.
- If the entry was present within the orthology prediction file for a given organism that position within the profile string was marked as 1 . Otherwise that position was marked as 0 . The order of the profiles is alphabetical. Therefore for example the profile 100000000000000000000000100000000000000000000000000000 indicates a protein with an orthologous group present in Anopheles gambiae and Homo sapiens but absent in all other species under consideration.
- Two sets of profiles were generated, one including the outgroup for use in ortholog selection and the other excluding the outgroup for use in prediction of functional linkage.

Table 2.2 lists the organisms under study along with their proteome sizes and number of orthologous groups with reference to H. sapiens. Figure 2.7 shows the distribution of proteome sizes in the organisms under consideration. Figure 2.8 shows the number of proteins clustered in each organism.

| Organism | No. of Proteins No. of Clusters |  | No. of Proteins Clustered |
| :---: | :---: | :---: | :---: |
| Methanosarcina acetivorans | 4544 | 408 | 822 |
| Anopheles gambiae | 13465 | 4889 | 5030 |
| Arabidopsis thaliana | 31915 | 2849 | 3389 |
| Ashbya gossypii | 4292 | 1590 | 1599 |
| Aspergillus fumigatus | 9630 | 2167 | 2175 |
| Aspergillus niger | 14102 | 2197 | 2210 |
| Bombyx mori | 21302 | 4060 | 4104 |
| Bos Taurus | 25379 | 15508 | 16567 |
| Caenorhabditis briggsae | 19553 | 4009 | 4212 |
| Caenorhabditis elegans | 23220 | 4149 | 4404 |
| Candida albicans | 14107 | 1909 | 3367 |
| Candida glabrata | 5192 | 1737 | 1784 |
| Canis familiaris | 33527 | 16372 | 17681 |
| Ciona intestinalis | 15852 | 6097 | 6117 |
| Cryptococcus neoformans | 6594 | 1985 | 2045 |
| Cryptosporidium hominis | 3886 | 953 | 957 |
| Cryptosporidium parvum | 3805 | 1080 | 1086 |
| Danio rerio | 36083 | 11598 | 13007 |
| Debaryomyces hansenii | 6317 | 1861 | 1887 |
| Dictyostelium discoideum | 13377 | 2551 | 2650 |
| Drosophila melanogaster | 20071 | 5037 | 6847 |
| Drosophila pseudoobscura | 9871 | 4230 | 4240 |
| Encephalitozoon cuniculi | 1996 | 711 | 725 |
| Entamoeba histolytica | 9772 | 1282 | 1572 |
| Gallus gallus | 18710 | 11489 | 11769 |
| Homo sapiens | 33473 | 26634 | 33473 |
| Kluyveromyces lactis | 5339 | 1745 | 1751 |
| Leishmania major | 8302 | 1494 | 1706 |
| Macaca mulatta | 37856 | 17025 | 22774 |
| Magnaporthe grisea | 14010 | 2099 | 4420 |
| Monodelphis domestica | 20194 | 13633 | 13880 |
| Mus musculus | 35048 | 16607 | 18448 |
| Neurospora crassa | 10082 | 1998 | 2002 |
| Oryza sativa | 26887 | 2699 | 2753 |
| Ostreococcus lucimarinus | 7603 | 2152 | 2241 |
| Pan troglodytes | 51517 | 18916 | 28944 |
| Paramecium tetraurelia | 39642 | 2147 | 2740 |
| Pichia stipitis | 5816 | 1883 | 3826 |
| Plasmodium falciparum | 5270 | 1090 | 2198 |
| Plasmodium knowlesi | 4958 | 1125 | 2254 |
| Plasmodium yoelii | 7861 | 1058 | 2140 |
| Populus trichocarpa | 45555 | 3037 | 3363 |
| Rattus norvegicus | 35903 | 16414 | 21279 |
| Saccharomyces cerevisiae | 5883 | 1803 | 3706 |
| Schizosaccharomyces pombe | 5045 | 2066 | 4272 |

Table 2.2: List of organisms used in study along with data source and proteome size as well as number of orthologous groups in alphabetical order. The outgroup is placed at the top.

| Strongylocentrotus | 42373 | 6638 | 13387 |
| :--- | :--- | :--- | :--- |
| Takifugu rubripes | 22428 | 10891 | 11046 |
| Tetrahymena thermophila | 26235 | 2014 | 2051 |
| Theileria annulata | 3795 | 1000 | 1016 |
| Theileria parva | 4079 | 1005 | 2044 |
| Trichomonas vaginalis | 59681 | 1823 | 2567 |
| Trypanosoma brucei | 8772 | 1531 | 3596 |
| Trypanosoma cruzi | 19606 | 1737 | 2079 |
| Ustilago maydis | 6548 | 3756 | 3762 |
| Yarrowia lipolytica | 6545 | 4092 | 4132 |

Table 2.2: List of organisms used in study along with data source and proteome size as well as number of orthologous groups in alphabetical order (cont).


Figure 2.7: Distribution of proteome sizes in organisms under consideration. $\mathrm{N}=55$.


Figure 2.8: Distribution of number of proteins placed within clusters in each organism. $\mathrm{N}=55$.

Table 2.3 shows the top ten profiles within the human genome and provides an interpretation.

| Profile | Count | Interpretation |
| :---: | :---: | :---: |
| 000000000000000000000000100000000000000000000000000000 | 5150 | Present in species Homo sapiens. |
| 000000000000000000000000100000000010000000000000000000 | 2281 | Present in Tribe <br> Hominini. |
| 000000100001000000000001100101100010000001000000000000 | 1089 | Present in Phylum Chordata. |
| 000000000000000000000000100100000010000000000000000000 | 622 | Present in Order <br> Primate. |
| 000000100001000000000000100101100010000001000000000000 | 495 | Present in Infraclass Eutheria. |
| 000000000000000000000000100100000000000000000000000000 | 466 | Present in species <br> Homo sapiens and <br> Macaca mulatta. |
| 000000100001000000000001100101100010000001000000000000 | 378 | Present in Class |
| 000000100001000010000001100101100010000001000000000000 |  | Mammalia and Class Aves |
| 000000100001000000000000100100100010000001000000000000 | 342 | Present in Class <br> Mammalia |
| 000000100001000010000001100101100010000001000000000000 | 281 | Present in the Phylum Chordata with the exception of the Species Takifugu rubripes |
| 000000100001100010000001100101100010000001000100000000 | 235 | Present in in Class <br> Mammalia ,Class <br> Actinopterygii and <br> Class Aves |

Table 2.3: Top ten occurring phylogenetic profiles ranked by counts.

### 2.2.5.1 Single copy proteins

In order to select orthologous proteins, which most accurately reflected the evolutionary histories of the species under consideration, it was decided to focus on orthologous groups that were present in a single copy across all 55 taxa. This focus on single copy proteins excluded potential comparisons between paralogous proteins.

A set of proteins with ubiquitous profiles was extracted from the data and these were sifted for proteins that were present in single copy in each organism in the dataset. 10 proteins were present in single copy over all the organisms under study. Table 2.4 shows details of these proteins.

| NCBI RefSeq GI Number | Description in NCBI annotation. Entrez Gene Name |  |
| :---: | :---: | :---: |
| 116805340 | glycyl-tRNA synthetase | GARS |
| 32307132 | NFS1 nitrogen fixation 1 precursor | or NFS1 |
| 4506605 | ribosomal protein L23 | RPL23 |
| 4506743 | ribosomal protein S8 | RPS8 |
| 4507215 | signal recognition particle 54 kDa isoform 1 | SRP54 |
|  | excision repair crosscomplementing rodent repair deficiency,complementation group | ERCC3 |
| 4557563 | 3 |  |
| 5031815 | lysyl-tRNA synthetase isoform 2 | KARS |
|  | $\mathrm{H}(+)$-transporting two-sector | ATP6V1D |
| 7706757 | ATPase |  |
| 5803092 | Methioine aminopeptidase 2 | METAP2 |
| 24430151 | 26s protease regulatory subunit 4 | PSMC1 |

Table 2.4: Single copy ubiquitous genes extracted via analysis of profiles.

### 2.2.5.2 Proteome content data/tree

The phylogenetic profiles as developed provided a matrix of presence and absence of every human protein in the dataset across the remaining 53 eukaryotes and 1 outgroup archeon. As this form of data also contains phylogenetic signal, i.e. shows the divergence in the proteomes of the given species over time, it was decided to subject this data to a phylogenetic analysis as well as the main phylogenetic analysis to be carried out on the multiple sequence alignments of the homologous proteins. The tree was reconstructed using Dollo parsimony (Farris 1977) via the program DOLLOP (Felsenstein 1989). This tree can be seen in Figure 2.17.

In order to carry this out the profiles were transposed, so that rather than showing the distribution of human proteins over a set of species they showed the pattern of presence and absence of human proteins over a single species. Thus instead of a matrix of 33,473 proteins by 55 species the end product was a matrix of 55 species by 33,473 proteins. In other words each species was assigned a binary string of length 34,373 where 1 indicated the presence of a particular human protein and 0 its absence.

This matrix was converted into PHYLIP format through the truncation of species names to 10 characters and the addition of header information about the size of the matrix. Finally this formatted file was input to the DOLLOP program (Felsenstein 1989). The program was run with its default settings.

In order to examine the level of support for the initial outputted tree 100 bootstrap replicates were created with SEQBOOT (Felsenstein 1989). These 100 replicates were resubmitted to DOLLOP to produce 100 bootstrap trees. These trees were unified using CONSENSE (Felsenstein 1989).

### 2.2.6 Multiple alignment

As an initial step to generate a phylogenetic tree using the orthologous proteins selected a multiple alignment of each of the 10 proteins was constructed utilising Mafft (Katoh et al. 2002) (Multiple Alignment By Fast Fourier Transform) using the L-INS-i algorithm for 1000 iterations. Each alignment was then subjected to Gblocks filtration (Talavera and Castresana 2007) to remove columns that were poorly aligned. Gblocks was run in its relaxed mode. These alignments were then concatenated to form a super matrix measure of divergence in
order to generate a measure of divergence across the genomes as opposed to at a single locus, as has been suggested by (Rokas et al. 2003). The full alignment can be seen in Appendix D.

### 2.2.7 Model selection

In order to select an evolutionary model which provided a statistically accurate measure of genetic divergence ModelGenerator (Keane et al. 2006) was used to select the model that best fitted the concatenated multiple alignment. It requires as an argument a number for gamma categories, to account for heterogeneity in substitution rates. The argument was given a value of 4 gamma categories as this has been observed to be sufficient number to create a "nearoptimum fit" of a model (Yang 1994).

It has been observed that individual gene trees can be highly incongruent with species trees (Cranston et al. 2009). Thus it is possible for the inference of a species tree to be misled by "non-phylogenetic signal" from the individual genes (Cranston et al. 2009). In order to examine potential incongruence between gene alignments each individual alignment was first analysed separately. The model selected for the complete supermatrix was the LG substitution matrix (Le and Gascuel 2008). The LG matrix was predicted with the additional parameter $\Gamma$ that allows different rates of evolution across the sequence. Each gene alignment was also matched by the LG matrix along with different variations as shown in Table 2.5. The LG matrix is generated by a model of evolution that takes into account mutation rate heterogeneity over sites, thus yielding better results then its predecessors WAG and JTT (Le and Gascuel 2008). The models selected were a best fit judged by both the Aikake Information Criterion (AIC) as well as the Bayesian Information Criterion (BIC) in all but two cases (GARS and ERCC3). Where they disagreed the BIC was selected over the AIC (Yang 2008) thus lowering the possibility of overfitting a more complex model to the data.

The models selected for the individual orthologs were exactly the same as the model for the concatenated alignment except in the case of SRP54 where the additional parameter I indicating that a proportion of the alignment was invariant.

| Entrez Gene Name | Substitution model of best fit (Selected by <br> GIC |
| :--- | :--- |
| GARS | LG $+\Gamma$ |
| NFS1 | LG $+\Gamma$ |
| ATP6V1D | $\mathrm{LG}+\Gamma$ |
| KARS | $\mathrm{LG}+\mathrm{I}+\Gamma$ |
| SRP54 | $\mathrm{LG}+\mathrm{I}+\Gamma$ |
| PSMC1 | $\mathrm{LG}+\Gamma$ |
| METAP2 | $\mathrm{LG}+\Gamma$ |
| ERCC3 | $\mathrm{LG}+\Gamma$ |
| RPL23 | $\mathrm{LG}+\Gamma$ |
| RPS8 | $\mathrm{LG}+\Gamma$ |

Table 2.5: Substitution model selected by ModelGenerator for each ortholog.

### 2.2.8 Phylogeny reconstruction

After selection of the model each alignment was individually input to PhyML (Guindon and Gascuel 2003) with its individual substitution model as well as the concatenated alignment. A total of 11 trees were generated one per orthologous gene and one for the concatenated alignment as a whole. For the concatenated alignment 1000 replicates were also generated using SEQBOOT (Felsenstein 1989) and rerun using PhyML (Guindon and Gascuel 2003). These 1000 replicated trees were input to CONSENSE (Felsenstein 1989) in order to acquire an estimate of the overall bootstrap support for the tree from within the data. The topology of the bootstrapped trees was identical to the topology of the primary ML tree.

### 2.2.9 Comparison of protein content tree with super matrix tree

In order to place a measure on whether the protein content tree was a significantly worse hypothesis of the evolutionary relationships of the species under study the PHYLIP program PROML (Felsenstein 1989) was utilised. PROML was given the supermatrix alignment as a dataset and the two trees as user inputted trees to evaluate against the dataset. PROML then ran the KH test against the two trees to examine the differences in the likelihood of the trees relative to the dataset.

### 2.3 Results

The individual gene trees can be seen in Appendix B. In order to examine potential differences in phylogenetic signal between the individual genes trees TREEDIST (Felsenstein 1989) was used to generate a distance matrix between the 10 gene trees. TREEDIST uses the Branch Score distance (Kuhner and Felsenstein 1994) to calculate the distance between two trees. This distance takes into account the branch lengths of the trees input as well as the overall topology. This matrix was then used to generate a dendogram using UPGMA (Unweighted Pair Group Method with Arithmetic mean) clustering as shown in Figure 2.9.


Fig 2.9: Cluster diagram of tree distances of individual gene phylogenies.

The gene trees fell into three main clusters. In order to examine the degree of incongruence of each cluster from the super matrix species tree the trees contained with each cluster were then submitted to CONSENSE (Felsenstein 1989) in order to view the consensus trees. CONSENSE was run with the majority rule setting where a group has to appear more than $50 \%$ of the time in the input trees in order to be conserved in the consensus tree. Figures 2.10 shows the outlier tree estimated from ERCC3 and Figures 2.11 and 2.12 show the consensus trees.


Figure 2.10: Gene tree for gene ERCC3.


Figure 2.11: Consensus Tree for Cluster 2 containing genes: RSP8 ATP6V1D PSMC1 and METAP2.


Figure 2.12: Consensus Tree 2 for Cluster 3 containing genes: GARS, NFS1, RPL23, SRP54 and KARS.

Both consensus trees preserve the kingdoms of Plantae, Animalia and Fungi though the order of branching is lost. However both clusters demonstrate a broad congruence with the fully concatenated ML tree, which can be seen in Figure 2.13. This is a useful measure of the degree of overlap of phylogenetic signal contributed from each of the individual genes. Figure 2.13 is an illustration of the topology of the tree with the animals, fungi and plants highlighted. Figure 2.14 also overlays the bootstrap support values for each proposed clade. Figure 2.15 presents the reconstructed phylogeny with a measure of support, which is the proportion of the individual gene trees that supported a clade (Bratke 2009). Figure 2.16 shows the topology of the tree in combination with branch lengths.


Figure 2.13: ML tree of 54 eukaryotes without branch lengths created from a super matrix of the concatenated alignments of all genes listed in Table 2.5. The clades containing animals, fungi and plants are coloured blue, red and green respectively.


Figure 2.14: ML tree of 54 eukaryotes without branch lengths created from a super matrix of the concatenated alignments of all genes listed in Table 2.5. Bootstrap support values are only shown at each node where support was less than 1000 (not universally supported across 1000 bootstrap replicates).


Figure 2.15: ML tree of 54 eukaryotes created from the concatenated alignments of genes listed in Table 2.5. Support Values are proportion of individual gene trees, which show a given clade (Bratke 2009).


Fig 2.16: ML tree of 54 eukaryotes with proportional branch lengths.


Figure 2.17: Proteome content phylogeny with bootstrap support only shown at each node which was not $100 \%$ supported out of 100 bootstraps.

## Tree Comparison

The results of the KH test carried out using PROML (Felsenstein 1989) showed that the proteome content tree is a significantly worse hypothesis of the evolutionary relationships between the organisms as shown in the table below.

| Tree | Log likelihood |
| :--- | :---: |
| Protein alignment supermatrix ML tree | -131725.8 |
| Proteome content parsimony tree | -133941.0 |

PROML (Felsenstein 1989) reported that the log likelihood of the proteome content tree was significantly worse than that of the protein supermatrix tree.

### 2.4 Discussion

### 2.4.1.ML tree

In terms of current thought about super groups within eukaryotes the phylogeny reconstructed as seen in Figure 2.13 is incongruent. However given that it is based on a concatenation of nuclear genes this is not surprising (Parfrey et al. 2006). This work showed that there is generally weak support for most putative eukaryotic super groups in phylogenies built using proteins coded by nuclear genes. The super group Opisthokonta is however supported as would be expected from the work (Parfrey et al. 2006) though it does subsume the Amoebozoa. The ML tree is consistent with known eukaryotic trees (Baldauf et al. 2000) in placing plantae as an outgroup to fungi and metazoans. The base of the tree is inconsistent with known trees in its placement of $E$. histolytica as an early braching eukaryote with $T$. vaginalis when it is thought that $E$. histolytica branches higher up in the tree as a member of the Amoebozoa super group (Parfrey et al. 2006). However there is no clear synapomorphy (shared derived character) which defines the group Amoebozoa. There is also a lack of unambiguous support for the existence of the group as a whole within the nuclear genome (Parfrey et al. 2006). This placement is also not novel as both organisms lack mitochondria and have been grouped together at the base of eukaryota by phylogenetic analyses of small subunit (SSU) RNA genes though their ultimate placement is not certain (Vanacova et al. 2003).

Within the animals the tree is consistent with the Coelomata hypothesis which places the nematoda as an outgroup to both arthropods and vertebrates. This grouping is fairly common in phylogenies derived using molecular data (Wolf et al. 2004) despite being held to be false (Aguinaldo et al. 1997). This is thought to be an artefact of long branch attraction due to rapid evolution along the C. elegans line (Telford 2004). The classes Mammalia (H. sapiens, P. troglodytes, B.taurus, M. domestica, C. familiaris, M. musculus and R. norvegicus), Aves (G. gallus), Osteichthyes (T. rubripes, D. rerio) are all maintained in the order that they are generally found in most broad vertebrate phylogenies, e.g. work by Stuart (Stuart et al. 2002).

The tree also arranges its four plant species as expected (Rodriguez-Ezpeleta et al. 2005) with the algae $O$. lucimarinus forming an outgroup to the monocot $O$. sativa and the two dicots $A$. thaliana and $P$. trichocarpa.

Within the fungi the tree is consistent with known fungal phylogenies (Fitzpatrick et al. 2006). The kingdom Dikarya is a separate clade. Within Dikarya in the phylum Ascomycota the subphyla Saccharomycotina (S. cerevisiae, C. albicans, C. glabrata, $P$ .stipitis, D. hanseii, Y. lipolytica, K. lactis , A. gossypii), Taphrinomycotina (S. pombe) and Pezizomycotina (A. fumigatus, A. niger, N. crassa, M. grisea) are grouped as separate clades. Another phylum in Dikarya Basidiomycota (U. maydis, C. neoformans) is a separate clade within the tree. The microsporidium E. cuniculi branches out as an outgroup to the Dikarya. Within Saccharomycotina the WGD (fungi which have undergone whole genome duplication) (C. glabrata, S. cerivisae) are presented as a clade. Also the CTG group (fungi which utilise the codon CTG to encode serine instead of leucine) (P. stipitis, D. hanseii, C. albicans) (Fitzpatrick et al. 2006) is proposed as a separate clade within the tree. Within the CTG group the tree shows disagreement with some published trees (Wang et al. 2009a) by placing $P$. stipitis and $C$. albicans together with $D$. hanseii as an outgroup. Figure 2.6 shows that this fungal topology was highly supported by the bootstrap analysis.

The Chromoalveolates are grouped together in one clade. Within this clade the Apicomplexa form a monophyletic group within that clade with the Ciliates as an outgroup. These groupings are congruent with published trees (Burki et al. 2008; Rodriguez-Ezpeleta et al. 2007).

### 2.4.2 Proteome content phylogeny

The proteome content phylogenetic tree as presented in Figure 2.17 shows a degree of topological congruence with the tree based on concatenated protein sequences shown in

Figure 2.13 in that it preserves the animals as a monophyletic group. The branching order of the taxa is however different. It also shows the fungi as a clustered group (though not monophyletic). However given that PROML reports that it is a significantly worse fit to the alignment of homologous proteins it is clearly a worse representation of the dataset then the ML supermatrix tree.

### 2.4.3 Conclusion

The reconstructed phylogeny via an application of maximum likelihood to the concatenated supermatrix of 10 eukaryotic proteins appeared to be a plausibly accurate reflection of the relationships between the taxa. This plausibility was assessed by both by inspection by eye and comparisons to previously published eukaryotic phylogenies.

The proteome content phylogeny reconstructed by Dollo parsimony on the other hand was shown to be a significantly worse representation of the evolutionary relationships between the species. As such it was the ML supermatrix tree that was utilised as the framework for comparative analysis of protein function within the species.
The work described in this chapter also produced phylogenetic profiles for each human protein across 54 other eukaryotes.

## Chapter 3

## Comparison of methods of prediction of functional linkage in proteins

### 3.1 Introduction

This chapter presents the comparison of four systems of inferring functional links between proteins using phylogenetic profiles. These systems were:

1) Hamming distance: Phylogenetic profiling was initially used without taking into account species phylogeny and treating the state of each point in the profile as independent (Pellegrini et al. 1999). Using profiling in this way entailed comparison of profiles using the string comparison algorithm Hamming distance (Hamming 1950), which is a count of the points at which two strings differ.
2) Use of the comparative method (Barker and Pagel 2005; Pagel 1994) in the context of phylogenetic profiling (Pellegrini et al. 1999) with constrained rates of gene gain (Barker et al. 2007) over the phylogeny developed in Chapter 2 to detect protein interactions. An implementation of the method, BayesTraits (Pagel et al. 2004a) was used in order to calculate the relevant likelihoods.
3) Co-expression of mRNAs corresponding to given proteins: Proteins that physically interact or are required to be produced in some of form of spatio-temporal order tend to show correlations (positive or negative) in the expression of their underlying mRNA molecules. This method has been shown to be effective in detecting interactions in Saccharomyces cerevisiae (von Mering et al. 2002) as well as in Arabidopsis thaliana in combination with examination of other genomic features (De Bodt et al. 2009). Use of this system presents a comparison of an un-curated highthroughput physical experimental system with an equivalently un-curated computational system.
4) Use of a Bayesian classifier to combine disparate sources of evidence comprising gene co-expression, orthology, post translational modification, co-localisation, intrinsic disorder, domain co-occurrence and network analysis data in order to predict protein interactions (McDowall et al. 2009).

### 3.1.1 Hamming distance

The distance measure Hamming distance is named after its creator Richard Hamming who introduced it in his work (Hamming 1950). As mentioned above and in previous chapters, it is the distance between two strings of equal length calculated as a count of the points where
they differ. A string can be represented as a vector of characters. As an illustrative example given the two strings x and y as defined below:

$$
\begin{aligned}
& x=[c, a, t] \\
& y=[h, a, t]
\end{aligned}
$$

The hamming distance between the two strings is 1 as they vary by 1 character.

In the context of phylogenetic profile analysis, a hamming distance of 1 between two profiles would indicate that the gene/protein under consideration differed by only one species in its pattern of distribution.

### 3.1.2 Comparative method

As mentioned in the introductory chapter the comparative method involves the examination of the association of a given trait in an organism with another variable in the context of a phylogenetic tree (Harvey and Pagel 1991). This variable can be another trait or potentially an environmental factor. A prerequisite for this form of analysis is reconstructing putative ancestral states at each hypothetical ancestral node within the tree (Pagel 1994). Using this method in the context of phylogenetic profile analysis involves testing whether the presence or absence of a given gene is associated with the presence or absence of a second gene or protein.

The implementation of the comparative method used in this chapter is based on the approach introduced by Mark Pagel (Pagel 1994) and utilised in the context of phylogenetic profiling by Barker and Pagel (Barker and Pagel 2005). This approach is preferred to other applications of the comparative method (e.g. work by Wayne Maddison (Maddison 1990) and work by Mark Ridley (Ridley 1983)) due to the fact that it does not depend on a single set of reconstructed ancestral states over a tree but instead calculates test statistics based on all possible ancestral states (Barker and Pagel 2005; Pagel 1994). An explanation of how Barker and Pagel (Barker and Pagel 2005) utilised the framework established in previous work by Pagel (Pagel 1994) in order to analyse phylogenetic profiles follows.

### 3.1.2.1 Phylogenetic profile analysis using the comparative method

Imagine a gene or protein $G_{l}$, which exists in a group of species and a phylogenetic tree representing the evolutionary relationships between the members of the group. At any given internal node (hypothetical ancestor) within the tree $G_{l}$ can either be present or absent. This state will be denoted as 0 for absent and 1 for present.

Over a given branch if the state of $G_{I}$ at the ancestral node is 0 then there is a probability of a gain, i.e. moving to state 1 at the descendant node over the time period represented by the branch. Conversely if the ancestral state is 1 then there is a corresponding probability of a loss. These probabilities are represented as $P_{01}(t)$ and $P_{10}(t)$ where $t$ is equal to the time interval represented by the branch. There are also the probabilities of no transitions which are represented by $P_{00}(t)$ and $P_{11}(t)$. The probabilities $P_{01}(t)$ and $P_{10}(t)$ can also be considered the rate of transitions.

The comparative method as applied to phylogenetic profiling is an examination of whether the state of a second gene/protein has an effect on the state of the first. Thus introducing a second gene / protein $G_{2}$ there is a corresponding set of probabilities for $G_{2}$. In order to examine whether the state of $G_{2}$ has an effect on the state of $G_{1}$ the probabilities or transition rates $P_{01}(t), P_{10}(t), P_{00}(t)$ and $P_{1 l}(t)$ for $G_{l}$ can be split in order to factor in the state of $G_{22}$. Thus for example $P_{01}(t)$ can be split into two probabilities one corresponding to the rate of gain of $G_{I}$ if $G_{2}$ is present and the other corresponding to the rate of gain of $G_{1}$ if $G_{2}$ is absent. These transition rates were the basic parameters used by Barker and Pagel (Barker and Pagel 2005) as represented in the following figure.


Figure 3.1: Parameters for modelling state transitions for pairs of genes as used by Barker and Pagel (Barker and Pagel 2005) (Figure is directly reproduced).


Table 3.1: Description of rate parameters used by Barker and Pagel (Barker and Pagel 2005).
Given these rates it is possible to investigate whether the state of $G_{2}$ has an effect on the transition rates for $G_{l_{1}}$ In order to carry out this investigation two competing models / hypotheses were constructed using these parameters (Barker and Pagel 2005). One was a dependant model where the presence of $G_{l}$ is somehow contingent on the presence of the absence of $G_{2}$ or an independent model where there was no connection (Barker and Pagel 2005). The dependent model makes an assumption that the rate of gain/loss of $G_{l}$ is somehow affected by the state of $G_{2 \text {. Thus for example the rate of gain of } G_{l} \text { in the presence }}$ of $G_{2}\left(q_{24}\right)$ will be different from the rate of gain of $G_{1}$ in the absence of $G_{2}\left(q_{13}\right)$. Conversely the dependant model makes the assumption that there is no effect on the transition rates of $G_{l}$ by the state of $G_{2}$. To detect gains and losses over a phylogenetic tree Barker and Pagel reconstructed the likelihood of these two competing hypothesis about the distribution of pairs of proteins in the constituent species (Barker and Pagel 2005). The premise of the work was that the dependent model would prove a better fit to observed data if the transition rate for a given protein were affected by the state of the other. In order to detect correlated evolution the two competing models were thus defined as follows

- The independent model of evolution where the probabilities of gain and loss of $A$ were independent of the state of $B$. In order to create this model the parameters involving gain and loss of $A$ were constrained to be equal irrespective of the state $B$ and vice versa. Using the symbols to define the rates shown in Figure 3.1 this entailed setting the transition rates as $q_{13}=q_{24}, q_{42}=q_{31}, q_{31}=q_{43}$ and $q_{12}=q_{34}$. This reduced the number of parameters for the independent model to four.
- A dependant or correlated model of evolution that utilised all eight-transition rate parameters.

Thus given a set of phylogenetic profiles and a phylogenetic tree, the values of the parameters that maximised the likelihood of each of the two models was calculated in turn between pairs of profiles. These likelihoods were calculated summing the likelihoods of all possible ancestral reconstructions at each internal node of the tree thus removing the need for the reliance on a single set. Having calculated the likelihoods of both models, the goodness of fit of the models to the observed data was compared using the likelihood ratio statistic, $L R$. This can be calculated using the following equation (Yang 2006):

$$
\begin{equation*}
L R=-2\left(\ln \left(H_{0}\right)-\ln \left(H_{1}\right)\right) \tag{1}
\end{equation*}
$$

As applied to detection of correlated evolution $H_{0}$ was the likelihood of the independent model of evolution and $H_{l}$ was the model of dependent evolution (Barker et al. 2007; Barker and Pagel 2005).

Further work showed that constraining the rate of gain of a gene to a preset low level as a more potentially realistic representation of actual biological reality improved on the ability of the method to detect functional linkages in Saccharomyces cerevisiae (Barker et al. 2007). This entailed specifying values for the parameters connected to gene gain specifically $q_{31}, q_{24}, q_{34}$ and $q_{12}$. Rate of gene gain is specifically low in eukaryotes where horizontal gene transfer is relatively rare (Whitaker et al. 2009). The rate of de novo generation of genes over a relatively short time scale has also been observed to be extremely low. A study detected three potential genes in the human genome, which were generated de novo since the split with P. troglodytes (Knowles and McLysaght 2009) (estimated at around 4 million years (Hobolth et al. 2007)).

This method of fitting models of correlated and uncorrelated models of evolution for pairs of proteins using maximum likelihood (ML) (Barker and Pagel 2005) while constraining the rate of protein gain (Barker et al. 2007) shall from here on be referred to as constrained ML.

### 3.1.3 Co-expression as measured by microarray

As mentioned in Chapter 1 a microarray is a chip usually made of glass, with fluorescently labelled oligonucleotide probes representing subsections of genes. The degree of florescence from these probes corresponds to the abundance of a given mRNA in a sample and therefore the level of expression (Quackenbush 2002; Wodicka et al. 1997). In a typical microarray experiment cells are subjected to different treatments, or harvested from organisms with differing phenotypic or disease states (e.g. cancerous tissue vs. normal in human cancer patients). The probes on a given microarray chip are generally designed to map to a within a coding region on a given gene (Brown 2006).

To establish whether a given experimental condition corresponds with differential expression of a given gene, a descriptive statistic illustrating the central tendency of expression (usually the median) is calculated for the entire set of samples. If a given gene is found to be expressed at a statistically significant higher level than the central value, this gene is interpreted as up-regulated. Similarly if a gene is expressed at a significantly lower level then that gene is interpreted as down-regulated. Probes on a microarray chip can map to a single gene, or members of a gene family depending on the specificity of the probe design (Heyer et al. 1999).

### 3.1.4 Bayesian classifier

The fourth system of prediction of functional linkage to be considered was that utilised by the PIPs server (http://www.compbio.dundee.ac.uk/www-pips)(McDowall et al. 2009). This system (Scott and Barton 2007) utilised a combination of sources of evidence for protein interactions. These sources were:

- Gene co-expression: As described and used above a correlated shift in gene expression patterns in response to a given environmental stimulus can be used to detect given protein interactions.
- Orthology: If a given pair of proteins is orthologous to a pair of proteins in another species that are known to interact then this interaction annotation can be ported from one species to the other (Yu et al. 2004a).
- Subcellular localisation, domain co-occurrence, and posttranslational modification cooccurrence: These features of a protein can also be informative as to its interaction partners. The PIPs system (Scott and Barton 2007) combines these as a joint source of evidence.
- Protein disorder: This measure is based on the observation that the unstructured regions within protein molecules are often involved in transient protein interactions. (Singh et al. 2007) showed that intrinsic disorder is enriched in "date hubs", proteins that maintain multiple interactions but at different times.
- Network topology similarity: This measure utilises the principle that proteins that interact will share other interacting partners.

These five predictors are combined using a naïve Bayesian classifier to generate a single score based on the posterior odds ratio of interaction after calculation of likelihood ratios over each of the individual predictor modules (Scott and Barton 2007).

In order to explain what an odds ratio is it is necessary first to define odds. Odds are a method of presenting the probability of an event by relating this probability to the probability of the event not occurring. Thus the odds of an event are simply the probability of an event occurring divided by the probability of the event not occurring (Sokal and Rohlf 1995). The odds of a given event can be calculated by the equation:

$$
\begin{equation*}
o d d s(e)=\frac{p(e)}{1-p(e)} \tag{2}
\end{equation*}
$$

An odds ratio is thus is the ratio of multiple odds. It can be used to measure the effect size of a given factor on the probability of an event. Thus if for example the probability of heart disease in people who consume a high fat diet is calculated as $\frac{1}{4}$ and the converse probability of heart disease in individuals who do not consume a high fat diet is calculated as $\frac{1}{8}$, the odds of having heart disease with a high fat diet are thus via Equation 2 equal to 0.3 and the odds of having heart disease with a low fat diet are 0.14 . Thus the odds ratio would be calculated as $\frac{0.3}{0.14}$ which is roughly equal to 2 . Thus in this example a high fat diet roughly doubles the probability of heart disease.

The posterior odds ratio utilised by PIPs (Scott and Barton 2007) was calculated by utilising a prior odds ratio calculated by using a prior probability of interaction estimated as $\frac{1}{400}$. This prior odds ratio was then multiplied by the likelihood ratios yielded by each of the individual predictor modules. The product of this calculation is the posterior odds ratio.

As in the example, the posterior odds ratio corresponds to the posterior probability of interacting, e.g. a score of 2 translates to the probability of interacting being twice as high as the probability of not interacting (McDowall et al. 2009; Scott and Barton 2007).

### 3.2 Methods

### 3.2.1 Assessing quality

In terms of classification of the accuracy of a binary classification system a common method of measurement is the use of sensitivity and precision. Sensitivity can be defined as the probability of predicting a true positive and precision as the probability of that prediction being correct (Baldi and Brunak 2001). In order to calculate these measures some terminology must be introduced.

True positives (TP): The number of positive predictions made by a binary classifier that lie within the positive training set.

False positives (FP): The number of positive predictions made by a binary classifier that lie within the known negative training set.

False negatives (FN): The number of items in the known positive set, which were not predicted by a binary classifier.

Given these values precision and sensitivity can be calculated as follows (Baldi and Brunak 2001; Barker et al. 2007; von Mering et al. 2003):

$$
\begin{align*}
& \text { precision }=\frac{(T P)}{(T P+F P)}  \tag{3}\\
& \text { sensitivity }=\frac{(T P)}{(T P+F N)} \tag{4}
\end{align*}
$$

### 3.2.2 Training and test data.

In order to calculate these values it was necessary to acquire data on known positive interactions. Thus the data used by Scott in her development of PIPs (Scott and Barton 2007) which was in turn derived from the HPRD (Human Protein Reference Database) (Mishra et al. 2006) was acquired. This dataset contained 25,013 predicted protein interactions. The dataset was then compared with the set of human proteins contained in the version of RefSeq (Pruitt et al. 2005) downloaded in Chapter 2 and the overlap was kept leaving a positive dataset of 6,106 proteins and 18,322 protein pairs.

A negative dataset was generated by creating a set of all possible pairs of proteins from the full set of human proteins. In order to filter out proteins which could potentially interact the full set of GO (Gene Ontology)(Ashburner et al. 2000) terms associated with each protein was downloaded. All pairs with any overlaps in associated GO terms were then excluded. As with the positive set the negative dataset was compared to the set of human proteins contained in the locally held version of RefSeq (Pruitt et al. 2005) and the overlap preserved. As a final check the negative set was compared to the positive set in order to examine whether there was any overlap. There was an overlap of 9,568 proteins pairs between the positive and negative datasets or $52 \%$ of the positive set. This suggests that the use of specific GO terms is not better than the selection of random pairs as a procedure for the generation of a negative set. However as previous work such as the PIPs procedure (Scott and Barton 2007) considered in this chapter, utilised solely random pairs, it was decided that this procedure of GO + HPRD filtration was an improvement on this process. The process resulted in a negative dataset of 3,216 proteins and 207,952 protein pairs.

To use the datasets effectively as an objective measure of quality as well as a training tool for the evaluation of optimal rates at which to constrain levels of gene gain (Barker et al. 2007) for BayesTraits (Pagel et al. 2004a) the datasets were randomly split into two halves. This was done in order to cross validate the predictive power of any proposed optimal rate of gain. This process yielded a positive training set of 4,868 proteins / 9,161 protein pairs and a negative training set of 3,216 proteins / 103,971 protein pairs. The second half of the dataset was marked as testing data and contained a positive testing set of 4,796 proteins / 9,161 protein pairs and a negative testing set of 3,215 proteins / 103,974 protein pairs. The sizes of the two negative sets are uneven as 5 pairs of proteins had to be removed from the negative
training set, as they were present in a $B-A$ orientation in the positive set. Similarly two pairs of proteins had to be removed from the negative testing set.

The ratio of the size of negative to positive datasets in this case was roughly 11 to 1.This is biologically unrealistic as current estimates of the size of the full human interactome range from 154,000-369,000 (Hart et al. 2006) to 650,000 (Stumpf et al. 2008). Stumpf estimated the potential size of the interactome by treating known experimentally verified data as a sub-network of the true network and extrapolating from the sub-network to the full network (Stumpf et al. 2008). Hart on the other hand employed the idea that two independent samples (experiments) from the complete interactome or subspace of the interactome of size $N$ would be expected to share $k$ interactions by random chance under the hypergeometric distribution (Hart et al. 2006). Thus Hart estimated the size of $N$ using actually observed intersections between experiments (Hart et al. 2006).

If these numbers are subtracted from the size of all potential interactions 112,044,172,9 (calculated as all possible pairs from version of RefSeq held) the remaining ratios of negative to positive range from 1722:1 to $8617: 1$. For any full genome-wise survey it would be necessary to scale all the precision and sensitivity scores from the training set ratios to ratios constructed from estimates of the interactome size. This issue is addressed more fully in Chapter 5.

### 3.2.3 Hamming distance

The ability of Hamming distance to differentiate between the positive and negative training set was measured with a lower distance corresponding to a higher score. Precision/sensitivity were evaluated at every integer within a range of Hamming distance cut-offs ranging from 0 to 54 .

### 3.2.4 Constrained ML

To use phylogenetic profiling in a phylogenetically aware manner to detect correlations in gain and loss the software package BayesTraits was utilised (Pagel et al. 2004a). This has been used in previous work (Barker and Pagel 2005) to demonstrate that detection of correlations in gain and loss of particular genes can be used as a tool with which to detect functional interactions.

The script bms_runner (Barker et al. 2007) was used to examine the performance of different rates of gain in predicting functional interactions amongst the training sets in order to select an optimal rate. The script utilised the phylogenetic profiles and phylogeny
described in Chapter 2 as well the positive and negative training sets to evaluate the performance of different rates of gain. bms_runner (Barker et al. 2007) creates input for the program BayesTraits (Pagel et al. 2004a) to evaluate the relative likelihood of correlated evolution at a range of rates of gain. bms_runner creates a non-redundant set of profiles (Barker et al. 2007) before passing them on to BayesTraits for comparisons. Thus 113,132 protein pairs in the training set were reduced to a set of 54,906 non-redundant pairs of profiles.

A number of rates of gain were evaluated for precision and sensitivity over the training data ranging from $1 \times 10^{-6}$ up to placing no restriction on gain. An LR score was calculated for each profile pair at rate of gain and assigned to each protein pair corresponding with that profile pair. bms_runner then evaluates precision and sensitivity at a range of cutoffs commencing at the minimum LR encountered and moving up by a decreasing interval until a value close to the maximum LR is reached (Barker et al. 2007). The program then provides a table that includes the following information for this range of cut-offs.

## LR cutoff No of predictions Precision Sensitivity

Table 3.2: Column headings for data matrix returned by bms_runner (Barker et al. 2007).

### 3.2.5 Co-expression of mRNA

The co-expression of two genes in association with a given environmental condition can be considered a potential indictor of functional linkage. In order to examine the performance of the ML reconstruction method in predicting protein functional interactions against the coexpression of mRNA the results of all microarray experiments held in the EBIs ArrayExpress database were downloaded.

This data was pre-processed and thus contained expression data at the gene level rather than at the probe level. As oligonucleotide probes only map to small subsections of a gene and also can hybridise with multiple targets the relationship between probe to gene is many-to-many. This many-to-many relationship was collapsed by data processing carried out by ArrayExpress on each individual experiment.

Thus a total of 377 experiments were downloaded. Each experiment record contained information on genes whose expression level varied significantly in response to the experimental treatment/tissue state. The size of individual experiments ranged in size from 1
gene to a maximum of 15,987 . The mean number of genes showing significant variation per experiment was approximately 3,143 .

A sample line from the downloaded data is shown below for illustrative purposes:

| Gene Symbol | Ensembl ID | Species | Factor | Value | Accession | Expression |  | $p$ Value |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| STAT1 | ENSG00000115415 | Homo | Disease | normal | E- | DOWN | 0.0423247888020165 |  |
|  |  | sapiens | state |  | GEOD- |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |

Table 3.3: Sample processed data from ArrayExpress for experiment E-GEOD-3790 (Hodges et al. 2006).

E-GEOD-3790 is a study on gene expression in brain tissue afflicted with Huntington's disease (Hodges et al. 2006). The factor column has a number of potential values corresponding to the annotation of the individual samples. In this case it corresponds to whether the tissue comes from a patient diagnosed with Huntington's disease as opposed to normal tissue. The value column shows the value of the factor (in this case normal). The $p$ value column shows the significance of the identified differential expression. Thus the data presented above shows that the gene STAT1 is significantly ( $p<0.05$ ) down-regulated in tissue annotated as normal.

To use the training datasets to measure the ability of gene co-expression to predict protein interactions it was necessary to convert the training data from protein pairs to gene pairs. Using a translation key provided by the International Protein Index (IPI) (Kersey et al. 2004) each RefSeq Gi number was mapped to the associated gene name. As individual genes can produce multiple proteins via the process of alternative splicing there isn't a one to one correspondence between the number of genes and the number of protein in both training sets. With some Gi entries missing in the translation key this created a positive gene training set of 4319 genes / 8057 gene pairs and a negative set of 2833 genes / 89549 pairs of genes.

In order to predict functional linkage the following procedure was followed.

- The data in each microarray experiment was split according to experimental condition.
- For each experimental condition pairs of genes were marked as functionally linked if their expression level went up or down in response to the given experimental condition.
- True positives were counted if the genes existed as a pair in the positive set.
- False positives were counted if the genes existed as a pair in the negative set.
- False negatives were counted as the complement of the set of predictions and the positive set.
This data was processed using a program implemented in Java, which followed the steps presented below. The input to the program was a file containing a set of lines as shown above for a single experiment.
- Create a non-redundant set of factors present in the experiment.
- Create a non-redundant set of values for each factor.
- Create an empty set $L$ to hold functionally linked pairs
- For each factor $F$ :

For each factor value $F V$ :
For each expression value $E$ (UP or DOWN)
Create a set $S$ of all genes possessing the attributes $\{F, F V, E\}$ where the associated $p$ value for $E$ is less than 0.05
All genes within $S$ are declared functionally linked.
All possible pairs of genes within $S$ are added to $L$

- $L$ is now evaluated against the training data for precision/sensitivity.


### 3.2.6 Bayesian classifier

The PIPs server offers its data split into six score cut-offs: $0.25,1,2.5,25,250$ and 2500 (McDowall et al. 2009). These datasets were downloaded and evaluated for precision and sensitivity at each of these cut-offs. For each cut-off the server provides a file with pairs of proteins and the associated posterior odds ratio score. Each pair of proteins within each file was declared to be functionally linked and then evaluated against the training data. This provided associated precision/sensitivity scores for each cut-off.

### 3.3 Results

### 3.3.1 Hamming distance

Figure 3.2 shows the performance of phylogenetic profiling over the training sets using Hamming distance.


Figure 3.2: Performance of phylogenetic profiling using Hamming distance over the training data.

Hamming distance as a measure does not perform well over the training data achieving a maximum precision of 0.08796 .

### 3.3.2 Constrained ML

The ability of constrained ML (Barker et al. 2007; Barker and Pagel 2005) to distinguish between the training data was tested at a number of rates of gain. The results of this can be seen in Figure 3.3.


Figure 3.3: Performance of constrained ML (Barker et al. 2007; Barker and Pagel 2005) over training data at different rates of gain.

Figure 3.3 shows points at a range of sensitivity between $0-1$. Sensitivity over the whole training set ranges between 1 at LR cut-offs of 0 where all pairs of proteins are
predicted to be functionally linked to 0 at the points where no pairs from the positive set are predicted to be functionally linked.

Precision ranges from 0.0809 (a base level that is derived from the ratio of the size of the positive set to the size of the negative set) up to 1 which is the point at which all predictions made at a given LR cut-off lie in the positive set, i.e. are true positives.

Of the two metrics (precision/sensitivity) it is precision, which appears to be the strong suit of constrained ML. This is probably due to the fact that correlated evolution will not occur in all cases of protein interactions. A large number of protein interactions will contain members that are phylogenetically ubiquitous. In some of these cases the interaction will be essential to maintenance of normal eukaryotic cellular function. In other cases even if an interaction is being lost or gained in an organism, its individual members might still be present (the interaction being lost due to some form of temporal/spatial separation of the members). Thus as low sensitivity is inevitable with this method; it was decided to focus on rates of gain that achieve $100 \%$ precision. Figure 3.4 places all rates of gain on a single plot and zooms into a range of sensitivities between 0 and 0.001 .


Figure 3.4: Performance of constrained ML (Barker et al. 2007; Barker and Pagel 2005) over training data magnified to a scale of sensitivity ranging from 0 to 0.001 . For clarity some of the worse performing rates of gain are removed.

Figure 3.4 shows that the rate 0.025 is the clear best performer as it delivers predictions with a precision of 1 at the highest sensitivity. The LR cut-off at this point is 58.54. The sensitivity at this cut-off for the rate 0.025 is 0.000545 . This rate is thus chosen as the exemplar rate to represent the method in comparisons and to utilise for further analysis. The findings of (Barker et al. 2007) were borne out in this investigation as lower rates of gain were generally seen as the best performers.

Over the training data constrained ML (Barker et al. 2007; Barker and Pagel 2005) with the rate of gain constrained to 0.025 makes five predictions from the positive set shown in Table 3.1
RefSeq Accessions Annotation Protein A Annotation Protein B Interaction type Verified By

| NP_001789 | Cell division protein | Origin recognition | Direct | Protein microarray <br> NP_004144 <br> complex subunit 1 |
| :--- | :--- | :--- | :--- | :--- |
|  | kinase 2 |  | (Ramachandran et |  |


| NP_005617 | Splicing factor, arginine/serine-rich | Splicing factor, arginine/serine-rich 6 | Complex | Site-directed mutagenesis |
| :---: | :---: | :---: | :---: | :---: |
| NP_006266 |  |  |  | (Monsalve et al. 2000) |
| NP_001789 | Cell division protein kinase 2 | Cyclin dependent kinase 7 | Direct | In-vitro experimentation |
| NP_001790 |  |  |  | (Garrett et al. 2001) |
| NP_001347 | DEAD/H (Asp-Glu- | Exportin 1 | Direct |  |
| NP_003391 | Ala-Asp/His) box polypeptide 3 |  |  | experimentation <br> (Yedavalli et al. |
|  |  |  |  | 2004) |
| NP_066953 | Peptidyl-prolyl cistrans isomerase A | Serine/threonineprotein phosphatase | Direct | Yeast 2-hybrid (Stelzl et al. 2005) |
| NP_000935 |  | 2B catalytic subunit alpha |  |  |

Table 3.1: True positive proteins predicted by constrained ML (Barker et al. 2007; Barker and Pagel 2005) from the training data with rate of gain constrained to 0.025 at an LR cutoff of 58.54 .

The examination of the training datasets yielded a similar result to (Barker et al. 2007) in so much as lower rates of gain tended to perform better.

As 0.025 was selected as the optimum rate of gene gain over the training data it was also tested on the testing data to cross validate this selection. Figure 3.5 shows the results of this cross validation check.


Figure 3.5: Performance of constrained ML (Barker et al. 2007; Barker and Pagel 2005) over test data with rate of gain constrained to 0.025 .

Figure 3.5 shows that constraining the rate of gain to 0.025 also achieves a precision of 1 over the testing data. This precision occurs at an LR cutoff of 53.3. The sensitivity at this point is 0.00054 . At this cutoff constrained ML (Barker et al. 2007; Barker and Pagel 2005) makes 5 predictions that are true positives as shown in Table 3.2.

| RefSeq Accessions | Annotation Protein A | Annotation Protein B | Interaction type | Verified By |
| :---: | :---: | :---: | :---: | :---: |
| NP_002583 | proliferating cell | ATP-dependent | Direct | In-vitro |
| NP_001347 | nuclear antigen | RNA helicase |  | experimentation |
|  |  | DDX3X isoform 1 |  | (Ohta et al. 2002) |
| NP_001118 | adaptor-related | AP-1 complex | Direct | Yeast 2-hybrid |
| NP_001119 | protein complex 1 | subunit gamma-1 |  | (Takatsu et al. |
|  | beta 1 subunit | isoform b |  | 2001) |
|  | isoform a |  |  |  |
| NP_000391 | TFIIH basal | cell division protein | Direct | In-vitro |
| NP_001790 | transcription factor | kinase 7 |  | experimentation |
|  | complex helicase |  |  |  |
|  | XPD subunit |  |  | (Coin et al. 1998) |
|  | isoform 1 |  |  |  |
| NP_005517 | heat shock factor | heat shock factor | Direct | In-vivo/in-vitro |
| NP_004497 | protein 1 | protein 2 isoform a |  | experimentation |
|  |  |  |  | (He et al. 2003) |
| NP_066953 | peptidyl-prolyl cis- | calcineurin subunit | Complex | In-vitro |
| NP_000936 | trans isomerase A | B type 1 |  | experimentation |
|  |  |  |  | (Huai et al. 2002) |

Table 3.2: True positive proteins predicted by constrained ML (Barker et al. 2007; Barker and Pagel 2005) from the testing data with rate of gain constrained to 0.025 at an LR cutoff of 50.217 .

### 3.3.2.1 Likelihood ratio statistic

The likelihood ratio statistic (LR) derived from the comparison of the independent and dependant models of evolution is asymptotically distributed as a $\chi^{2}$ variate with degrees of freedom equal to the difference of numbers in parameters between the two models which in this case equals 4 under assumptions about the size of the phylogeny and the speed of evolution of the character under consideration (Barker and Pagel 2005; Pagel 1997). Thus if the LR falls within the critical region of the distribution it is considered significant. A histogram showing the theoretical $\chi^{2}$ distributions with 4 degrees of freedom is shown below in Figure 3.6.


Figure 3.6: Theoretical $\chi^{2}$ distribution with 4 degrees of freedom.

The distribution of LRs in the positive and negative set as well as over the combined training data differs from this theoretical distribution as can be seen in Figures 3.7, 3.8, 3.9 and 3.10.


Figure 3.7: Distribution of likelihood ratio statistic for constrained ML within the rate of gain 0.025 over the positive training set.


Figure 3.8: Distribution of likelihood ratio statistic for constrained ML within the rate of gain 0.025 over the negative training set.


Figure 3.9: Distribution of likelihood ratio statistics for constrained ML (Barker et al. 2007; Barker and Pagel 2005) within the rate of gain 0.025 over the complete training dataset.

| Minimum | 1st Quartile | Median | Mean | 3rd Quartile | Maximum |
| :--- | :---: | :--- | :---: | :---: | :---: |
| 0.08932 | 7.85 | 10.51 | 11.14 | 13.79 | 74.80 |

Table 3.3: Descriptive statistics for the distribution of likelihood ratios for the rate of gain 0.025 over the complete training data.


Figure 3.10: Distribution of likelihood ratio statistics for constrained ML within the rate of gain 0.025 over the complete training dataset, the positive training dataset and the negative training dataset compared with the theoretical $\chi^{2}$ distribution with 4 degrees of freedom.

The distribution of LR statistics over the training data seems to differ from the theoretical $\chi^{2}$ distributions with 4 degrees of freedom.

This distribution was also tested via a two-sample Kolmogorov-Smirnov test for goodness of fit between a generated theoretical $\chi^{2}$ distribution with 4 degrees of freedom and the LR statistic score distribution over the training data using R (R Development Core Team 2011). This also showed a difference between the two distributions ( $D=0.9993$, $p$-value $<2.2 \mathrm{e}^{-}$ ${ }^{16}$ ).

This may be due to a violation of assumptions of the model with regards to the speed of character transition.

The overall frequency of higher LR statistics does appear to be higher in the positive set which is further validation for the constrained ML method.

### 3.3.3 Co-expression of mRNA

The results for each microarray experiment measured over the training data are given below in Figure 3.11.


Figure 3.11: Precision/ sensitivity results for 377 microarray experiments over the training datasets.

As before the area of interest in Figure 3.11 is the point at which precision equals 1. This is because the average correlation between transcript abundance and peptide abundance has been observed to be fairly low in primates at around 0.33 (Fu et al. 2007). Thus mRNA co-expression is unlikely to be capable of high sensitivities in protein-protein interaction detection. Figure 3.12 is a magnification of this area.


Figure 3.12: precision/ sensitivity results for microarray experiments over the training datasets magnified to a scale of sensitivity ranging from 0 to 0.01 .

Mean precision over all 377 microarray experiments was 0.2141 and mean sensitivity was 0.1195 . Out of the 377 total 18 experiments achieved a precision of 1 . Details of these experiments are shown in Table 3.4.

| Accession | Size | Description of experiment Sensitivity |  |
| :---: | :---: | :---: | :---: |
| E-GEOD-4567 | 166 | Transcription profiling of human pulmonary artery endothelial cell culture treated with Chapel Hill Ultrafine particle. | 0.0006547359 |
| E-GEOD-2280 | 168 | Transcription profiling of oral cavity samples from human squamous cell carcinoma patients (O'donnell et al. 2005). | 0.0003274752 |
| E-GEOD-3183 | 255 | Transcription profiling of human bronchial cell line treated with IL-13 to better understand early cytokine-mediated mechanisms that lead to asthma. | 0.0002183168 |
| E-GEOD-994 | 266 | Transcription profiling of human intra-pulmonary airways and buccal mucosa to identify the effects of cigarette smoke on the human airway epithelial cell transcriptome (Spira et al. 2004). | 0.0002183168 |
| E-GEOD-2152 | 474 | Transcription profiling of human uterine fibroids mith mutated or wild type fumarate hydratase gene (Vanharanta et al. 2006). | 0.0008728860 |
| E-GEOD-2504 | 28 | Transcription profiling of untreated, HIV-1 vector-infected and TNFalpha-treated human Jurkat T cells (Lewinski et al. 2005). | 0.0002182929 |
| E-GEOD-4748 | 191 | Transcription profiling of human dendritic monocytes treated with LPS (lipopolysaccharide) or CyP (Cyanobacterial Product) (Macagno et al. 2006). | 0.0004366336 |
| E-MEXP-1224 | 538 | Transcription profiling of human colon samples from patients who have colorectal cancer recurrence or are recurrence-free (Garman et al. 2009). | 0.0030511060 |
| E-GEOD-7664 | 254 | Transcription profiling of human PBMC response to benzene metabolites (Gillis et al. 2007). | 0.0008728860 |
| E-GEOD-2361 | 87 | Transcription profiling of 36 normal human tissue types to identify tissue-specific genes (Ge et al. 2005). | 0.0003274752 |
| E-GEOD-1739 | 212 | Transcription profiling of blood samples from human patients with severe acute respiratory syndrome (SARS) (Reghunathan et al. 2005). | 0.0002182929 |

Table 3.4: Microarray experiments achieving a precision of 1.

| E-TABM-577 | 95 | Transcription profiling of human placenta from women presenting at term with villitis of unknown etiology (Kim et al. 2009). | 0.0001091584 |
| :---: | :---: | :---: | :---: |
| E-GEOD-2018 | 129 | Transcription profiling of human bronchoalveolar lavage samples collected from lung transplant recipients with rejection states determined at the time of sample collection (Lande et al. 2003). | 0.0001091584 |
| E-GEOD-1786 | 126 | Transcription profiling of human male vastus lateralis muscle samples from healthy and COPD subjects before and after 3 months of training (Radom-Aizik et al. 2005). | 0.0008728860 |
| E-GEOD-2624 | 293 | Transcription profiling of human tetracycline-regulated cell line expressing an NF-kB inhibitor to systematically identify NF-kB dependent genes (Tian et al. 2005). | 0.0014181302 |
| E-MEXP-714 | 55 | Transcription profiling of human hepatitis C virus replicon cell line treated with interferon-alpha 2 a in a time series. | 0.0002182929 |
| E-GEOD-9770 | 1225 | Transcription profiling of human neurons from different brain regions derived from individuals with mild cognitive impairment. | 0.0008731718 |
| E-GEOD-403 | 333 | Transcription profiling time series of the cAMP-induced decidualization of human endometrial stromal cells (Tierney et al. 2003). | 0.0007641087 |

Table 3.4: Microarray experiments achieving a precision of 1 (cont).

The highest scoring microarray experiment was E-MEXP-1224, an investigation into whether there was a difference in expression profiles between the colorectal tissue of patients who has recurrent cancer and those who remained clear (Garman et al. 2009). The sensitivity of this experiment was $3.0511 \times 10^{-3}$ with a precision of 1 .

### 3.3.4 Bayesian classifier

The ability of the PIPs (McDowall et al. 2009) server to predict functional interaction over the training set was evaluated at 6 cutoffs. The results can be seen in Figure 3.13


Figure 3.13: Precision/ sensitivity results for predictions from the PIPs server over six cut-offs over the training dataset.


Figure 3.14: precision/ sensitivity results for predictions from the PIPs server over six cutoffs over the training dataset zoomed in to a maximum sensitivity of 0.15 .

None of the score cut-offs over the predictions from the PIPs server achieved a full precision of 1 . However none of them fell under 0.9 either as seen in Table 3.5.

| Cutoff | Predictions | Precision | Sensitivity <br> 0.25 |
| :--- | :---: | :---: | :---: |
| 79441 | 0.9135546 | 0.14366504 |  |
| 1.00 | 37606 | 0.9395973 | 0.11068444 |
|  |  |  |  |
| 2.50 | 25598 | 0.9533333 | 0.09825928 |
| 25.00 | 5394 | 0.9949239 | 0.04742318 |
| 250.00 | 1232 | 0.9865772 | 0.01832689 |
|  |  |  |  |
| 2500.00 | 498 | 0.9883721 | 0.01067973 |



Figure 3.12: All methods compared over training dataset. Legend explanation (PIPs=PIPs server, MA= microarray experiment and $\mathrm{PP}=$ phylogenetic profiling measuring correlation in gain and loss over a phylogeny with constrained rate of gain).

### 3.4 Discussion

Arguably the best performing method out of all three methods is the PIPs server (McDowall et al. 2009) as it achieves the highest rates of combined precision and sensitivity over the training data. The success of the PIPs server in terms of accuracy and coverage is attributable to its use of multiple, disparate sources of evidence. The other two methods both focus on particular types of interactions.

Phylogenetic profiling measured with constrained ML over a phylogeny is limited to proteins that have been gained and lost in a correlated fashion over a phylogeny. Thus protein interactions between phylogenetically ubiquitous partners cannot be detected. Similarly it cannot detect interactions between interactors with potentially redundant partners.

Microarrays are more flexible in the types of interaction they are capable of detecting. However individual experiments are limited in the types of interactions that they can uncover by the experimental conditions under which their constituent mRNAs were extracted. They are also biased toward stable complexes (von Mering et al. 2002). Another limitation in the use of microarray experiments in the prediction of protein interactions is the fact that expression levels of a gene at the transcription level do not correlate strongly with overall levels of protein production at the translational level (Gygi et al. 1999). This is due to regulation at the posttranscriptional level by factors such as mRNA half-life, codon usage and ribosome occupancy and density (Wu et al. 2008). The best performing microarray experiments outperformed constrained ML in terms of sensitivity.

However given the difference in cost and labour intensiveness between a microarray experiment and a computational analysis employing phylogenetic profiling, the latter can clearly be a useful tool in the functional annotation of identified genes within a newly sequenced genome.

### 3.4.1 Low Sensitivities

None of the methods as described and utilised above can are particularly sensitive in detecting protein-protein interactions. Constrained ML and gene co-expression are insensitive to protein-protein interactions for the reasons described above.

The PIPs server as the best performer achieves a sensitivity of 0.14 at a high level of precision. However this still corresponds to a $14 \%$ chance of detecting a possible protein interaction despite its integration of various forms of supporting evidence. It is possible that it is this integration of evidence that renders PIPs insensitive. If for example the likelihood ratio returned by one of its predictor modules was high with the rest all being low, the overall posterior odds ratio score would be low. Thus the individual sensitivities of the module predictors are averaged out.

It seems that maximising coverage of the interactome is beyond the scope of each of the predictive methods considered in this chapter. To use the analogy of the interactome as a dark room, none of these methods are equivalent to an overhead light that illuminates every corner of the room. Rather each method is more like a lamp that casts a pool of light on its immediate surroundings. It is only by lighting a number of these lamps that the entire room can be illuminated.

## Chapter 4

## Design and implementation of data filter

### 4.1. Introduction

The constrained maximum likelihood (ML) method used to detect proteins which share correlated evolutionary histories as described in Chapter 3 and in work by Barker et al. (Barker et al. 2007; Barker and Pagel 2005) estimates values for parameters which model the transition rates of the gain and loss of discrete characters (Pagel 1994) by integrating over all possible ancestral states at each node within the phylogenetic tree.

As pointed out by Barker (Barker et al. 2007) placing a constraint on the rate of acquisition of new proteins increases the ability of the likelihood method to discriminate between proteins that interact and those that do not. The determination of an optimum rate of gain reduces the scale of the problem of parameter estimation (Barker and Pagel 2005) as it reduces the numbers of parameters to be fitted to 2 for the independent model and 4 for the dependent model.

The detection of potential functional interactors for a single given protein using this method is possible, however given the low sensitivity of the method (see Chapter 3) the probability of detecting a functional interaction for any given single protein or even a set of proteins is low. A complete genome-wide survey however would detect all protein pairs that displayed evidence of correlated evolution.

The procedure is however prohibitively slow for a complete genome-wide survey without access to a significant amount of computing power. A timed training run over the training dataset for a single rate of gain took approximately 110 CPU-hours to conclude 54,906 comparisons of non redundant phylogenetic profile pairs on a single core of a 3 GHz dual-core Intel Xeon processor (see Section 4.5). As there are $60,615,555$ possible nonredundant pairs of phylogenetic profiles in the version of the human proteome currently held; a full genome comparison would take 121,825.05 CPU-hours or 13.9 CPU-years on the single core of a dual-core 3 GHz Intel Xeon processor. The speed of constrained ML (Barker et al. 2007; Barker and Pagel 2005) was also measured in work presenting a genome order based approach to phylogenetic profiling (Cokus et al. 2007). In this case it was found to range between 5-15 seconds per pair of proteins (Cokus et al. 2007). This caused the authors to utilise a subset of their data in their benchmarking study of constrained ML (Cokus et al. 2007).

Potentially access to multi-core CPUs and/or computing clusters could ameliorate this to a certain extent. As application of constrained ML (Barker et al. 2007; Barker and Pagel 2005) involves sequential comparison of pairs of phylogenetic profiles, it is a process that is easily amenable to parallelisation via splitting the task into a smaller set of tasks, which can be launched in parallel. Task farming is applied in computational biology to tasks that are potentially intractable if tackled serially, e.g. analysis of gel electrophoresis data (Dowsey et al. 2003) or analysis of microarray data (Hill et al. 2008). However even with the application of task farming it is clear that a full genome-wide survey is not feasible for this method on any averaged sized eukaryotic genome.

This chapter details the development of a data filter to remove protein pairs that display little or no evidence of correlated evolution. There are two main types of filter evaluated. The first type is a simple distance based test (Hamming distance) as shown in Chapter 3 and utilised in early work on phylogenetic profiling (Pellegrini et al. 1999). Potentially proteins that display evidence of correlated evolution will have phylogenetic profiles that have a lower Hamming distance from each other. Thus even though Hamming distance applied in isolation performs poorly as seen in Chapter 3, it may serve as a filter for proteins which do not display evidence of correlated evolution in combination with the second type of filter.

The second type of filter will utilise a single set of reconstructed ancestral states. By using a single set of reconstructed states and a simpler method for the detection of evidence of correlated evolution proteins that do not display any such evidence may be filtered out. This chapter describes the implementation and comparison of five filters, which utilise a single set of reconstructed ancestral states to detect signs of correlated evolution. As a large amount of the computations performed by constrained ML (Barker et al. 2007; Barker and Pagel 2005) involve estimation of the transition rate parameters by integrating over all possible ancestral states, use of a single set of reconstructed ancestral states reduces the scope of the problem. Through the use of an effective and accurate data filter a genome-wide survey for an average eukaryotic organism could be rendered feasible.

The end product of this research described in this chapter is just such a filter based on logistic regression of a set of empirically evaluated predictors/parameters, which reflect correlated evolution between a pair of proteins. The filter is approximately 2208 times faster then constrained ML and achieves a reasonable degree of precision/sensitivity over the training data in its own right. Thus application of this filter can facilitate a heuristic search
for genes/proteins displaying evidence of correlated evolution over an entire genome/proteome.

In order to describe the process of filter development/evaluation it will firstly be necessary to present an overview of ancestral state reconstruction.

### 4.1.1 Ancestral state reconstruction

The procedures involved in the reconstruction of the states of characters and traits in extinct ancestral species are similar to those involved in phylogeny reconstruction. This is due to the similarity of the issues involved. The reconstruction procedures for character states thus utilise similar criteria with which to judge putative reconstructions. Ancestral reconstruction is a useful tool for investigating hypothetical evolutionary scenarios having been used to investigate many biological questions such as for example the demonstration of homoplasy in the evolution of lysozyme (Malcolm et al. 1990; Messler and Stewart 1997; Stewart et al. 1987). It is also a prerequisite step for a number of comparative method tests (Maddison 1990; Ridley 1983).

### 4.1.1.1 Parsimony

A parsimonious reconstruction of ancestral states over a phylogenetic tree would entail the selection of the internal state that minimised change. Thus if for example two terminal nodes within a given clade had the same internal state the same state would be assigned to the node immediately preceding them.

Algorithms such as the Fitch (Fitch 1971) and Sankoff (Sankoff 1975) algorithms as described in Chapter 2 are used employed as a step within phylogeny reconstruction (Albert 2006). However given a particular already constructed phylogenetic tree they can be employed to reconstruct a set of ancestral node values which minimises evolutionary change over that particular tree (Felsenstein 2004). The algorithms themselves do not reconstruct individual states at each internal node but instead construct sets of potential states at each node. These potential states can be resolved into a singular state reconstruction through the application of algorithms such as ACCTRAN (Accelerated transformation) (Swofford and Maddison 1987), which reconstructs ancestral states by placing points of change as close to the root of the tree as possible (Agnarsson and Miller 2008). The converse approach to ACCTRAN is DELTRAN (delayed transformation)(Swofford and Maddison 1987), which reconstructs ancestral states by placing points of change as close to the tips of the tree as possible (Agnarsson and Miller 2008). ACCTRAN and DELTRAN are the most commonly
used methods for collapsing node state sets into individual node states though of the two ACCTRAN is the more widely employed (Agnarsson and Miller 2008).

Parsimony methods fail to consider different branch lengths in different parts of the tree (Yang et al. 1995). Parsimony based methods have also been criticised for their lack of statistical soundness (Elias and Tuller 2007). Parsimony methods are also unable to distinguish between reconstructions that are equally parsimonious (Koshi and Goldstein 1996).

### 4.1.1.2 Likelihood

In a similar fashion as likelihood is employed as an optimality criterion for phylogeny generation, it can also be used in the context of ancestral state reconstruction. Maximum likelihood techniques are used to estimate the parameters of the specified model of evolution (Yang 2006). Once these parameters are estimated they can be utilised to calculate the posterior probability of ancestral states using Bayes theorem (Yang 2006). The state with the highest posterior probability is then assigned to the node under consideration. This procedure has been defined as empirical Bayes (Yang 2006). Empirical Bayes can be used to either assign a character state to a set of nodes in a tree via a process known as marginal reconstruction or it can be used to assign a set of possible characters to each node (Yang 2006). This latter process is known as joint reconstruction (Yang 2006).

Empirical Bayes can be contrasted with hierarchical Bayes where rather than estimating a single value for the parameters of a model of evolution a prior probability distribution is assigned for each unknown parameter (Yang 2006). The posterior probability for a given ancestral state is then calculated by integrating over all possible values of parameters (Huelsenbeck and Bollback 2001). Again the putative state with the highest posterior probability is then assigned to each ancestral node.

Work by Koshi and Goldstein used the empirical Bayes method to reconstruct the sequence of ancestral ribonuclease (Koshi and Goldstein 1996). The performance of parsimony and the empirical Bayes method was also compared in a reconstruction of lysozyme c by Yang et al. (Yang et al. 1995). This work found that empirical Bayes outperformed parsimony but both methods suffered when the sites within the multiple alignments being reconstructed were highly variable and the distance from the ancestral nodes to the extant species was high (Yang et al. 1995).

An interesting application of empirical Bayes reconstruction was carried out by Gashen (Gaschen et al. 2002). This work entailed reconstruction of the reconstruction of the
sequence of the ancestor to various regional variants of the HIV-1 virus in order to contribute to the creation of a potential vaccine (Gaschen et al. 2002).

### 4.2 Filters

### 4.2.1 Hamming distance filter

The original work which introduced the methodology of phylogenetic profiling as a means of detection of functional interaction between genes (Pellegrini et al. 1999) utilised Hamming distance (Hamming 1950) as a measure of similarity of profiles. Phillip Kensche also examined this method in a review of phylogenetic profiling methods, and found it to perform reasonably well over a dataset composed of the proteins sequences of 25 fungi (Kensche et al. 2008). Hamming distance did not perform well over the training data as seen in Chapter 3 however it was possible that it could reduce the possible search space for an application of constrained ML. As a potential heuristic it offers speed, as Hamming distance is one of the simplest comparisons that can be carried out between two strings. Hamming distance therefore was investigated as a potential filter to be used possibly in conjunction with a filter based on a single set of reconstructed states.

### 4.2.2 Ancestral state reconstruction filter

The first consideration in the development of a heuristic/filter based on a single set of reconstructed characters was which criterion to use to reconstruct that set. Likelihood as a criterion yields more accurate results as discussed above. However as the aim of this heuristic approach was to develop a method that reduced the search space for an application of the computationally intensive constrained ML (Barker et al. 2007; Barker and Pagel 2005) to phylogenetic profiling, it was decided to use the simpler though less accurate criterion of parsimony.

### 4.2.2.1 Dollo parsimony

Dollo parsimony operates under the assumption that once a complex trait has been lost it cannot be re-acquired (Albert 2006). Given that the character under investigation is the presence and absence of genes/proteins in eukaryotic organisms it was decided that Dollo parsimony was the appropriate variant to use. Dollo parsimony has been previously used to investigate the propensity of particular genes to be lost over the course of evolutionary time in eukaryotes (Krylov et al. 2003). It was chosen by the authors due to the relative rarity of lateral gene transfer events in eukaryotes (Krylov et al. 2003).

Dollo parsimony has also been utilised to investigate gene gain in poxviruses (McLysaght et al. 2003). The results of this use however may have been affected by the fact that poxviruses were later observed to acquire genetic material from infected hosts (Hughes and Friedman 2005). Kensche also evaluated the efficacy of Dollo reconstructions of profiles as a method of phylogenetic profiling (Kensche et al. 2008). Kensche utilised a distance measure $d(A, B)$ between the Dollo parsimonious reconstructions of the phylogenetic profiles of two (orthologous groups of ) proteins $A$ and $B$ calculated as:

$$
\begin{equation*}
d(A, B)=\sum_{i \in b \text { branches }}\left|\left(\operatorname{anc}\left(a_{i}\right)-\operatorname{desc}\left(a_{i}\right)\right)-\left(\operatorname{anc}\left(b_{i}\right)-\operatorname{desc}\left(b_{i}\right)\right)\right| \tag{1}
\end{equation*}
$$

where branches denoted the set of branches in the phylogenetic tree, anc $\left(a_{i}\right)$ was defined as the state of orthologous group $A$ at the ancestral node of branch $i, \operatorname{desc}\left(a_{i}\right)$ was defined as the state of orthologous group $A$ at the descendant node of branch $i, \operatorname{anc}\left(b_{i}\right)$ was defined as the state of orthologous group $B$ at the ancestral node of branch $I$ and $\operatorname{desc}\left(b_{i}\right)$ was defined as the state of orthologous group $B$ at the descendant node of branch $i$ (Kensche et al. 2008). The distance $d(A, B)$ was a count of branches where either orthologous group was gained or lost independently. The method performed as well as more sophisticated techniques on the data analysed by Kensche (Kensche et al. 2008).

One of the methods evaluated by Barker as a potential source of signal for correlated evolution was also examination of Dollo parsimony based reconstructions of phylogenetic profiles over a phylogeny (Barker et al. 2007). Dollo parsimony was utilised as it reflected the idea of setting the rate of acquisition of a complex trait (in this case a protein) to a preset low level (Barker et al. 2007). Pairs of proteins were scored on branches of the tree where they were jointly lost and jointly gained to form a score referred to as Dollo-pos (Barker et al. 2007). Branches where proteins were not gained or lost together were also counted and subtracted from Dollo-pos to form a score referred to as Dollo-overall (Barker et al. 2007). Both these scores however did not perform particularly well over the data examined (Barker et al. 2007). Dollo-overall however performed significantly better than Dollo-pos (Barker et al. 2007).

Thus given the fact that Dollo parsimony based tests had been moderately successful at detecting correlated evolution, a series of potential data filters /heuristics for examination of phylogenetic profiles using constrained ML (Barker et al. 2007; Barker and Pagel 2005)
based on a single set of reconstructed ancestral states over the phylogeny using Dollo parsimony were investigated.

### 4.2.2.2 Maddison Test for correlated evolution

To use the reconstructed ancestral state data a test to detect correlated evolution using the comparative method that utilised a set of reconstructed ancestral states over a given phylogenetic tree was needed. One candidate test was a contingency table based test presented by Ridley where a gain or loss of a character was considered in the light of whether it occurred in the presence or absence of another character over a phylogenetic tree (Ridley 1983). This test however does not separate which character is dependent and which is independent.

A second candidate test considered was a procedure described by Wayne Maddison (Maddison 1990) for the comparison of the association of changes in one binary character with the given state of another. This test was designed to carry out this analysis assuming a given phylogenetic tree and a set of reconstructed characters (Maddison 1990). This test has been referred to as a test for concentrated changes (Felsenstein 2004).

The fundamental idea behind the Maddison test is to test whether changes in one trait or character are concentrated in an area of a tree where a second trait or character in a given state. As an illustrative example consider a fictional monophyletic group of related cow-like animals. These animals do not possess horns. The phylogenetic relationships of these animals are fully resolved and understood as well as the ancestral states for all morphological and molecular traits. Now imagine that this group overall has no ability to metabolise valine. Finally imagine the ability to metabolise valine is independently acquired by a sub-clade of our fictional group and this leads to the development of horns in this sub-clade.

If we wished to test whether the ability to metabolise valine leads to horn development, the Maddison test would return the probability of the observed configuration of valine metabolism / horn presence. This probability would be calculated by firstly calculating the total number of ways to acquire horns in the presence of the ability to metabolise valine over the phylogenetic tree. Secondly the number of ways to acquire horns over the entire tree irrespective of the state of the ability to metabolise valine are calculated. By dividing the first value by the second a probability can be calculated. If horns are concentrated in parts of the tree where valine metabolism is also present this probability will be lower.


Figure 4.1: Illustrative example tree.

Thus imagine in the above figure only Cow1 has the ability to metabolise valine and also possesses horns. Thus a reasonable hypothesis/reconstruction could be that both abilities were gained in the branch leading to Cow1. There is 1 gain of horns. Over the entire tree there are 6 branches (not counting the root branch) and thus 6 ways to have 1 gain of horns. However there is only one way of having a gain of horns in the presence of valine metabolism and that is on the branch leading to Cowl. Thus the probability of the observed configuration is $\frac{1}{6}$.

To reiterate the test works through counting all possible ways of having a set of observed changes in a character over a phylogeny and then counting how many ways there are of having the same number of changes in parts of the tree where a second character is in a given state. Thus if correlated evolution is occurring changes in the first character will be concentrated in areas of the tree where the second character is in the causative state. Consider as a second example two proteins, which carried out the same function. If the presence of the first protein made the second protein redundant then losses in the second protein could be concentrated in areas where the first protein was present.

The drawbacks of the test are the fact that it treats all forms of evolutionary change as equally likely and its inability to take into account branch lengths (Pagel 1994). However as the motivation behind the implementation of the test was its use as a simple data filter to remove protein pairs that showed little or no evidence of correlated evolution it was decided that the Maddison test (Maddison 1990) was an appropriate test.

### 4.3 Methods

To create Dollo parsimony based reconstructions over each phylogenetic profile over the phylogeney presented in Chapter 2, the program DOLLOP from the PHYLIP package (Felsenstein 1989) was used. The program implements the Dollo parsimony reconstruction algorithm described in work by Farris (Farris 1977).

Given a binary trait $T$ that can take on 2 possible values coded as [ 0,1 ], DOLLOP implements Dollo parsimony by seeking to explain a given observed configuration of presence and absence for $T$ over a set of taxa over a phylogenetic tree by allowing one gain (transition from 0 to 1) and multiple reversions (transition from 1 to 0 ) (Felsenstein 1989).

As an illustrative example consider the tree below and a trait with the distribution 010101.


Figure 4.2: Example tree.

DOLLOP will reconstruct the trait as initially gained at the root of the tree and lost at the braches leading to $\mathrm{A}, \mathrm{B}$ and E . This is as opposed to allowing multiple gains on the branches leading to D and G .

The process followed was similar to the process followed to generate the genome content tree produced in Chapter 2. The main difference in this case was that DOLLOP was run with the -U option, which instructed it to produce Dollo parsimonious reconstructions over a user-supplied tree. The program was supplied with the phylogeny generated in Chapter 2 as well as the phylogenetic profile for each protein under consideration.

Apart from the -U option the program was run with its default settings. The output from DOLLOP contained data on the state of the protein at every node in the phylogeny as well the branches within the phylogeny at which transitions occurred. In order to record this data DOLLOP assigns an identifying number to each internal node of the phylogeny.

An example of the outputted data from DOLLOP is given below. For a human protein with the profile 000000000000000000000000100000000000000000000000000000 (Only present in Homo sapiens) over the species used for the phylogeny produced in Chapter 2, the following reconstruction was provided for species close to the root of the tree.

| From | To | Changed | State |
| :--- | :--- | :--- | :---: |
| root | 1 | No | Absent |
| 1 | Entamoeba <br> histolytica | No | Absent |
| 1 | 2 | No | Absent |
| 2 | Trichomonas <br> vaginalis | No | Absent |
| 2 | 3 | No | Absent |
| 3 | 4 | No | Absent |

Table 4.1: Sample output from DOLLOP.

Clearly the parsimonious reconstruction for this protein would only contain one gain. This gain occurs between the ancestral node immediately preceding Homo sapiens.

The output files from DOLLOP were stored for further use.
In order to process this data and utilise it as input for various tests of correlated evolution two Java objects were defined and implemented.

| TransitionMatrix | Transition |
| :--- | :--- |
| theStates : ArrayList <br> thePositionMap : TreeMap | theStart : String <br> theEnd : String <br> changed : Boolean <br> theState : String |
| calculateClade(Node) <br> createPositionMap0 |  |

Figure 4.3: Class diagram illustrating classes underpinning Dollo analyses.

The main object in the preceding figure is the Transition Matrix object. This object has 2 main attributes.

- The States: This is a list of Transition objects. Transition objects contain the same 4 attributes as shown in Table 4.1
- The Position Map: This is a Tree Map, which contains a position within the tree as a key and the state of a given trait at that position as a value. Thus this attribute can be queried for the state (present or absent) of a given trait at any point in the tree.

The Transition Matrix object also has 2 main operations.

- Calculate clade: This function returns all parts of a tree descended from a given node. Thus if a trait is gained or lost at Node $n$, the function will return the monophyletic group consisting of $n$ and all its descendants.
- Create Position Map: This function traverses the States list and utilises the Calculate clade function to populate the Position map.

These objects underpin all further analyses described in this chapter.

### 4.3.1 Maddison test for correlated evolution

Given the set of the ancestral reconstructed state the Maddison test as described above and in the original work by Wayne Maddison (Maddison 1990) was implemented. The following description of the algorithm utilised is based entirely upon the work presented by Maddison (Maddison 1990). A modification to the test for correlated evolution defined by Maddison (Maddison 1990) to fit the constraints imposed by Dollo parsimony is presented in Section 4.3.2

### 4.3.1.1 Algorithm

Assume two discrete binary characters $A$ and $B$ and a phylogenetic tree $T$ and a set of reconstructed states for $A$ and $B$ for each node $N$ within $T$. Possible states for characters $A$ and $B$ lie within the closed interval $[0,1]$. A gain is defined as a transition from 0 to 1 and conversely a loss is defined as a transition from 1 to 0 .

Define character $B$ as reference trait. Define state $s$ as the relevant state of character $B$. Define subset $k(k \subset T)$ as the area(s) of the tree where $B$ is in state $s$.

Define $W_{\text {root }}(x, y \mid b)$ as the total number of ways to have $x$ gains and $y$ losses of character $A$ over the tree starting at the root node given that state of character $A$ is $b$ at the root of the tree.

Define $B_{\text {root }}(p, q \mid x, y, b)$ as the total number of ways to have $p$ gains and $q$ losses of character $A$ in subset $k$ given $x$ gains and $y$ losses over the entire tree starting at the root node given that state of character $A$ is $b$ at the root of the tree.

The test for correlated evolution is thus calculated by

$$
\begin{equation*}
p(o b s)=\frac{B_{\text {root }}(p, q \mid x, y, 0)+B_{\text {root }}(p, q \mid x, y, 1)}{W_{\text {root }}(x, y \mid 0)+W_{\text {root }}(x, y \mid 1)} \tag{2}
\end{equation*}
$$

Solving Equation 2 provides the probability $p$ (obs) of having $p$ gains and $q$ losses of character $A$ in subset $k$ given a total of $x$ gains and $y$ losses occur over the whole tree under the null hypothesis of no correlated evolution. If gains and losses of character $A$ are in some way dependent on whether character $B$ is in state $s$ then we could expect those gains and losses to be concentrated in subset $k . W_{\text {root }}(x, y \mid b)$ and $B_{\text {root }}(p, q \mid x, y, b)$ are calculated through the use of a dynamic programming approach starting at the tips of the tree and proceeding in a post order fashion (Maddison 1990).

### 4.3.1.2 Calculation of total number of ways of having $\mathbf{x}$ gains and $\mathbf{y}$ losses over the tree

 In order to calculate $W_{\text {root }}(x, y \mid b)$ over the entire tree for a character $A$, a matrix containing the number of ways of having 0 to $x$ gains, 0 to $y$ losses for either potential values of $b$ ( 0 or 1) has to be calculated for each node in the tree.For a leaf node there are 0 ways of having $x$ gains and $y$ losses at the node for all values of $x$ and $y$ which are greater than 0 . There is one way of having 0 gains and 0 losses at a leaf node.

For a non-leaf node $K$ there are four calculations to make. Firstly assume all gains and losses occur post the node's immediate descendants $L$ and $M$ and that the state of character $A$ is 0 . A non-leaf node is only processed after both its descendants have been visited. The number of ways of having $x$ gains and $y$ losses at node $K$ given a state of 0 can be calculated by the expression $\sum_{i=0}^{x} \sum_{j=0}^{y} W_{L}(i, j \mid 0) \times W_{M}(x-i, y-j \mid 0)$. This counting system operates on the principle that for every way of having $i$ gains and $j$ losses on node $L$ there are ( $x-i$ ) gains and $(y-j)$ gains on node $M$. Thus if for example there was 1 gain and 1 loss to distribute over node
$K$ then if both of them occurred post descendent $L$ then no changes would occur post descendent $M$. If only the gain occurred post $L$ then the loss would occur post node $M$.

The second part of this calculation is based on the assumption that one of the changes occurs between $K$ and one of its child nodes for example $M$. Thus as one of the changes has occurred (the change is a gain as the state of the character is of character $A$ at node $K$ is 0 ) the state of character $A$ at node $M$ is now 1 and one of $x$ gains has already occurred. Thus the number of ways to have the remaining number of gains and losses can be calculated by the expression. $\sum_{i=0}^{x-1} \sum_{j=0}^{y} W_{L}(i, j \mid 0) \times W_{M}(x-i, y-j \mid 1)$. The third part of the calculation covers the eventuality that the change happens between $K$ and its other child $L$. Thus the number of ways remaining to have $x$ gains and $y$ losses are calculated by the expression

$$
\sum_{i=0}^{x-1} \sum_{j=0}^{y} W_{L}(i, j \mid 1) \times W_{M}(x-i, y-j \mid 0) . \text { Finally assume changes occur between } K \text { and both of its }
$$

child nodes $L$ and $M$. The states of both nodes will be 1 and there will be two fewer gains to distribute over the remainder of the tree. Thus the fourth part of the calculation is:

$$
\sum_{i=0}^{x-2} \sum_{j=0}^{y} W_{L}(i, j \mid 1) \times W_{M}(x-i-2, y-j \mid 1) .
$$

Summing up the results of these four expressions will provide the number of ways of having $x$ gains and $y$ losses at non-leaf node $K$ given that the state of character $A$ is 0 . The calculation of the number of ways of having $x$ gains and $y$ losses if the state of character $A$ is 1 at node $K$ is a mirror image of the process described above (Maddison 1990).

### 4.3.1.3 Calculation of total number of ways of having $\boldsymbol{p}$ gains and $\boldsymbol{q}$ losses in subset $\boldsymbol{k}$ given $\boldsymbol{x}$ gains and $\boldsymbol{y}$ losses over the entire tree

This calculation of $B_{\text {root }}(p, q \mid x, y, b)$ is very similar to the one described above. As above a matrix containing the number of ways of having 0 to $p$ gains in subset $k, 0$ to $q$ losses in subset $k$ given 0 to $x$ gains and 0 to $y$ losses over the whole tree for either potential values of $b$ ( 0 or 1 ) has to be calculated for each node in the tree.

For a leaf node there are 0 ways of having $p$ gains and $q$ losses in subset $k$ given $x$ gains and $y$ losses overall for all values of $p, q, x$ and $y$ which are greater than 0 . There is one way of having 0 gains and 0 losses in subset $k$ given 0 gains and 0 losses overall.

As above a non-leaf node is only processed when both its children have been visited. For a non-leaf node $K$ with character $A$ having state 0 with children $L$ and $M$ there are again
four calculations to be made. The first calculation counts the possibilities where both changes occur post the child nodes. This number is calculated through the expression

$$
\sum_{i=0}^{x} \sum_{j=0}^{y} \sum_{f=0}^{p} \sum_{g=0}^{q} B_{L}(f, g \mid i, j, 0) \times B_{M}(p-f, q-g \mid x-1, y-j, 0) \text {. The second calculation counts }
$$ the possibilities where one of the changes occurs between node $K$ and node $M$. Whether this change is counted as within subset $k$ depends on whether node $M$ lies within subset $k$. To facilitate calculation (Maddison 1990) defined a number $Z_{M}$ as set to 1 if $M$ lies within $k$. Thus the second calculation is evaluated by the expression

$$
\sum_{i=0}^{x-1} \sum_{j=0}^{y} \sum_{f=0}^{p-Z_{m}} \sum_{g=0}^{q} B_{L}(f, g \mid i, j, 0) \times B_{M}\left(p-f-Z_{M}, q-g \mid x-i-1, y-j, 1\right) \text {. The third calculation }
$$

counts the possibility of one of the changes occurring between node $K$ and node $L$. This is evaluated via the expression

$$
\sum_{i=0}^{x-1} \sum_{j=0}^{y} \sum_{f=0}^{p-Z_{L}} \sum_{g=0}^{q} B_{L}(f, g \mid i, j, 1) \times B_{M}\left(p-f-Z_{L}, q-g \mid x-i-1, y-j, 0\right) \text {. The fourth calculation }
$$

counts the possibilities where changes occur between $K$ and $L$ as well as $K$ and $M$. This is evaluated with the expression:

$$
\sum_{i=0}^{x-2} \sum_{j=0}^{y} \sum_{f=0}^{p-Z_{L}-Z_{M}} \sum_{g=0}^{q} B_{L}(f, g \mid i, j, 1) \times B_{M}\left(p-f-Z_{L}-Z_{M}, q-g \mid x-i-2, y-j, 1\right)
$$

The summation of the solutions of the four expressions yields the total number of ways to have $p$ gains and $q$ losses of character $A$ within subset $k$ given $x$ gains and $y$ losses of character $A$ overall under node $K$ given the state of character $A$ is 0 . As above this process is mirrored when the state of character $A$ is 1 (Maddison 1990).

### 4.3.1.4 Permutation effects

The Maddison test for correlated evolution (Maddison 1990) is potentially susceptible to two effects in the context of examination of protein phylogenetic profiles. Maddison's test was designed to test specific hypotheses about correlated evolution. For example one of the first applications of the test was on data testing the association of gregariousness in butterflies with unpalatable larvae (Sillentullberg 1988; Maddison 1990). Thus in a pairwise comparison of characters one character is held static as a reference while the location of changes in the other dynamic character are examined over the tree. The terms static and dynamic shall be used in this context in all subsequent references.

In the case of examinations of correlated evolution in phylogenetic profiles however it is not possible to state whether we are testing for the dependence of the distribution of protein $A$ with the state of protein $B$ or vice versa. The first effect is thus permutation.

The second effect is based on how subset $k$ is defined. As phylogenetic profiles compare patterns of presence and absence of genes subset $k$ can either be defined as the presence of protein $B$ or the absence of protein $B$.

This second effect is however precluded as defining subset $k$ as the absence of protein $B$ shifts position of the number of changes sought. Consider the tree shown in Figure 4.4. If for example protein $A$ was gained once within the clade containing Species 1 and Species 2 and protein $B$ was present in that clade but no where else within the tree. Thus the ancestral state of $B$ would be reconstructed parsimoniously as shown in Figure 4.5.

If $k$ is defined as the presence of $B$ then the test is investigating the probability of 1 gain within $k$ with 1 gain over the entire tree. If $k$ was defined as the absence of $B$ then the test is investigating the probability of 0 gains within $k$ with 1 gain over the entire tree.

Thus over the sample tree if $k$ is defined as the presence of $B$, then there are 3 ways to have 1 gain of $A$ within $k$. That is 1 on the branch leading to Species 1,1 on the branch leading to Species 2 and 1 on the branch leading to the clade. There are 9 ways of having 1 gain over the entire tree. Thus the probability of 1 gain in $k$ is $\frac{3}{9}$ or 0.33 . If on the other hand $k$ is defined as the absence of $B$ then there are 0 gains within $k$ with 1 gain overall. As there is one gain to be accounted for and this gain can only occur within the clade containing Species 1 and Species 2. Thus as before there are 3 ways of having one gain within that clade and 9 ways of having one gain over the whole tree thus the associated probability remains the same, i.e. 0.33 .

Protein A is
gained within
this clade
excluding the
branch which
leads to
Species 3


Figure 4.4: Sample phylogeny of 5 hypothetical species. The numbers on the tree represent presence and absence of protein $B$. The arrow points out the point post which protein A was acquired.


Figure 4.5: Sample phylogeny of 5 hypothetical species. The black area of the tree corresponds to a Fitch parsimonious reconstruction (Fitch 1971) carried out by Mesquite (Maddison and Maddison 2010) of protein $B$ if protein $B$ has the phylogenetic profile 11000 (where the order of species in the profile is the same as the numerical order of the species). It is also the Dollo parsimonious reconstruction. This black area corresponds to subset $k$ if it is defined as the presence of $B$. Conversely the white area of the tree corresponds to $k$ if it is defined as the absence of $B$.

A further example of this concept can be considered by using the initial example provided in 4.2.2.2 involving our fictional cow like species. In that case the probability of acquiring horns in the presence of valine was calculated as $\frac{1}{6}$. If we were to examine the probability of acquiring horns in the absence of valine then the denominator remains the same and there are 0 gains of horns in $k$ (the area of the tree where the ability to metabolise valine is absent). There is 1 way of having 0 gains of horns. Thus the probability of the observed configuration remains the same.

Thus the results shown in by the Maddison-Dollo test over the training data as described in Chapter 3 were identical with respect to the choice of how subset $k$ is defined.

### 4.3.1.5 Evaluation of Maddison test as heuristic for constrained ML

The Maddison test (Maddison 1990) as described above modified to accommodate the assumptions of Dollo parsimony (Section 4.3.2) was implemented using Java. This entailed writing 3,563 lines of code. The implementation was supplied with the set of reconstructed states for each protein pair and the phylogenetic tree on which they are reconstructed.

In order to remove permutation based effects from the analysis the training dataset as described in Chapter 3 was doubled so it included proteins pairs in both the $A-B$ and the $B-A$ orientations. The Maddison test was run on this expanded training set. Thus each protein pair in the training set had two associated probability scores. The lower of these two scores was selected as the lower the probability of the observed distribution of gains and losses the stronger the evidence for correlated evolution. In order to use this probability as an ascending score the score was defined as 1-p. The test was run with subset $k$ defined as the absence of trait B.

The ability of the test to detect protein interactions was then judged according to the criterion of precision/sensitivity as defined in Chapter 3.

This process was then repeated with $k$ defined as the presence of trait B in order to verify the observation that there was no effect on whether $k$ was defined as the absence or the presence of trait B.

### 4.3.2 Modification of test to match Dollo constraints

The calculation of the null distribution under the standard Maddison approach (Maddison 1990) allows for all sequences and permutations of gains and losses as allowed by Fitch parsimony (Fitch 1971). In order to reconcile the test to the assumptions of Dollo parsimony the test was modified to remove all possibility of a gain following a loss. This was achieved by examining the state of the character under consideration at the root node. If the root node was 0 then the standard test as described in Equation 2 was utilised. However if the state was 1 then the following test was used.

$$
\begin{equation*}
p(o b s)=\frac{B_{\text {root }}(p, q \mid x, y, 0)+B_{\text {root }}(0, q \mid 0, y, 1)}{W_{\text {root }}(x, y \mid 0)+W_{\text {root }}(0, y \mid 1)} \tag{3}
\end{equation*}
$$

If a character is acquired at the root of the tree then no gains can be allowed to occur post a loss thus a is calculated for 0 gains and $y$ losses from the root.

To illustrate this difference consider the following tree.


Figure 4.6: Example tree to illustrate the imposition of Dollo parsimonious constraints on the Maddison test for correlated evolution.

Assume the state of a character $C$ was reconstructed as 1 at the root node of the tree and there was one gain and one loss to be distributed over the tree. If gains were allowed to follow losses then there are 4 ways of having one gain and one loss over the tree. These are:

- A loss between node 2 and node 6 followed by a gain between node 6 and node 7.
- A loss between node 2 and node 6 followed by a gain between node 6 and node 8.
- A loss between node 2 and node 3 followed by a gain between node 3 and node 4.
- A loss between node 2 and node 3 followed by a gain between node 3 and node 5.

However with the added Dollo parsimony constraint there are no ways having one gain and one loss over the tree in Figure 4.6 if the state at the root node is 1.

### 4.3.3 Differential parsimony

The distance as defined in work by Phillip Kensche (Kensche et al. 2008) and reiterated in Section 4.3.1 was implemented and its performance examined over the training data.

### 4.3.4 Dollo-pos/ Dollo-overall

Both measures as described by Barker et al. (Barker et al. 2007) were implemented and examined in the light of the testing data.

### 4.3.5 Test based on logistic regression

In order to test for correlated evolution the reconstructed ancestral states allowed the calculation of potential predictor variables, which bore a correspondence to the transition rate parameters, used by Barker and Pagel (Barker and Pagel 2005; Pagel 1994) as described in Chapter 3. These parameters represent the rates of transition in state for a discrete binary character given a particular state for a second discrete binary character.

Given a single set of reconstructed ancestral states these transitions can be empirically counted over the reconstructed states. For example a protein is lost on a given branch of the phylogeny and a second protein is present on that branch according to the reconstructed states this can be counted as a loss of one protein in the presence of the other.

The Dollo parsimony based reconstructions of each phylogenetic profile contain within them the state of the associated protein at every given point in the tree. It was thus possible to compare the state of any given ancestral branch within the tree for any two given proteins. The Dollo reconstruction data is framed in terms of transitions between two nodes. This meant that it was possible to compare a transition in the state of a given protein with the state of the other protein at the same point in the tree.

In order to use the Dollo reconstructions of each phylogenetic profile as potential predictors of functional interaction using logistic regression the reconstructions had to be framed in terms of being potential predictors of correlated evolution. The possible states that a protein could be in at any given transition in the tree were coded as:

- 0: Absent
- 1: Present
- 2: Lost
- 3:Gained

Each profile was associated with a matrix of transitions, which was constructed using the reconstruction of the ancestral states of the profile over the tree. Pairwise comparisons were then carried out. Thus at each transition point the state of protein $A$ was compared to that of protein $B$.

In order to avoid permutation effects the order in which protein pairs were considered was made redundant. This was performed by framing transitions in terms of changes in proteins as opposed to changes in a particular protein. If for example protein $A$ was lost in the presence of protein $B$ this would be counted as the loss of a protein in the presence of another, not the loss of $A$ in the presence of $B$. Pairwise comparisons were thus carried out at each node of the tree to create the predictors shown in Table 4.2. The lower case $s$ stands for scenario.

| Predictor | Description |
| :--- | :--- |
| $s_{00}$ | A point in the tree where both proteins are absent. |
| $s_{01}$ | A point in the tree where one protein is present and the <br> other is absent. |
| A point in the tree where one protein is absent and the |  |
| other is lost. |  |
| $s_{02}$ | A point in the tree where one protein is present and the <br> other is gained. |
| $s_{03}$ | A point in the tree where both proteins are present. |
| $s_{12}$ | A point in the tree where one protein is present and the |
| $s_{13}$ | A point in the tree where one protein is present and the <br> other is gained. |
| $s_{22}$ | A point in the tree where both proteins are lost. |
| $s_{23}$ | A point in the tree where one protein is lost and the |
| other is gained. |  |

Table 4.2: Description of predictor parameters to be utilised in regression model.

Whether or not two proteins interact is a binomially distributed variable. Models for the calculation of the probability of a dichotomous outcome include:

- Two group discriminant function analysis: given two predictors $X_{1}$ and $X_{2}$ discriminant functional analysis constructs a variable $Z$ which is a linear function of $X_{I}$ and $X_{2}$. This function is the equation of a line that separates the data under consideration into two groups. One of these groups will have high values for $Z$ and the other will have low values (Sokal and Rohlf 1995). Novel data with measured values for $X_{I}$ and $X_{2}$ can thus be classified by solving the equation for $Z$ (Sokal and Rohlf 1995).
- Logistic regression: logistic regression relates the probability of a successful outcome in this case that of an interaction with estimated coefficients for a set of predictors via the following application of the logistic function (Sokal and Rohlf 1995).

Preliminary trials were carried out on the training data, which found the results of logistic regression and a linear discriminant function to be broadly similar. However the values of predictors associated with gains are not distributed normally as the number of gains is restricted to 1 . In such cases logistic regression is the recommended technique (Lei and Koehly 2003) as it makes no assumptions of normality regarding the distribution of the predictor variables. Thus logistic regression was selected as an appropriate method of testing whether the predictor variables contribute to the outcome of interaction as well as determining the degree to which they contribute. Logistic regression is carried out via an application of the logistic function, which can be defined as shown in Equation 4

$$
\begin{equation*}
p(\text { Interaction })=\frac{e^{a+b x}}{1+e^{a+b x}} \tag{4}
\end{equation*}
$$

Where $a$ is the $y$-intercept of a regression line, $e$ is the base of the natural logarithm and $b$ is the coefficient of a predictor variable $X$ for a set of predictors ( $X_{0,}, X_{1}, X_{2} \ldots \ldots \ldots . X_{i}$ ) . The logistic function, which is also known as the sigmoid function returns a value within the closed interval $[0,1]$ for values in the range of real numbers from $-\infty$ to $+\infty$.

This probability is converted into the odds of an interaction versus no interaction with the expression: $\frac{p}{(1-p)}$ where $p$ is equal to the probability of an interaction (Sokal and Rohlf 1995).

Solving the expression substituting Equation 4 as a value for $p$ yields the following equation (Sokal and Rohlf 1995).

$$
\begin{equation*}
\frac{p}{(1-p)}=e^{a+b x} \tag{5}
\end{equation*}
$$

Finally the odds of an interaction are converted into the log odds or logit of an interaction via Equation 6 (Sokal and Rohlf 1995):
$\ln \left(\frac{p}{1-p}\right)=a+b X$

The optimal values for the coefficients and intercept of an optimal regression line are estimated by maximum likelihood (Sokal and Rohlf 1995). The full positive training set of 9,161 protein pairs was used as examples of proteins, which interact. A random subsample of 9,161 proteins was then selected from the negative training set as examples of proteins, which do not interact. The size of negative and positive sets were set to be equal to allow the linear model to create a regression line which matched the distribution of the predictors in both sets rather than being biased by a larger negative set.

Counts were then calculated for each parameter for each pair of profiles within the new training dataset. The statistical package R (R Development Core Team 2011) was then used to fit a generalised linear model between the two binary variables using a binomial (logit) link function. The predictor variables were considered to be continuous.

Predictor variables $s_{13}, s_{22}, s_{23}$ and $s_{33}$ were found to cause singularities within the model. $s_{13}$ was found to be perfectly correlated with $s_{03}$ and $s_{33}$ as Dollo parsimony only allows one acquisition of a complex trait. $s_{22}$ and $s_{33}$ only occur rarely as seen below.

| Minimum | 1st Quartile | Median | Mean | 3rd Quartile | Maximum. |
| :--- | :--- | :--- | :--- | :--- | :--- |
| 0 | 0 | 0 | 0.5259 | 1 | 14 |

Table 4.3: Descriptive statistics for counts of predictor $s_{22}$.

| Minimum | 1st Quartile | Median | Mean | 3rd Quartile | Maximum. |
| :--- | :--- | :--- | :--- | :--- | :--- |
| 0 | 0 | 0 | 0.1 | 0 | 1 |

Table 4.4: Descriptive statistics for counts of predictor $s_{33}$.

The predictor $s_{23}$ never occurs at all within the data. The results of the initial regression are shown in Table 4.5.

| Predictor | Coefficient | $p$ value $<0.05$ |
| :---: | :---: | :---: |
| $\mathrm{S}_{00}$ | 0.02179 | 0.1278 (Not |
|  |  | significant) |
| $\mathrm{S}_{01}$ | 0.02607 | 0.0845 (Not |
|  |  | significant) |
| $\mathrm{S}_{02}$ | 0.03775 | 0.0244 |
| $\mathrm{S}_{03}$ | -0.24041 | $2.66 \times 10^{-05}$ |
| $\mathrm{S}_{11}$ | 0.06446 | $5.46 \times 10^{-5}$ |
| $\mathrm{S}_{12}$ | 0.04350 | 0.0189 |

Table 4.5: Coefficients of the initial logistic regression equation.

After removing all predictors that caused singularities as well as all insignificant predictors the analysis was repeated. This led to the following coefficients as shown in Table 4.6.

|  | Predictor | Coefficient |
| :--- | :---: | :--- |
| $\mathrm{s}_{02}$ | 0.019429 | $2.05 \times 10^{-7}$ |
|  | -0.177835 | 0.00115 |
| $\mathrm{~s}_{03}$ | 0.043359 | $<2 \times 10^{-16}$ |
| $\mathrm{~s}_{11}$ | 0.018787 | 0.00107 |
|  |  |  |

Table 4.6: Coefficients of the logistic regression equation derived via examination of the reduced set of predictors.

The full equation of the regression line is shown below as Equation 7.

$$
y=0.019429_{s 02}-0.177835_{s 03}+0.043359_{s 11}+0.018787_{s 12}-0.791849
$$

$y$ is equal to the logit score of the probability of two proteins interacting versus the probability of them not interacting. The logit scores were then transformed into a probability of interaction via an application of the logistic function. This probability was used as the score.

The predictor, which contributes the most to the probability of an interaction, is $s_{03}$. This would suggest that a protein being gained in the absence of the other is indicative of the two proteins not being functionally linked (as the coefficient is negative). The other significant terms, which contribute positively to the probability of an interaction, are $s_{02}, s_{11}$ and $s_{12}$. Losses appear to be a defining event for correlated evolution whether a loss in the presence of a protein or a loss in the absence of a protein. This could potentially be confirmation of work postulating that gene loss is relatively the most important event shaping gene content and determining phenotype. However this may be attributable to the use of Dollo parsimony for ancestral state reconstruction. A loss of a protein in the presence of another might suggest some form of redundancy-based loss. A loss of a protein in the absence of another on the other hand might suggest a cascade of losses of a group of proteins that carry out a particular function that is no longer needed in a particular lineage.
$\boldsymbol{s}_{11}$ is the final significant predictor which implies that two proteins coexisting for periods of time at different points in the phylogeny are also more likely to be functionally linked. This predictor can be thought of as similar to the method used by Cokus (Cokus et al. 2007) except whereas while that work carried out a horizontal comparison of the distribution of presence and absences of proteins across a set of genomes clustered by similarity, the predictor $s_{l l}$ measures co-occurrence both horizontally across species and vertically over a set of putative ancestors as reflected by the phylogenetic tree.

### 4.3.5.1 Evaluation of logistic regression as a heuristic for constrained ML

In order to examine the efficacy of the derived regression equation it was applied to the full training dataset. The performance of the method was then evaluated using through the calculation of precision/sensitivity.

### 4.4 Results

The results for the five tests carried out were measured in terms of precision and sensitivity over the training data. The intersection of the predictions made by the tests with the 6 predictions made by constrained ML method (Barker et al. 2007; Barker and Pagel 2005) at its optimum rate of gain 0.025 and its optimum likelihood ratio cut-off of 58.54 over the training dataset was also examined. As pointed out in Chapter 3 this combination of rate of gain and likelihood ratio cut-off yielded 5 predictions all of which were true positives.

Maintenance of this intersection is judged to be a key criterion for any data filter to form part of a heuristic approach. Thus in order to use a test as a data filter for constrained ML (Barker et al. 2007; Barker and Pagel 2005) the highest acceptable cut-off for the test was considered to be the point at which all 5 predictions made by the ML method were preserved.

### 4.4.1 Maddison test for correlated evolution



Figure 4.7: Performance of the Maddison test for correlated evolution on the training data with $k$ as the absence of trait B. The figure shows the performance between cut-offs of 0.9999 and 1 rising by increments of $1 \times 10^{-7}$.


Figure 4.8: Performance of the Maddison test for correlated evolution on the training data with $k$ defined as the presence of trait B. The figure shows the performance between cut-offs of 0.9999 and 1 rising by increments of $1 \times 10^{-7}$.

The Maddison test for correlated evolution (Maddison 1990) performs reasonably well on the training data as can be seen in Figure 4.7. There is a maximum intersection of 1 in this range of cut-offs with the 5 predictions made by constrained ML (Barker et al. 2007; Barker and Pagel 2005) at its optimum rate of gain and optimum LR cut-off. It is however a computationally intensive process with each node in the tree being evaluated for all combinations of states, gains and losses respectively firstly for the calculation of the null distribution and secondly for all combinations of gains within subset $k$, losses within subset $k$, gains over the whole tree, losses over the whole tree and states. Binary tree traversal is carried out in linear time (Felsenst.J 1973). The amount of time taken as the number of gains and losses to be accounted for increase however rises at a much steeper rate (Maddison 1990).

The mirror property of the algorithm, i.e. that the calculations mirror each other for any given combination of states (Maddison 1990) cannot be utilised in this particular case as gains are restricted to 1 by the use of Dollo parsimony.

### 4.4.2 Differential parsimony



Figure 4.9: Precision and sensitivity measure of differential parsimony over training data.

Differential parsimony (Kensche et al. 2008) was not very successful on the training data as shown in Figure 4.9. It reached a maximum precision of 0.22547 with a sensitivity of 0.023 . There was an intersection of 3 at this point with the predictions made by ML constrained at its optimum rate of gain and LR cut-off.

### 4.4.3 Dollo-pos



Figure 4.10: Precision and sensitivity measure of Dollo-pos over training data. Range of cutoffs at which predictions are made: 0 tol4.

Dollo-pos (Barker et al. 2007) was fairly successful over the training data with a maximum precision of 0.57 at a cut-off of 13 as shown in Figure 4.10. The sensitivity at this point was 0.00043 . There was an intersection of 2 with the 6 predictions made by constrained ML (at its optimum rate of gain and LR cut-off) at this cut-off. In order to use Dollo-pos as a potential data filter the lowest acceptable cut-off that maintained an intersection with all 6 predictions made by constrained ML was 1 . At this cut-off the precision of Dollo-pos was 0.105 and the sensitivity was 0.389 .

### 4.4.4 Dollo-overall



Figure 4.11: Precision and sensitivity measure of Dollo-overall over training data. Range of cut-offs at which predictions are made: -23 to 14 .

The results for Dollo-overall (Barker et al. 2007) were found to be broadly similar to Dollo-pos (Barker et al. 2007) as seen in Figure 4.11. However the effective range of cut-offs for Dollo-overall is shifted down. The maximum precision achieved was 0.5 at a cut-off of 11. The sensitivity at this point was 0.00002 . The lowest cut-off at which the intersection with the 6 predictions made by constrained ML (at its optimum rate of gain and LR cut-off) was maintained was -21. At this cut-off the precision of Dollo-overall was 0.08074 and the sensitivity was 0.99 .

### 4.4.5 Logistic regression

Logistic regression performed well on the training data as can be seen below in Figure 4.11.


Figure 4.12: Precision and sensitivity measure of logistic regression over training data. Range of probability cut-offs at which predictions are made: 0 to 1 . Cut-offs were incremented by 0.001 .

The maximum precision achieved by logistic regression was 0.736 at a sensitivity of 0.01 . This was achieved at a probability cut-off of 0.967 . The lowest probability cut-off at which the intersection with predictions made by constrained ML was maintained was 0.85 . The precision achieved at this point was 0.479 and the sensitivity was 0.0598 . The method that achieved the highest precision while maintaining a full intersection with the predictions made by constrained ML was thus logistic regression.

In order to cross validate the ability of the logistic regression based filter to discriminate between proteins that interact and those that do not, the filter was run on the testing data. Figure 4.13 shows the performance of the filter over the testing data.


## Precision

Figure 4.13: Precision and sensitivity measure of logistic regression over testing data. Range of probability cut-offs at which predictions are made: 0 to 1 . Cut-offs were incremented by 0.001 .

As the $\boldsymbol{s}$ predictors used to determine the logit score used in the logistic regression filter are based on the transition rate parameters used to detect correlated evolution by the constrained ML technique (Barker et al. 2007; Barker and Pagel 2005; Pagel 1994) it was expected that there is a correlation between a high logit score for a protein pair and a high LR
(likelihood ratio statistic) score using the exemplar rate of gain elucidated in Chapter 3 (0.025). The distribution of LR scores is extremely skewed towards the lower end as can be seen in Chapter 3. All true positive protein pairs detected by constrained ML (Barker et al. 2007; Barker and Pagel 2005) at the optimum rate of gain lie within the $99^{\text {th }}$ percentile of LR scores. This is due to the fact that many protein pairs within the training dataset do not display little or no evidence of correlated evolution, i.e. the low sensitivity of the method as shown in Chapter 3.

It is only in the upper ranges of the LR scores that protein pairs that interact are distinguished from those that do not interact. This phenomenon was also observed by Barker as well as by Kensche (Barker et al. 2007; Kensche et al. 2008). Thus in order to display the relationship between the two prediction systems the logit derived probability scores of protein pairs with corresponding LR scores that lay in the $95^{\text {th }}$ percentile of the distribution of LR scores were selected. This came to a set of 5,658 protein pairs.

The logit derived probability scores of these pairs were plotted against their corresponding LR score. Figure 4.14 shows the logit derived probability scores of these 5,658 proteins pairs plotted against their LR scores. The relationship between the values is displayed via an overlaid regression line. There is a large amount of scatter around the regression line however the relationship between the two variables is found to be significant ( $p$ value $<0.001$ ).


Figure 4.14: Linear regression line (Adjusted $\mathrm{R}^{2}=0.1049$ ) drawn over a plot of logit derived probability scores against likelihood ratio statistics over the training data. Vertical dotted line shows optimum cut-off for likelihood ratio statistic. Horizontal dotted line shows optimum cut-off for the logit derived probability score.

Figure 4.14 shows that there is a positive relationship between the LR score generated by constrained ML (Barker et al. 2007; Barker and Pagel 2005) and the logit derived probability scores generated by application of Equation 6 to the set of reconstructed ancestral states.

Figure 4.15 shows that a similar relationship is observed over the testing data set ( $p$ value $<0.001$ ).


Figure 4.15: Linear regression line (Adjusted $\mathrm{R}^{2}=0.0995$ ) drawn over a plot of logit derived probability scores against likelihood ratio statistics over the testing data. Vertical dotted line shows optimum cut-off for likelihood ratio statistic determined by the training data. Horizontal dotted line shows optimum cut-off for the logit derived probability score determined by training data.

### 4.4.6 Hamming distance

In order to further improve the quality of the heuristic hamming distance was applied to the set of predictions generated at each logit score cut-off. The goal of the additional filter was to reduce the size of the search space while still preserving the 6 true positive predictions made by constrained ML.

This yielded an unexpected result. The expectation with the application of Hamming distance was that as the Hamming distance increased the number of true positives predicted actually went up. This would suggest that as evidence of correlated evolution goes down, the probability of predicting a protein-protein interaction goes up. This anomalous result is probably due to the fact that Hamming distance does not take in the phylogenetic relationships between the organisms in the study as pointed out by Barker (Barker and Pagel 2005).

### 4.5 Discussion

The Maddison-Dollo test proved not to be an effective filter as evidenced by its speed. The method achieved a high level of accuracy however it did not maintain an intersection with constrained ML (Barker et al. 2007; Barker and Pagel 2005) at its higher cut-offs.

Differential parsimony (Kensche et al. 2008) was unable to differentiate between the negative and positive training data to an adequate degree.

Dollo-pos and Dollo-overall (Barker et al. 2007) were both able to differentiate between the negative and positive data however they did not preserve the intersection with constrained ML (Barker et al. 2007; Barker and Pagel 2005) at a reasonable level of precision/sensitivity.

The application of logistic regression on the other hand is a viable filter for constrained ML analysis (Barker et al. 2007; Barker and Pagel 2005) as evidenced by both its relative speed as well as its relationship with the LR scores provided by constrained ML (Barker et al. 2007; Barker and Pagel 2005).

The filter utilising logistic regression on a single set of reconstructed ancestral states as implemented above is quick as compared to both constrained ML (Barker et al. 2007; Barker and Pagel 2005) as well as the Maddison test for correlated evolution (Maddison 1990). The comparative CPU time taken by each method over the training data as measured by the Mac OS X utility time is given in Table 4.7.

| Process | Duration: Minutes/Hours |
| :--- | :--- |
| Maddison Dollo test (Farris | 3988 minutes and 20 |
| 1977; Felsenstein 1989; | seconds/ 66 hours and 28 <br> maddison 1990) |
| minutes (approximately). |  |
| Logistic regression using | 3 minutes and 48 seconds. |
| Dollo parsimony |  |
| reconstructions (Farris 1977; |  |
| Felsenstein 1989) |  |
| Constrained ML (Barker and | 6624 minutes and 37 |
| Pagel 2005;Barker et al. | seconds/ 110 hours and 24 |
| 2007) | minutes. |

Table 4.7: Times taken by each of the three methods on the training data. The time given for the Maddison Dollo test is an extrapolation from a run on $12.5 \%$ of the training data. This $12.5 \%$ was selected randomly. Times given in minutes are rounded to the nearest second. Times given in hours are rounded to the nearest minute. All tests were run on an Intel Xeon 3.0 GHz processor.

The reduction in potential protein pairs to be investigated via an application of the logistic regression filter using the probability score cut-off of 0.85 is $111,902(113,132-$ 1132). This is a reduction of $98.9 \%$. As the filter discriminates between proteins, which show evidence of correlated evolution, the remaining $1.1 \%$ should be enriched for proteins amenable to investigation via phylogenetic profiling using ML reconstructions with constrained rates of gain (Barker et al. 2007; Barker and Pagel 2005). Thus an application of the filter to the full human genome followed by an application of phylogenetic profiling using constrained ML (Barker et al. 2007; Barker and Pagel 2005) will potentially yield a large set of interactions from within the human genome some of which may be novel.

All code implementing the procedures described in this chapter is available on request from the author.

## Chapter 5

## Genome-wide prediction of protein functional interactions in humans using a heuristic approach

### 5.1 Introduction

The interactome of an organism can be defined as the complete set of molecular interactions that occur within its full complement of cell types (Yu et al. 2008). This study focuses on the elucidation of interactions between proteins (both direct and indirect) in the human proteome (PPIs). PPIs have been defined as physical interactions between proteins (De Las Rivas and Fontanillo). PPIs are detected by methods such as the yeast 2-hybrid and tandem affinity purification coupled to mass spectrometry (TAP-MS) (De Las Rivas and Fontanillo 2010) as well as co-immunoprecipitation (De Las Rivas and Fontanillo 2010). Interactions between proteins that are indirect can be detected by gene co-expression as investigated in Chapter 3 or techniques like double mutant synthetic lethality (De Las Rivas and Fontanillo 2010). Indirect protein interactions are also detected by TAP-MS as proteins that share membership of a complex do not necessarily maintain a direct physical interaction. Computational interaction detection methods as described in Chapter 1 can contribute to this effort by pointing out putative interactions, which can then be further verified. This chapter describes the application of the logistic regression-based data filter developed in Chapter 4 in combination with constrained ML (Barker et al. 2007; Barker and Pagel 2005) phylogenetic profile analysis to detect potential novel protein-protein interactions as well as novel indirect interactions between proteins.

### 5.1.1 PPI databases

As experimental data has accumulated on protein-protein interactions there have been a number of attempts to organise and annotate accumulated data on PPIs. There are thus a number of databases, which contain data on human protein-protein interactions. As any attempt to examine the quality of predicted PPIs comparison with known data, a brief overview of the major PPI databases is presented below.

### 5.1.1.1 MIPS

MIPs (Mammalian Protein-Protein Interaction Database)(Pagel et al. 2005), is a PPI database which contains 1,812 experimentally verified human protein-protein interactions (PPIs). It
only includes published data from individual experiments as opposed to large scale highthroughput surveys (Pagel et al. 2005).

### 5.1.1.2 BIND

BIND (Biomolecular Interaction Network Database) contains data on three main interactions types (Bader et al. 2001). These are binary interactions, molecular complexes and pathway data (Bader et al. 2001).

### 5.1.1.3 MINT

MINT (Molecular INTeraction) in contrast to MIPS contains data from large scale highthroughput experiments (Chatr-aryamontri et al. 2007). As of 2009 it contains data derived from more than 19,000 experiments and 25,105 curated human PPIs (Ceol et al. 2010).

### 5.1.1.4 INTACT

IntAct is one of the larger PPI databases containing over 200,000 curated binary protein interactions (Aranda et al. 2010). IntAct follows an extremely specific curation process with information from experiments being recorded in high detail using a number of controlled vocabularies to facilitate further data analysis (Aranda et al. 2010).

### 5.1.1.5 HPRD

The HPRD (Human Protein Reference Database) is a human specific PPI database. There are currently 45,207 interactions held in the HPRD (Prasad et al. 2009). It contains manually curated data on protein interactions derived from both high throughput surveys as well as single experiments (Prasad et al. 2009).

### 5.1.1.6 DIP

The DIP (Database of Interacting Proteins) is one of the earlier PPI databases. It contains data derived from manual curation of the literature as well as from structural information on complexes derived from the PDB (Protein Data Bank) (Salwinski et al. 2004).

### 5.1.1.7 REACTOME

The REACTOME database holds data on PPIs in the context of the biological pathways that underpin cellular processes and is also manually curated (Haw et al. 2011).

### 5.1.1.8 STRING

The STRING database holds data on PPIs that are experimentally verified and also adds a set of computationally predicted PPIs (von Mering et al. 2005). It contains PPI information on 630 organisms (Jensen et al. 2009). The total number of interactions held by STRING exceeds 50,000,000 (Jensen et al. 2009).

### 5.1.1.9 I2D

The I2D (Interologous Interaction) database contains the full literature derived predictions from the databases HPRD, BIOGRID, InTact, BIND and MINT as well as computationally predicted interactions (Brown and Jurisica 2005). The sources of evidence utilised for computational predictions include domain co-occurrence, gene co-expression and intersection of GO terms. I2D contains 133,250 unique entries for detected protein interactions between 13,490 proteins.

### 5.1.1.10 KEGG

KEGG as mentioned in Chapter 1 localises gene products within functional pathways (Kanehisa 1997; Kanehisa et al. 2006). This is similar to REACTOME.

### 5.1.1.11 BIOGRID

The BIOGRID database also contains curated data. It has 49,378 interactions involving human proteins (Stark et al. 2006).

### 5.1.1.12 Discussion

There is an overlap between these databases as they are all based on examination of similar experimental data (De Las Rivas and Fontanillo 2010). Given that current estimates of the human interactome size are around 650,000 including non-direct functional interactions (Stumpf et al. 2008) there are still a large number of interactions still to be characterised.

### 5.1.2 Power law

In order to examine the statistical properties of PPI networks, these networks are usually analysed as graphs (Jeong et al. 2001). An interesting observation of the degree distribution within some of these graphs (the degree of a vertice within a graph is the number of edges connected to that given vertice) appear to follow a power law (Jeong et al. 2001). That is the number of vertices within a graph with degree $k$ is approximately $k^{-x}$ where $x$ is a constant (Alon 2007). What this entails is that for any given protein within the PPI network the probability of having a large degree (many interactions) is low. There will however be
proteins within the network that will have a large number of interacting partners. These proteins have been referred to as "hubs" (Han et al. 2004). It has been hypothesised that there are two forms of protein hub (Han et al. 2004). These are "date" hubs, which interact with different partners at different times, and "party" hubs, which interact with multiple partners simultaneously (Han et al. 2004).

Networks with similar degree distributions have been observed in both natural and man-made networks such as the neural arrangements of C. elegans and the power grid of the western United States (Watts and Strogatz 1998).

In the case of PPI networks however as there is a clear physical limit to the number of interacting partners that a given molecule can interact with, the power law distribution over a PPI network will sharply decay at the upper ends of the distribution, as hub proteins reach saturation point with a given number of interaction partners. Similarly the lower end of the distribution may not match a power law, as the cellular environment and other physiochemical factors will affect the probability of being an entirely monogamous interacting partner in a binary interaction. Examples of PPI networks that do not exhibit a power law in degree distribution have been pointed out in the literature, e.g. in work by Tanaka (Tanaka et al. 2005).

### 5.2 Methods

The first step in carrying out a full genome-wide survey was to develop a list of all possible ordered pairs of proteins within the version of RefSeq (Pruitt et al. 2005) used. This came a total of to $560,237,601$ pairs. The logistic regression-based filter implemented in Chapter 4 was applied to the ordered pairs of profiles at its optimum probability cut-off of 0.85 . Removing all pairs that scored beneath this threshold resulted in a total of $5,312,880$ pairs of proteins. This was a reduction of approximately $90 \%$ of the total search space. This set of reduced profile pairs was then analysed by constrained ML (Barker et al. 2007; Barker and Pagel 2005) with the rate of gain parameters restricted to the optimum rate of 0.025 . This analysis was carried out using a cluster consisting of 2602 GHz dual core Opteron 270 processors.

The results of the constrained ML analysis were then filtered for pairs with a likelihood ratio (LR) statistic score of higher than 58.54 (this was the optimum LR score determined in Chapter 3). This led to a set of 20,605 predicted interactions between protein
pairs, consisting of 2,188 individual proteins. In order to examine predicted interactions between members of the same orthologous group predictions were then converted to predictions between orthologous groups. These orthologous groups were identified as the groups clustered by the Inparanoid (Remm et al. 2001) implementation described in Chapter 2 resulting in a predicted set of 7,150 interactions between orthologous group pairs, consisting of 1,417 individual orthologous groups.

### 5.2.1 Short Branch filtration

Examination of the distribution of interactions amongst the individual proteins showed that some of the individual proteins were predicted to have an extremely large number of interacting partners. The maximum number of interactions partners was predicted for the protein with RefSeq GI number 148613856 (described as "probable ATP-dependent RNA helicase DDX17 isoform 3" on the NCBI website). This was predicted to have 1,503 interactions. However given the overall distribution of interaction partners within the set of predictions as shown below in Figure 5.1 these extreme numbers seem to be implausible.

Distribution of predicted interaction counts


Figure 5.1: Distribution of number of predicted interaction partners/protein in constrained ML predictions.

Thus the profiles of these highly connected proteins were investigated.

Another protein with the RefSeq GI numbers 29029591 (labelled "putative ribosomal RNA methyltransferase 1 isoform $b$ " on the NCBI website) was predicted to take part in 1,430 interactions. The phylogenetic profile of this protein was:

001101001001010001111010100010000001001000010001111010
This translates to these proteins being present in Ashbya gossypii, Aspergillus fumigatus, Bombyx mori, Caenorhabditis elegans, Canis familiaris, Cryptococcus neoformans, Debaryomyces hansenii, Dictyostelium discoideum, Drosophila melanogaster, Drosophila pseudoobscura, Entamoeba histolytica, Homo sapiens, Magnaporthe grisea, Paramecium tetraurelia, Plasmodium knowlesi, Schizosaccharomyces pombe, Theileria annulata, Theileria parva, Trichomonas vaginalis, Trypanosoma brucei and Ustilago maydis. This is an extremely unbalanced distribution over the tree as illustrated in Figure 5.2.


Figure 5.2: Distribution of protein labelled "putative ribosomal RNA methyltransferase 1 isoform b". The character 1 indicates presence of the orthologous group.

It was hypothesised that this unbalanced distribution of this profile contributed to its display of a high likelihood ratio statistic (LR) score with a large number of proteins. In particular it was hypothesised that the profiles of prediction heavy proteins might contain losses on the branches leading to $P$. troglodytes and $M$. mulatta. As the branches leading to these taxa are short (see Chapter 2) this may contribute to spuriously high LR scores. In order to investigate this hypothesis a list of RefSeq Gis for proteins lost on either the branch leading to P. troglodytes or the branch leading to M. mulatta was sifted from the overall set of phylogenetic profiles.

The following procedure was then applied iteratively.

- Set cut-off to 0 .
- Select all protein Gis from predicted interactions where no. of predicted interactions for $>$ cut-off.
- Examine intersection of Gis in selected list with set of Gis of proteins lost on short branches.
- Increment cut-off by 1 .

At the point where the cut-off was equal to 298, the intersection between the two sets was $100 \%$ as shown in the Venn diagram below.

$$
\begin{array}{ll}
\text { Gis for proteins lost } & \text { Gis for proteins with }> \\
\text { on short branches } & 298 \text { predicted interactions }
\end{array}
$$



Figure 5.3: Intersection of proteins lost on in P. troglodytes and M. mulatta with proteins with $>298$ predicted interaction partners.

The 16 proteins in this intersection alone accounted for 13,082 of the total predicted protein interactions or $63 \%$. It is impossible to tell whether these proteins are genuinely evolving in a correlated fashion or merely an artefact of the loss on a short branch.

Thus a post-processing step was applied which removed any prediction involving proteins with profiles that matched this pattern. Thus 16,301 proteins with profiles that contained a 0 at either the position representing $P$. troglodytes or the position representing $M$. mulatta were removed from the set of predicted interactions. This led to the removal of 19,463 predicted interactions between proteins. This left a reduced set of 1,142 predictions.

An examination of the training data showed that 2 of the 5 predicted interactions by constrained ML (Barker et al. 2007; Barker and Pagel 2005) at its selected optimum rate of gain ( 0.025 ) during the training step (see Chapter 3 ) involved the protein ( $16936528 /$ NP_001789) which would have been removed by this post processing step. This reduces the sensitivity of the method by $40 \%$.

### 5.2.2 GO term enrichment

A plausible method to examine the potential accuracy of the predicted interaction is to test whether the GO terms associated with the predicted interaction partners are enriched for particular terms.

In order to subject the data to GO (Gene ontology) term analysis (Ashburner et al. 2000) the set of 1,142 predicted interactions between protein pairs was converted to predictions between gene pairs. This was carried out using IDconverter (Alibes et al. 2007). This produced a set of 273 interactions between pairs of genes and 183 individual genes.

In order to investigate the validity of predicted interactions the set of interactions between genes was converted into a network of interactions. The network can be represented as an undirected graph. The graph in this case would be undirected as there is no way of inferring any form of directional relationship between putative predictions.

The predicted interactions were converted into a graph through insertion of the characters ' $x x$ ' between each predicted pair. This converted the predicted interactions into a format known as the simple interaction format, which is usable by a platform known as Cytoscape (Shannon et al. 2003). Cytoscape is a program that allows visualisation and analysis of network data (Shannon et al. 2003) and is widely used for such analyses. The broad structure of the resultant graph is shown in Figure 5.4.


Figure 5.4: Graph of 273 interactions between genes as predicted by the application of constrained ML post data filtering. Each vertex in the graph is one gene. The edges in the graph represent a predicted functional interaction between two vertices.

In order to examine the quality of the predictions the graph was subjected to clique analysis to break up the network into sub-graphs, which are densely connected. The Cytoscape plugin ClusterViz (Cai 2010) was utilised to deconstruct the network into sub clusters. The plugin was used with the FAG-EC agglomerative hierarchical algorithm (Li et al. 2008), which builds up sub-clusters through analysis of the clustering coefficient of each edge in the graph. The clustering coefficient measures the density of connections between a given edge and its neighbours. It does this by calculating the number of triangles that a given edge is part of and dividing this number by the number of triangles that might potentially include it given the degree (number of incoming edges) of its adjacent nodes (Radicchi et al. 2004). FAG-EC was run with a specified cut-off of sub-cliques of at least size 3 . This was because GO terms can be found to be significantly enriched in pairs of proteins even if an annotation is attached to just one protein in the pair.

FAG-EC was also run with a "weak" module definition. This identifies modules as sub-cliques within graphs where the sum of in-degree of each node within a module is higher than the sum of out-degree (Li et al. 2008). The in-degree of a node within an undirected graph is defined as defined as the number of edges connecting it to other nodes in the same subgraph (Li et al. 2008). The out-degree of a node is defined as the number of edges connecting it to the rest of the graph excluding its subgraph (Li et al. 2008).

The application of this algorithm yielded 10 connected sub-cliques. GO term enrichment was examined through the use of the Cytoscape plugin Bingo (Maere et al. 2005). Bingo operates through examination of all GO terms associated with a given network. There are a number of sources of evidence by which a term may be associated with a gene (Ashburner et al. 2000). These are:

- IMP: inferred from mutant phenotype
- IGI: inferred from genetic interaction
- IPI: inferred from physical interaction
- ISS: inferred from sequence similarity
- IDA: inferred from direct assay
- IEP: inferred from expression pattern
- IEA: inferred from electronic annotation
- TAS: traceable author statement
- NAS: non-traceable author statement
- ND: no biological data available
- IC: inferred by curator

In order to utilise reliable sources of evidence terms that were determined using the evidence codes ISS, IEA, NAS and ND were excluded.

Bingo operates by calculating the probability of the association of a given set of terms with a cluster of genes given a background distribution of terms associated with a reference set of genes. This is calculated using the hypergeometric test (Maere et al. 2005). The probability of a given set of genes being associated with a given GO term follows the hypergeometric distribution, which is equivalent to the binomial distribution but utilising sampling without replacement (Sokal and Rohlf 1995). The probability of a cluster $C$ of $r$ genes being associated with a given GO term $g$ (if evaluated against a background set of $N$
genes and assuming the total number of genes associated with $g$ is $t$ ) can be calculated. The background probability of any given gene being associated with $g$ is $\mathrm{t} / N$ and the probability of $g$ not being associated with a gene is $(1-t / N)$. Thus the probability of $x$ genes inside $C$ being associated with $g$ can be calculated using the formula $\frac{\binom{\left(\frac{t}{N}\right)(N)}{x}\binom{\left(1-\frac{t}{N}\right)(N)}{r-x}}{\binom{\frac{t}{N}(N)}{r}}$ where $\binom{k}{Y}$ is the number of combinations of $k$ items taken $Y$ at a time (Sokal and Rohlf 1995).

The effects of multiple testing are reduced through application of the Bonferroni correction. This correction scales the point at which $p$ values are found to be significant down by dividing by $n$ the number of tests performed (Sokal and Rohlf 1995). A procedure involving the hypergeometric test is common in GO enrichment tools and is also utilised by ClueGO (Bindea et al. 2009), Gorilla (Eden et al. 2009) and GOEAST (Zheng and Wang 2008) amongst others.

Bingo was run against a background set of genes, which consisted of the full set of human genes held in Entrez Gene.

### 5.2.3 Intersection with other data sources

To examine the extent to which the predictions made by the filter in combination with constrained ML (Barker et al. 2007; Barker and Pagel 2005) intersected with known data it was decided to compare the predictions to a known set of PPIs. It was decided to utilise the I2D database (Brown and Jurisica 2005) as it contained data from all the other major PPI databases. Thus the Interologous Interaction Database (I2D) version 1.95 was downloaded.

Predictions were converted from RefSeq GI numbers to their corresponding Uniprot (Apweiler et al. 2010) primary accession. Only Swiss-Prot accessions were used, as these are high confidence protein molecules that have been manually annotated (Apweiler et al. 2010).

As mentioned above there were 1,142 predictions made between RefSeq GI pairs. This conversion reduced the set of predictions to 278 predictions as a complete mapping of RefSeq to Uniprot is lacking.

### 5.3 Results

### 5.3.1 GO Enrichment

Table 5.1 shows details of the sub-clusters generated by ClusterViz (Cai 2010) ordered in descending order by size.
Cluster No. No. of genes No. of interactions Genes
$1 \begin{array}{llll}1 & 97 & 188 & \text { PRPF31 RRP9 SNRPE SUPT4H1 TNPO1 }\end{array}$
COPB1 FASN GLRX5 RPS25 WWOX
DHDDS EXO1 H2AFY ATP5C1 RPS29
RER1 PHF5A PIGL RPS21 POLR2G PSMD8
TP53RK ABT1 ANAPC10 TCEA2 NOP10
POLR2L SF3B5 LZTR1 TUBGCP2 CDS1
MAK16 CTDSPL RBM34 KIFC1 GFPT1
PPP1CC UBE2D4 BYSL PSMA6 FDX1L
TFB1M C20orf118 KIAA1609 UBE2V1
NAPNAPB RLBP1 RPF1 PSMC4
TRAPPC6B RBMX2 RHOC TOP2B UBE2I
CDK5 FKBP4 CCT6A CDK7 CKS2 CTDSP1
DIMT1L FAM96B FKBP5 GNB1 GUK1
HSPE1 KIF 19LSM7 POLE2 PSMB1
RIOK2RPL13 RPL19 RPL30 RPL37A
RPS23 SEC11C SLC2A6
SMARCAL1TRAPPC1 TRAPPC4
TRAPPC6ATXNL4B UBE2V2 VBP1 VPS45
ZDHHC21 ERCC2 SPO11 TRIT1 SHMT2
GDPD1 DOLK DUSP5 LIG1 TRMT112

Table 5.1: Sub-cliques of predicted interactions generated through analysis of clustering coefficients.

| Cluster No. No. of genes No. of interactions |  |  | Genes |
| :---: | :---: | :---: | :---: |
| 2 | 4 | 6 | MTMR2 MTM1 |
|  |  |  | MTMR9 MTMR1 |
| 3 | 8 | 10 | VAPB ZNF516 STK17A |
|  |  |  | ZNF225 |
|  |  |  | MAZ ZNF286A ZNF304 |
|  |  |  | DNM1L |
| 4 | 4 | 4 | NLE1 CDC6 TEP1 GEMIN5 |
| 5 | 3 | 2 | ZRSR2 PPP1CB |
|  |  |  | RPL31 |
| 6 | 3 | 2 | DERL1 DERL2 |
|  |  |  | DNAJC12 |
| 7 | 3 | 2 | H2AFV ATG4A |
|  |  |  | UNG |

Table 5.1: Sub-cliques of predicted interactions generated through analysis of clustering coefficients (cont).

To investigate whether GO terms were enriched in the predicted sub-clusters; clusters were subjected to analysis for GO term enrichment using Bingo (Maere et al. 2005). All terms were judged significant at $p<0.05$ after application of the Bonferroni correction. Table 5.2 presents the clusters and the GO terms enriched in each cluster.
Cluster No. Enriched GO terms

1

44238 primary metabolic process
44237 cellular metabolic process
8152 metabolic process
44260 cellular macromolecule metabolic process
43170 macromolecule metabolic process
6414 translational elongation
6412 translation
44267 cellular protein metabolic process
6368 RNA elongation from RNA polymerase II promoter

6354 RNA elongation
10467 gene expression
19538 protein metabolic process
44265 cellular macromolecule catabolic process

No significant enrichment
19224 termination of RNA polymerase II transcription

43653 mitochondrial fragmentation during apoptosis
79 regulation of cyclin-dependent protein kinase activity

31981 nuclear lumen
No significant enrichment
30970 retrograde protein transport, ER to cytosol
30433 ER-associated protein catabolic process
6515 misfolded or incompletely synthesized protein catabolic process

6984 ER-nuclear signaling pathway
30176 integral to endoplasmic reticulum membrane
31227 intrinsic to endoplasmic reticulum membrane
51789 response to protein stimulus

Table 5.2: GO enrichment in sub-cliques within predicted interaction network.

## Cluster No. Enriched GO terms

| 6 | 31301 integral to organelle membrane <br> 31300 intrinsic to organelle membrane <br> 43161 proteasomal ubiquitin-dependent protein catabolic process <br> 10498 proteasomal protein catabolic process <br> 5789 endoplasmic reticulum membrane <br> 42175 nuclear envelope-endoplasmic reticulum network <br> 44432 endoplasmic reticulum part <br> 43632 modification-dependent macromolecule catabolic process <br> 51603 proteolysis involved in cellular protein catabolic process <br> 19941 modification-dependent protein catabolic process <br> 6511 ubiquitin-dependent protein catabolic process <br> 44257 cellular protein catabolic process <br> 30163 protein catabolic process <br> 9607 response to biotic stimulus <br> 6886 intracellular protein transport <br> 19060 intracellular transport of viral proteins in host cell <br> 30581 intracellular protein transport in host <br> 51708 intracellular protein transport in other organism during symbiotic interaction <br> 15031 protein transport <br> 44265 cellular macromolecule catabolic process <br> 45184 establishment of protein localization <br> 43285 biopolymer catabolic process <br> 8104 protein localization <br> 9057 macromolecule catabolic process <br> 46719 regulation of viral protein levels in host cell <br> 33036 macromolecule localization <br> 12505 endomembrane system <br> 6508 proteolysis <br> 46907 intracellular transport <br> 5783 endoplasmic reticulum |
| :---: | :---: |

Table 5.2: GO enrichment in sub-cliques within predicted interaction network (cont).

## Cluster No. Enriched GO terms

| 6 | 44248 cellular catabolic process <br> 31090 organelle membrane <br> 9056 catabolic process <br> 51649 establishment of localization in cell <br> 42288 MHC class I protein binding <br> 51641 cellular localization <br> 42287 MHC protein binding <br> 30307 positive regulation of cell growth <br> 42221 response to chemical stimulus <br> 45793 positive regulation of cell size <br> 65008 regulation of biological quality <br> 19048 virus-host interaction <br> 45927 positive regulation of growth <br> 51701 interaction with host <br> 7242 intracellular signaling cascade <br> 44419 interspecies interaction between organisms <br> 44404 symbiosis, encompassing mutualism through parasitism <br> 6950 response to stress <br> 6810 transport <br> 51234 establishment of localization <br> 22415 viral reproductive process <br> 16032 viral reproduction <br> 1558 regulation of cell growth <br> 51179 localization <br> 8361 regulation of cell size <br> 44267 cellular protein metabolic process <br> 19538 protein metabolic process <br> 40008 regulation of growth <br> 44260 cellular macromolecule metabolic process <br> 51869 response to stimulus <br> 16021 integral to membrane <br> 7154 cell communication <br> 43170 macromolecule metabolic process <br> 30968 endoplasmic reticulum unfolded protein response |
| :---: | :---: |

Table 5.2: GO enrichment in sub-cliques within predicted interaction network (cont).

## Cluster No. Enriched GO terms

| 6 | 31224 intrinsic to membrane |
| :--- | :--- |
|  | 44446 intracellular organelle part |
|  | 44422 organelle part |
|  | 43283 biopolymer metabolic process |
|  | 7165 signal transduction |
|  | 44425 membrane part |
|  | 51706 multi-organism process |
|  | 22414 reproductive process |
|  | 8284 positive regulation of cell proliferation |
|  | 6986 response to unfolded protein |
| 7 | No Significant enrichment. |

Table 5.2: GO enrichment in sub-cliques within predicted interaction network (cont).

### 5.3.2 Intersection with known data

The comparison of the set of protein interactions predicted by constrained ML (Barker et al. 2007; Barker and Pagel 2005) with the I2D database (Brown and Jurisica 2005) was carried out by determining the intersection between the two sets of interactions. There were 2 predictions in common between the two datasets, which were not self-interactions (there were 9 self interactions in the intersection). These were:

Table 5.2: Intersection between I2D (Brown and Jurisica 2005) and predictions by logistic regression/constrained ML (Barker et al. 2007; Barker and Pagel 2005).

There are thus 1,131 predictions made by logistic regression/constrained ML (Barker et al. 2007; Barker and Pagel 2005), which are potentially novel. All predictions made can be seen in Appendix C.

### 5.3.3 Network statistics

The degree distribution of nodes within the graph appears to follow a power law. This could potentially also be indicative of the correctness of the predictions made by constrained ML. This pattern is observed in both the full and the reduced graphs as shown in Figures 5.5 and 5.6 .


Figure 5.5: Degree distribution for full graph of protein interactions. Line is fitted power law of the form $y=a x^{b}$. Line is fitted by least squares regression $R^{2}=0.694$.


Figure 5.6: Degree distribution for graph of protein interactions post short branch filtration. Line is fitted power law of the form $y=a x^{b}$. Line is fitted by least squares regression $R^{2}=0.768$.

### 5.4 Discussion

A full genome wide investigation of human protein interactions by constrained ML (Barker et al. 2007; Barker and Pagel 2005) in combination with the logistic regression-based data filter seems to be a potentially fruitful source of new protein interactions. The enrichment of GO terms in some sub-cliques of the resultant network suggests that the system has an ability to make predictions with some basis in reality and thus a proportion of the set of predictions made are both novel and accurate.

### 5.4.1 GO enrichment

GO enrichment was investigated conservatively by excluding the GO evidence code IEA. This evidence code is associated with $90 \%$ of GO annotations (Buza et al. 2008). However
despite removing terms associated with this code as well as terms associated with the codes ISS, ND and NAS, a reasonable degree of enrichment was still observed.

The terms enriched appear to be associated with processes, which are divergent across eukaryotes such as transcription (enriched in sub-clique 1) (Coulson and Ouzounis 2003). This is a demonstration of the fact that it is only proteins that show a degree of variability in their distribution pattern that are susceptible to this line of investigation.

### 5.4.2 Intersection with known data

The level of intersection with the I2D database is fairly low. Using the estimate of interactome size provided by (Stumpf et al. 2008) and assuming every prediction in I2D (Brown and Jurisica 2005) is correct. This would correspond to a coverage level of $133,250 / 650,000$ or $20 \%$. Thus the probability of any given accurate prediction being within this database would be 0.2 . Thus the converse probability of an accurate prediction not being in the database would be 1-0.2 or 0.8 .

If every prediction made by the heuristic approach were accurate, then the observed result of an intersection of 11 and a complement of 1,131 would be highly improbable $0.8^{1131}$ or $\sim 0$ ). The lack of intersection between the two datasets could be due to the bias in PPI databases to particular physical detection systems such as yeast 2 hybrid. Approximately 37 $\%$ of the binary interactions held in HPRD (Mishra et al. 2006) were detected using yeast 2 hybrid.

The issue of RefSeq to Uniprot mapping is also pertinent in contributing to this lack of intersection as over $75 \%$ of the predictions were lost post mapping.

Finally it is also unlikely that there is $100 \%$ accuracy in all PPIs held in I2D.

### 5.4.3 Weaknesses

Clearly the result of a precision of 1 as achieved on the training and testing data cannot be extended to a full genome wide survey. The fact that predictions are made through comparisons of the phylogenetic distribution of proteins suggests that one weakness of the method could be an inability to distinguish between paralogs/isoforms and proteins showing evidence of correlated evolution. However this issue is far from clear-cut as there is evidence to show that homologous proteins are more likely to interact (Ispolatov et al. 2005; Orlowski et al. 2007). Thus it is possible that the success of the phylogenetic profile method is partly based on this observation. This is a potentially confounding issue for the method. However
examination of interactions between predicted orthologous groups can ameliorate this. In the case of this study of the 1,142 pairs of proteins predicted to be functionally linked by this study 221 lie within the same orthologous group as identified by the Inparanoid implementation (Remm et al. 2001).

Thus predictions between members of orthologous groups are not particularly widespread over the data examined.

The other weakness of phylogenetic profiling in general that applies to this set of predictions is potentially inaccurate profiles. Profiles can be inaccurate for a number of reasons including low coverage sequencing, poor annotation or incorrect assumptions in homolog identification. The short branch filtration step undertaken before further analysis is potentially attributable to this phenomenon.

### 5.4.3.1 Scaling

The precision and sensitivity results observed over the training data were based on a biologically unrealistic ratio of 10:1 of negative to positive examples of interacting proteins. The results observed can be adjusted for the whole genome by scaling to a more realistic ratio. A possibly more realistic ratio can be calculated using estimates of interactome size. These range from 154,000-369,000 (Hart et al. 2006) to 650,000 (Stumpf et al. 2008). If these numbers are subtracted from the size of all potential interactions 560,237,601 (calculated as all possible pairs from version of RefSeq held) estimated ratios of negative to positive range from approximately $860: 1$ to $3636: 1$. Assume for the sake of argument the ratio of $860: 1$ is adopted (via an assumption of an interactome size of 650,000 ). Recall that the size of the positive set in the training data is 9,161 pairs of known interactions. Thus as an illustrative example if a given predictive method yielded a precision of 0.5 and a sensitivity of 0.1 over the training data this would correspond to making 916 predictions of which $50 \%$ were correct ( $\mathrm{TP}=458, \mathrm{FP}=458$ and $\mathrm{FN}=8703$ ). In order to scale the data the following numbers need to be calculated:

- $\quad P(T P)=$ Probability of predicting a true positive.
- $P(F P)=$ Probability of predicting a false positive.
- $\quad P(F N)=$ Probability of predicting a false negative.

These numbers can be calculated by the following equations:

$$
\begin{align*}
& P(T P)=\frac{(T P)}{(P S)}  \tag{1}\\
& P(F P)=\frac{(F P)}{(N S)}  \tag{2}\\
& P(F N)=\frac{(F N)}{(P S)} \tag{3}
\end{align*}
$$

Where $P S=$ size of the positive set and $N S=$ size of the negative set.

For the example above $P(T P)=458 / 9161=0.049, P(F P)=458 / 103971=0.004$ and $P(F N)=8703 / 9161=0.95$. Thus by multiplying these probabilities by the estimated full interactome size (in this case) 650,000 the sizes of $T P$ and $F N$ can be calculated over the full interactome. In order to calculate the size of $F P$ the size of a potential negatome (proteins that do no interact) must be calculated. This can be calculated as the estimated size of the interactome subtracted from the number of all possible interactions (in this case 560,237,601$650,000=559,587,601)$. Given these numbers the values of $T P, F P$, and $F N$ over the full interactome for this example would be $31,850,2,238,216.5$ and 617,500 respectively leading to a scaled precision of 0.014 and a scaled sensitivity of 0.049 over the whole interactome. In cases where precision $=1$ scaling will not affect this value as there are no false positives predicted.

Probabilities of predicted interactions being genuine can also be calculated via an alternate route applying Bayes theorem with the prior probability of an interaction being derived from an estimate of interactome size. Thus applying Bayes theorem the posterior probability of an interaction can be calculated using the following parameters (Yang 2006):

- $\quad P(I)=$ prior probability of interaction. Calculated by division of interactome size estimate by total number of potential interactions.
- $\quad P(P o s)=$ Probability of making any positive prediction. Calculated as
$P($ Pos $)=P($ Pos $\mid I)(P(I))+P($ Pos $\mid \sim I)(P(\sim I))$
In cases where precision is $=1 P(\operatorname{Pos} \mid \sim I)=0$. Note $P(\operatorname{Pos} \mid \sim I)=P(F P)$
- Thirdly the probability of making a positive prediction given an interaction is calculated as: $P(\operatorname{Pos} \mid I)=$ Sensitivity of the method.

Thus the posterior probability of a predicted interaction being genuine can be calculated using Bayes theorem as presented in Equation 4:

$$
\begin{equation*}
P(I \mid \text { Pos })=\frac{P(I) \times P(\operatorname{Pos} \mid I)}{P(P o s)} \tag{4}
\end{equation*}
$$

Bayes theorem however is only applicable in cases where precision $<1$ as the posterior probability is 1 when precision $=1$.
This can be simply demonstrated using basic algebra and recasting the terms.
$P(I \mid$ Pos $)=\frac{(\text { Prior } \times \text { Sensitivity })}{(\text { Prior } \times \text { Sensitivity })+(P(F P)(1-\text { Prior }))}$

Thus as $P(F P)(1-\operatorname{Pr}$ ior $)=0$ the posterior probability is 1 .

### 5.4.4 Conclusions

Given results observed on the training data and testing data and the GO term enrichment observed in the sub-clusters, as well as the results of previous work on the phylogenetic profile method (Barker et al. 2007; Barker and Pagel 2005; Bowers et al. 2004; Cokus et al. 2007; Kensche et al. 2008; Pagel et al. 2004b; Pellegrini et al. 1999; Vert 2002) amongst others, it appears that the method is capable of discerning between proteins that are functionally linked and proteins that are not. Thus the novel predictions made could potentially be genuine interactions, which are of yet uncharacterised.

## Chapter 6

## Conclusions and further work

### 6.1 Summary of Project

The goal of this project has been an investigation into detection of human protein interactions using the comparative method. More specifically the development of a novel heuristic approach to allow application of the effective but computationally intensive constrained ML (Barker et al. 2007; Barker and Pagel 2005) approach to phylogenetic profile analysis on a genome-wide scale. This application was intended to allow the generation of novel predictions of protein interactions.

A database of all against all comparisons of the proteomes of 54 eukaryotic organisms plus 1 archaeon was created. This was used to as input to an implementation of the Inparanoid (Remm et al. 2001) procedure to cluster the contents of the proteomes into orthologous groups. Using the human proteome as a reference point phylogenetic profiles were then constructed for each protein within the human proteome.

10 proteins that were universally present in single copies in all organisms under consideration were then selected through analysis of the phylogenetic profiles and orthologous groups. The versions of these single copy proteins from each species were then aligned to create a multiple sequence alignment. Each multiple sequence alignment was then concatenated to create one single combined alignment. This combined alignment provides a measure of divergence between the 55 organisms under consideration. The concatenated multiple sequence alignment was then used to reconstruct a phylogenetic tree of the 54 eukaryotes under consideration using the archaeon as an outgroup with which to root the tree. This phylogeny was broadly congruent with current thought on eukaryotic evolution (see Chapter 2) .


Figure 6.1: Process flow for research carried out in Chapter 2.

In order to use constrained ML (Barker et al. 2007; Barker and Pagel 2005) to analyse the training data it was necessary to ascertain the optimum rates at which a character could be gained in order to constrain the models of evolution used by the method (Barker et al. 2007). In order to do this it was necessary to obtain training data, i.e. examples of protein pairs that interact and examples of protein pairs that are unlikely to interact. Positive data was acquired which was based protein interactions held with the HPRD database. Negative data was generated by creating a set of all possible pairs of human proteins. These pairs were filtered by removing all pairs that possessed any Gene Ontology (GO) (Ashburner et al. 2000) terms in common. Once these training sets were obtained different rates of protein gain were evaluated in terms of precision and sensitivity and an optimum rate of gain of 0.025 was selected. The highest sensitivity reached by constrained ML at this rate was 1 at a cut-off of 56.37. The sensitivity of the method at this cut-off was 0.000654 .

The efficacy of constrained ML (Barker et al. 2007; Barker and Pagel 2005) in detecting protein-protein interactions was then compared to a comparable high throughput laboratory based method for detecting interactions using the training data. This method was examination of gene co-expression in response to given experimental stimuli as measured by
microarrays. The highest performing microarray experiments also achieved a precision of 1. The highest performing microarray experiment E-MEXP-1224 (Garman et al. 2009) achieved a sensitivity of 0.003 .

Constrained ML (Barker et al. 2007; Barker and Pagel 2005) was also compared to the PIPs server (McDowall et al. 2009) which uses a semi-naive Bayesian classifier (Scott and Barton 2007) in order to evaluate multiple sources of evidence for potential protein interactions. At its highest cut-off the Bayesian classifier achieved a precision of 0.9883721 and a sensitivity of 0.01 .


Figure 6.2: Process flow for research described in Chapter 3.

These comparisons showed that constrained ML (Barker et al. 2007; Barker and Pagel 2005) showed comparable levels of precision to gene co-expression at an optimal level of constraint for rate of gain and outperformed the method that integrated multiple sources of evidence. In terms of sensitivity however constrained ML (Barker et al. 2007; Barker and Pagel 2005) was clearly the worst performer. However given that constrained ML (Barker et
al. 2007; Barker and Pagel 2005) achieved a precision of 1 over the training data it was utilised for further analysis.

The application of constrained ML (Barker et al. 2007; Barker and Pagel 2005) to a full genome-wide survey was found to be impractical due to time considerations. Thus a heuristic was developed which approximated the ability of constrained ML (Barker et al. 2007; Barker and Pagel 2005) to distinguish between proteins that interact and those that do not. This heuristic was based on the reconstruction of ancestral states using Dollo parsimony (Farris 1977) over the phylogenetic tree. Two novel potential heuristics were developed, implemented and tested using the Dollo parsimonious reconstruction. The first was an implementation of a test for correlated evolution which calculates the probability of the concentration of a set of gains and losses of a protein in the areas of a phylogenetic tree where a second protein was either present or absent (Maddison 1990). The second potential heuristic was based on logistic regression using empirical counts of the presence, absence, gain or loss of one protein given the presence, absence, gain or loss of the other as predictor variables.

The Maddison test (Maddison 1990) based heuristic performed reasonably well in its own right as a method of detecting functional interactions. It achieved a maximum precision of 0.857 with a sensitivity of $6.54 \times 10^{-4}$ over the training data at a score cut-off of 0.9999997999999475 . However it proved not to be efficient enough in terms of speed to be justified for use as a heuristic. It also did not maintain an intersection with the 5 predictions made by constrained ML (Barker et al. 2007; Barker and Pagel 2005) at its optimum rate of gain (0.025) and at its optimum likelihood ratio (LR) statistic score cut-off (58.3).

Maintenance of an intersection with these predictions was considered a necessary property of an effective heuristic.

The heuristic that utilised logistic regression achieved a precision of 0.736 with a sensitivity of 0.01 at its optimum cut-off of 0.967 . It also maintained an intersection with the predictions made by constrained ML (Barker et al. 2007; Barker and Pagel 2005) (at its optimum rate of gain and LR cut-off) up to a cut-off of 0.85 . At a cut-off of 0.85 the heuristic made 1,230 predictions, which amounted to a reduction of the search space of potential proteins by $98.9 \%$.

The heuristic based on logistic regression was then applied to the full human genome in order to filter out protein pairs that displayed little or no evidence of correlated evolution. The heuristic reduced the size of the search space by $90 \%$ over the whole genome.


Figure 6.3: Process flow for research carried out in Chapter 4. Note: Validation sets were used to validate all methods. The connectors have been left out for clarity.

Having applied the heuristic to the method a full genome-wide survey was launched. The results of the genome-wide survey found that a large majority of predicted protein interactions involved proteins, which had been lost on short branches in the phylogeny. These predictions were removed from the overall set of predictions. The prediction set was then recast as a network of interactions.

The results of the genome-wide survey were then examined by generating subnetworks from the complete network generated and examining these sub networks for enrichment in Gene Ontology (GO) (Ashburner et al. 2000) terms. GO term enrichment was found in $57 \%$ of the clusters generated. The intersection of the predictions made by constrained ML (Barker et al. 2007; Barker and Pagel 2005) with the I2D database (Brown and Jurisica 2005) was also examined. The intersection with the I2D (Brown and Jurisica 2005) database was low suggesting that any correct predictions generated by this project are
also novel predictions of protein interaction. The genome-wide survey yielded a final set of 1,131 predictions of protein interaction.


Figure 6.4: Procedure for research carried out in Chapter 5.

### 6.1.1 Repeat Analysis

To apply this procedure to a new dataset, the following procedure would have to be followed.
Prerequisites needed:

- Phylogenetic tree for species of interest.
- Phylogenetic profiles for proteins of interest.
- Positive and negative examples of protein interaction data. An automated procedure for the acquisition of training/testing data is found in (Chen et al. 2011).

Having acquired these, the programs BayesTraits (Pagel et al. 2004a) and bms_runner (Barker et al. 2007) should be downloaded.

To determine the optimum rate of protein gain for use in the constrained ML procedure bms_runner should be used to evaluate multiple rates of gain. The LR scores for all proteins for the optimum rate of gain should be kept.

Once this rate is determined, the next step is the ancestral state reconstructions. In order to carry out these reconstructions it will be necessary to download the program DOLLOP held in the PHYLIP package (Felsenstein 1989).

DOLLOP should be run with the -U option, which will allow it to utilise the phylogenetic tree. (Note bms_runner uses a NEXUS formatted tree while DOLLOP will need a PHYLIP format tree). DOLLOP should be run on every profile in the dataset. Thus the end product of this step is a set of ancestral reconstructions over the tree for each profile.

At this point code written by the author (available on request) can be used to process these reconstructions. This code will take in the reconstructions and return a dataset consisting of the $s$ parameters described in Chapter 4 calculated for each protein.

This data can then be processed using standard statistical package R (R Development Core Team 2011) in order to carry out logistic regression. Once regression has been carried out, this should yield a linear equation for calculating a logit based score for the probability of interacting.

Again code available from the author can now be utilised. This code will take in the specified coefficients for the $s$ parameters calculated in R, the Dollo reconstructions of the proteins, the LR scores of the proteins at the optimum rate of gain and the validation data and return the optimum logit cut-off for the data for preserving the performance of constrained ML (Barker et al. 2007; Barker and Pagel 2005).

At this point a dataset of all possible pairs of profiles should be prepared. Code from the author can be used to apply the linear equation to each of these pairs to calculate the logit score. These pairs can now be filtered by the optimum cut-off.

Once a reduced set has been created, constrained ML can be applied to this set (Barker et al. 2007; Barker and Pagel 2005).

### 6.2 Conclusion

This project has investigated use of the comparative method specifically constrained ML (Barker et al. 2007; Barker and Pagel 2005) as a means to detect protein-protein interactions. It has generated a set of predictions that if validated by further laboratory based investigation could contribute to knowledge about the human interactome. It has also developed a method
that allows the application of the computationally intensive constrained ML (Barker et al. 2007; Barker and Pagel 2005) approach to phylogenetic profiling on a genome-wide scale.

The ability of the comparative method to unearth protein interactions can only be enhanced by the current rate of data generation given the rapid uptake of next generation sequencing technologies such as the Roche 454 GS FLX sequencer, the Illumina Genome Analyser and the Applied Biosystems SOLID sequencer, which can generate gigabases of sequence data in a matter of days (Mardis 2008). As more organisms are sequenced the quality of reconstructed phylogenies and consequently the efficacy of the comparative method in detecting associations between traits should improve due to increased taxon sampling (Heath et al. 2008).

Given this increased pace of data generation it is also necessary to develop fast and effective computational techniques for functional annotation of proteins. Detection of protein interactions can be used to functionally annotate proteins via the principle of "guilt by association" (Aravind 2000). Thus the combination of the developed heuristic with constrained ML (Barker et al. 2007; Barker and Pagel 2005) can contribute to annotation efforts. It has been seen that this method is not very sensitive thus the probability of it making any predictions at all for a given protein are low. But used in a high throughput unsupervised context the method is potentially capable of detecting novel interactions as one tool amongst many.

Among the methods of detecting protein interactions examined over the course of this study was the PIPs server (McDowall et al. 2009), which as mentioned above combines multiple sources of evidence in order to detect potential protein interactions (Scott and Barton 2007) utilising a Bayesian classifier. A similar approach of using combined evidence in a Bayesian framework was previously taken by (Jansen et al. 2003). This combination of diverse sources of evidence as a means to elucidate protein interactions has also been applied by Mohamed (Mohamed et al. 2010) utilising a classifier based on a majority vote from a collection of decision trees. Other approaches such as support vector machines and singular decision trees have also been investigated by (Qi et al. 2006).

Potentially the application of constrained ML (Barker et al. 2007; Barker and Pagel 2005) in combination with the heuristic in a genome-wide manner could be utilised as a source of contributory evidence in a similar framework.

### 6.3 Future directions

Constrained ML (Barker et al. 2007; Barker and Pagel 2005) has been seen to be capable of detecting protein-protein interactions at a reasonable level of accuracy. With the data accumulated over the course of this project there are a number of further avenues of investigation and areas of extension.

### 6.3.1 Computational extensions

The procedure followed in order to utilise constrained ML (Barker et al. 2007; Barker and Pagel 2005) for a genome-wide survey of $H$. sapiens involved the use of bespoke scripts and various programs provided by a plethora of authors as cited throughout this text. To facilitate the application of this tool by other users it will be necessary to create an interface and combine the functionality of the programs utilised into one computational procedure.

The construction of phylogenetic profiles for all proteins in all species held in the current dataset and the provision of these profiles online via a web interface would also facilitate this process. The data generated by this project as presented in Appendix D could also be presented via an online database either an extant protein interaction database such as String or I2D or a bespoke database, which would have to be constructed.

### 6.3.2 Consensus profiles

The application of constrained ML (Barker et al. 2007; Barker and Pagel 2005) to detection of protein interactions is carried out in a pairwise fashion. Work by Bowers extended the idea of pairwise comparisons to three way comparisons using Boolean logic operators (Bowers et al. 2004). This method attempted to detect dependencies in the presence and absence of a given gene on the presence and absence of two other genes. A similar technique could be utilised to integrate matching profiles into consensus profiles. By classifying mismatches as missing information consensus phylogenetic profiles could be constructed to represent groups of proteins. The program BayesTraits (Pagel et al. 2004a), which is utilised to apply the constrained ML approach, handles missing data by reconstruction of the missing data as an extension of ancestral state reconstruction. Thus when a plausible reconstruction is reached at the immediate ancestral node of the taxon with the missing data the state of the taxa can be estimated using rate transition parameters (Pagel 1994). Consensus profiles will utilise a mismatch character $X$ to represent missing information. Thus if for example we compare the following four species profiles:

1110
The consensus profile of the above two profiles would be:
1X10
In comparisons of consensus profiles the X character will remain unchanged if matched against another X , shift to 0 if matched against a 0 and shift to 1 if matched against a 1 . Thus a 1 or a 0 in a consensus profile will always be present in more than $50 \%$ of its constituent profiles.

Some of these groups will represent clade specific distributions of proteins. Others will represent distributions of proteins correlated with the distribution a given function over the species under consideration. Comparison of a protein with an as yet unascertained function using consensus profiles would connect a protein to either a clade-specific group or a group, which possessed a function connected to the presence of the protein. Thus a protein that showed correlated evolution with a consensus profile could potentially be functionally linked to all constituent members of that profile. At a higher-level if two consensus profiles show evidence of correlated evolution with each other this could suggest functional linkage between two groups of proteins, e.g. the functional interaction of one pathway with another.

### 6.3.3 Correlated evolution of proteins with the presence or absence of phenotypes

Given the data currently generated an interesting avenue of investigation would be the comparison of the presence and absence of given phenotypes with the presence and absence of given proteins. This process can detect proteins that underlie the phenotype of interest. This method was developed by Levesque (Levesque et al. 2003) and used to detect genes associated with cell motility. It was also applied to associating a number of phenotypes with given proteins (Jim et al. 2004; Slonim et al. 2006). The method was found to be to be reasonably effective with traits that were evenly distributed among the organisms under consideration (Jim et al. 2004). A further application of the method by Gonzalez and Zimmer examined the association of optimal growth pH with given genotypes (Gonzalez and Zimmer 2008). Gonzalez and Zimmer utilised a threshold with which to discretise continuous phenotypes (Gonzalez and Zimmer 2008). If the measured value of a measured phenotype was over a given value then the phenotype was declared present. Applications of this method have so far utilised measures like string distance measures (Jim et al. 2004; Levesque et al. 2003) and mutual information (Gonzalez and Zimmer 2008; Slonim et al. 2006) to compare the phylogenetic profiles of genes and given phenotypes. Use of a phylogenetically aware method such as constrained ML (Barker et al. 2007; Barker and Pagel 2005) would enhance
the method and potentially yield more accurate results. Given the range of eukaryotic organisms currently held potential traits to be investigated could include multi-cellularity, aerobic respiration and parasitism.

### 6.3.4 Drug Targets

Keeping with the theme of parasitism there are a number of disease causing parasitic organisms in the dataset currently held. These are

- Plasmodium falciparum
- Plasmodium knowlesi
- Plasmodium yoelii
- Trypanosoma brucei
- Trypanosoma cruzi
- Leishmania major
- Trichomonas vaginalis
- Theileria annulata
- Theileria parva
- Encephalitozoon cuniculi

These include T. cruzi and T. brucei, which cause Chagas disease (Lescure et al. 2010) and sleeping sickness (Ralston et al. 2009) respectively. Also included in the dataset are three members of the malaria-causing genus Plasmodium. Take for example $P$. falciparum. There is currently resistance to all five groups of anti-malarial drugs (Hayton and Su 2004). The detection of protein interactions in P. falciparum could potentially aid in the development of new anti-malarial drugs. Using this species as a reference point, phylogenetic profiles for its proteome could be constructed. An application of the logistic regression based heuristic would make all against all comparisons using constrained ML (Barker et al. 2007; Barker and Pagel 2005) feasible. These studies could potentially detect novel protein interactions within P. falciparum. Disruption of protein-protein interactions is potentially one avenue for drug development. This could potentially be carried out via procedures such as peptidomimetics (Hruby 1997), which involves the construction of a molecule that mimics the properties of one of the interacting partners. The construction of phylogenetic profiles could also reveal proteins and protein interactions that are unique to P. falciparum. These molecules could potentially be targeted with a lower risk of side effects in the host organism. A similar procedure could be followed with all other parasitic organisms in the dataset.

## References

Agnarsson I, Miller JA (2008) Is Acctran Better Than Deltran? Cladistics 24:1032 Aguinaldo AM, Turbeville JM, Linford LS, Rivera MC, Garey JR, Raff RA, Lake JA (1997) Evidence for a Clade of Nematodes, Arthropods and Other Moulting Animals. Nature 387:489
Ahola V, Aittokallio T, Vihinen M, Uusipaikka E (2006) A Statistical Score for Assessing the Quality of Multiple Sequence Alignments. BMC Bioinformatics 7:484
Albert VA (2006) Parsimony, Phylogeny, and Genomics. Oxford University Press, Oxford Alberts B (1998) Essential Cell Biology : An Introduction to the Molecular Biology of the Cell. Garland, New York
Alberts B (2002) Molecular Biology of the Cell. Garland Science
Alberts B (2008) Molecular Biology of the Cell. Garland Science, New York ; Abingdon
Alberts B (2010) Essential Cell Biology. Garland Science, New York ; London
Alibes A, Yankilevich P, Canada A, Diaz-Uriarte R (2007) Idconverter and Idclight: Conversion and Annotation of Gene and Protein Ids. BMC Bioinformatics 8
Alon U (2007) An Introduction to Systems Biology Design Principles of Biological Circuits. Chapman \& Hall / CRC
Altenhoff AM, Dessimoz C (2009) Phylogenetic and Functional Assessment of Orthologs Inference Projects and Methods. PLoS Comput Biol 5:e1000262
Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic Local Alignment Search Tool. J Mol Biol 215:403
Altschul SF, Wootton JC, Gertz EM, Agarwala R, Morgulis A, Schaffer AA, Yu YK (2005)
Protein Database Searches Using Compositionally Adjusted Substitution Matrices. FEBS Journal 272:5101
C. elegans Sequencing Consortium (1998) Genome Sequence of the Nematode C. Elegans: A Platform for Investigating Biology. Science 282:2012

Antonov AV, Mewes HW (2008) Complex Phylogenetic Profiling Reveals Fundamental Genotype-Phenotype Associations. Computational Biology and Chemistry 32:412

Apweiler R, Martin MJ, O'Donovan C, Magrane M, Alam-Faruque Y, Antunes R, Barrell D, Bely B, Bingley M, Binns D, Bower L, Browne P, Chan WM, Dimmer E, Eberhardt R, Fedotov A, Foulger R, Garavelli J, Huntley R, Jacobsen J, Kleen M, Laiho K, Leinonen R, Legge D, Lin Q, Liu WD, Luo J, Orchard S, Patient S, Poggioli D, Pruess M, Corbett M, di Martino G, Donnelly M, van Rensburg P, Bairoch A, Bougueleret L, Xenarios I, Altairac S, Auchincloss A, Argoud-Puy G, Axelsen K, Baratin D, Blatter MC, Boeckmann B, Bolleman J, Bollondi L, Boutet E, Quintaje SB, Breuza L, Bridge A, deCastro E, Ciapina L, Coral D, Coudert E, Cusin I, Delbard G, Doche M, Dornevil D, Roggli PD, Duvaud S, Estreicher A, Famiglietti L, Feuermann M, Gehant S, Farriol-Mathis N, Ferro S, Gasteiger E, Gateau A, Gerritsen V, Gos A, Gruaz-Gumowski N, Hinz U, Hulo C, Hulo N, James J, Jimenez S, Jungo F, Kappler T, Keller G, Lachaize C, Lane-Guermonprez L, Langendijk-Genevaux P, Lara V, Lemercier P, Lieberherr D, Lima TD, Mangold V, Martin X, Masson P, Moinat M, Morgat A, Mottaz A, Paesano S, Pedruzzi I, Pilbout S, Pillet V, Poux S, Pozzato M, Redaschi N, Rivoire C, Roechert B, Schneider M, Sigrist C, Sonesson K, Staehli S, Stanley E, Stutz A, Sundaram S, Tognolli M, Verbregue L, Veuthey AL, Yip LN, Zuletta L, Wu C, Arighi C, Arminski L, Barker W, Chen CM, Chen YX, Hu ZZ, Huang HZ, Mazumder R, McGarvey P, Natale DA, Nchoutmboube J, Petrova N, Subramanian N, Suzek BE, Ugochukwu U,

Vasudevan S, Vinayaka CR, Yeh LS, Zhang J (2010) The Universal Protein Resource (Uniprot) in 2010. Nucleic Acids Research 38:D142
Aranda B, Achuthan P, Alam-Faruque Y, Armean I, Bridge A, Derow C, Feuermann M, Ghanbarian AT, Kerrien S, Khadake J, Kerssemakers J, Leroy C, Menden M, Michaut M, Montecchi-Palazzi L, Neuhauser SN, Orchard S, Perreau V, Roechert B, van Eijk K, Hermjakob H (2010) The Intact Molecular Interaction Database in 2010. Nucleic Acids Res 38:D525

Aravind L (2000) Guilt by Association: Contextual Information in Genome Analysis. Genome Research 10:1074
Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene Ontology: Tool for the Unification of Biology. The Gene Ontology Consortium. Nat Genet 25:25
Aubourg S, Rouze P (2001) Genome Annotation. Plant Physiology and Biochemistry 39:181
Avery OT, Macleod CM, McCarty M (1944) Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types : Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type Iii. J Exp Med 79:137
Bader GD, Donaldson I, Wolting C, Ouellette BF, Pawson T, Hogue CW (2001) Bind--the Biomolecular Interaction Network Database. Nucleic Acids Res 29:242
Baldauf SL, Roger AJ, Wenk-Siefert I, Doolittle WF (2000) A Kingdom-Level Phylogeny of Eukaryotes Based on Combined Protein Data. Science 290:972
Baldi P, Brunak S (2001) Bioinformatics : The Machine Learning Approach. MIT Press, Cambridge, Mass.
Barker D, Meade A, Pagel M (2007) Constrained Models of Evolution Lead to Improved Prediction of Functional Linkage from Correlated Gain and Loss of Genes. Bioinformatics 23:14
Barker D, Pagel M (2005) Predicting Functional Gene Links from PhylogeneticStatistical Analyses of Whole Genomes. PLoS Comput Biol 1:e3
Beadle GW, Tatum EL (1941) Genetic Control of Biochemical Reactions in Neurospora. Proc Natl Acad Sci U S A 27:499
Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2009) Genbank. Nucleic Acids Res 37:D26
Berg JM, Tymoczko JL, Stryer L (2001) Biochemistry. W. H. Freeman and CO., New York
Berg JM, Tymoczko JL, Stryer L (2007) Biochemistry. W. H. Freeman, New York
Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, Kirilovsky A, Fridman WH, Pages F, Trajanoski Z, Galon J (2009) Cluego: A Cytoscape Plug-in to Decipher Functionally Grouped Gene Ontology and Pathway Annotation Networks. Bioinformatics 25:1091
Birney E, Clamp M, Durbin R (2004) Genewise and Genomewise. Genome Res 14:988
Black DL (2003) Mechanisms of Alternative Pre-Messenger Rna Splicing. Annu Rev Biochem 72:291
Blair C, Murphy RW (2011) Recent Trends in Molecular Phylogenetic Analysis: Where to Next? J Hered 102:130
Blanchard JL, Lynch M (2000) Organellar Genes - Why Do They End up in the Nucleus? Trends in Genetics 16:315
Blow MJ (2004) A Survey of RNA Editing in the Human Brain Sanger Institute. University of Cambridge, Cambridge

Borodovsky M, Rudd KE, Koonin EV (1994) Intrinsic and Extrinsic Approaches for Detecting Genes in a Bacterial Genome. Nucleic Acids Res 22:4756
Bowers PM, Cokus SJ, Elsenberg D, Yeates TO (2004) Use of Logic Relationships to Decipher Protein Network Organization. Science 306:2246
Bratke K (2009) Comparative Analysis of Poxvirus Genome Evolution. University of Dublin, Trinity College, Dublin
Breathnach R, Benoist C, O'Hare K, Gannon F, Chambon P (1978) Ovalbumin Gene: Evidence for a Leader Sequence in mRNA and DNA Sequences at the ExonIntron Boundaries. Proc Natl Acad Sci U S A 75:4853
Brennan RG, Matthews BW (1989) The Helix-Turn-Helix DNA Binding Motif. J Biol Chem 264:1903
Brent MR (2008) Steady Progress and Recent Breakthroughs in the Accuracy of Automated Genome Annotation. Nat Rev Genet 9:62
Brown KR, Jurisica I (2005) Online Predicted Human Interaction Database. Bioinformatics 21:2076
Brown TA (2006) Genomes 3. Garland Science Pub., New York
Bruno WJ, Halpern AL (1999) Topological Bias and Inconsistency of Maximum Likelihood Using Wrong Models. Molecular Biology and Evolution 16:564
Burge C, Karlin S (1997) Prediction of Complete Gene Structures in Human Genomic DNA. Journal of Molecular Biology 268:78
Burki F, Shalchian-Tabrizi K, Pawlowski J (2008) Phylogenomics Reveals a New 'Megagroup' Including Most Photosynthetic Eukaryotes. Biology Letters 4:366
Buza TJ, McCarthy FM, Wang N, Bridges SM, Burgess SC (2008) Gene Ontology Annotation Quality Analysis in Model Eukaryotes. Nucleic Acids Research 36(2):e12
Cai JC, G. Wang , J (2010) ClusterViz: A Cytoscape Plugin for Graph Clustering and Visualization Central South University, Changsha

Camin JH, Sokal RR (1965) A Method for Deducing Branching Sequences in Phylogeny. Evolution 19:311

Capecchi MR (2005) Gene Targeting in Mice: Functional Analysis of the Mammalian Genome for the Twenty-First Century. Nat Rev Genet 6:507
Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T (2009) Trimal: A Tool for Automated Alignment Trimming in Large-Scale Phylogenetic Analyses. Bioinformatics 25:1972
Cavalli-Sforza LLE, Edwards A.W.F (1964) Reconstruction of Evolutionary Trees. Phenetic and Phylogenetic Classification 6:67-76
Ceol A, Aryamontri AC, Licata L, Peluso D, Briganti L, Perfetto L, Castagnoli L, Cesareni G (2010) Mint, the Molecular Interaction Database: 2009 Update. Nucleic Acids Research 38:D532
Chalfie M, Tu Y, Euskirchen G, Ward WW, Prasher DC (1994) Green Fluorescent Protein as a Marker for Gene-Expression. Science 263:802
Chatr-aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, Castagnoli L, Cesareni G (2007) Mint: The Molecular Interaction Database. Nucleic Acids Res 35:D572
Chen LF, Vitkup D (2006) Predicting Genes for Orphan Metabolic Activities Using Phylogenetic Profiles. Genome Biology 7:R17
Chen XW, Jeong JC, Dermyer P (2011) Kups: Constructing Datasets of Interacting and Non-Interacting Protein Pairs with Associated Attributions. Nucleic Acids Res 39:D750

Coin F, Marinoni JC, Rodolfo C, Fribourg S, Pedrini AM, Egly JM (1998) Mutations in the Xpd Helicase Gene Result in Xp and Ttd Phenotypes, Preventing Interaction between Xpd and the P44 Subunit of Tfiih. Nature Genetics 20:184
Cokus S, Mizutani S, Pellegrini M (2007) An Improved Method for Identifying Functionally Linked Proteins Using Phylogenetic Profiles. BMC Bioinformatics 8:S7
Coulson RMR, Ouzounis CA (2003) The Phylogenetic Diversity of Eukaryotic Transcription. Nucleic Acids Res 31:653
Cranston KA, Hurwitz B, Ware D, Stein L, Wing RA (2009) Species Trees from Highly Incongruent Gene Trees in Rice. Systematic Biology 58:489
Cranston KA, Rannala B (2007) Summarizing a Posterior Distribution of Trees Using Agreement Subtrees. Systematic Biology 56:578
Crick FH, Barnett L, Brenner S, Watts-Tobin RJ (1961) General Nature of the Genetic Code for Proteins. Nature 192:1227
Cunningham FX, Lafond TP, Gantt E (2000) Evidence of a Role for Lytb in the Nonmevalonate Pathway of Isoprenoid Biosynthesis. Journal of Bacteriology 182:5841
Curwen V, Eyras E, Andrews TD, Clarke L, Mongin E, Searle SM, Clamp M (2004) The Ensembl Automatic Gene Annotation System. Genome Res 14:942
Dandekar T, Snel B, Huynen M, Bork P (1998) Conservation of Gene Order: A Fingerprint of Proteins That Physically Interact. Trends Biochem Sci 23:324
Davey R, Savva G, Dicks J, Roberts IN (2007) Mpp: A Microarray-to-Phylogeny Pipeline for Analysis of Gene and Marker Content Datasets. Bioinformatics 23:1023
Dayhoff MO, Schwartz. RM, Orcutt. BC (1978) A Model of Evolutionary Change in Proteins. Atlas of Protein Sequence and Structure 5:345
De Bodt S, Proost S, Vandepoele K, Rouze P, Van de Peer Y (2009) Predicting ProteinProtein Interactions in Arabidopsis Thaliana through Integration of Orthology, Gene Ontology and Co-Expression. BMC Genomics 10:288
De Las Rivas J, Fontanillo C (2010) Protein, Protein Interactions Essentials: Key Concepts to Building and Analyzing Interactome Networks. PLoS Comput Biol 6:e1000807
Dereeper A, Guignon V, Blanc G, Audic S, Buffet S, Chevenet F, Dufayard JF, Guindon S, Lefort V, Lescot M, Claverie JM, Gascuel O (2008) Phylogeny.Fr: Robust Phylogenetic Analysis for the Non-Specialist. Nucleic Acids Research 36:W465
Dowsey AW, Dunn MJ, Yang GZ (2003) The Role of Bioinformatics in Two-Dimensional Gel Electrophoresis. Proteomics 3:1567
Durbin R (1998) Biological Sequence Analysis : Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press, Cambridge New York
Eddy SR (1998) Profile Hidden Markov Models. Bioinformatics 14:755
Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z (2009) Gorilla: A Tool for Discovery and Visualization of Enriched Go Terms in Ranked Gene Lists. BMC Bioinformatics 10:48
Edgar RC (2004) Muscle: Multiple Sequence Alignment with High Accuracy and High Throughput. Nucleic Acids Res 32:1792
Edgar RC, Batzoglou S (2006) Multiple Sequence Alignment. Curr Opin Struct Biol 16:368
Edgell DR, Belfort M, Shub DA (2000) Barriers to Intron Promiscuity in Bacteria. J Bacteriol 182:5281
Edwards AWF (1992) Likelihood. Johns Hopkins University Press, Baltimore ; London Elias I, Tuller T (2007) Reconstruction of Ancestral Genomic Sequences Using Likelihood. Journal of Computational Biology 14:216

Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA (1999) Protein Interaction Maps for Complete Genomes Based on Gene Fusion Events. Nature 402:86
Farrar M (2007) Striped Smith-Waterman Speeds Database Searches Six Times over Other Simd Implementations. Bioinformatics 23:156
Farris JS (1977) Phylogenetic Analysis under Dollo's Law. Systematic Zoology 26:77
Farris JS (1978) Inferring Phylogenetic Trees from Chromosome Inversion Data. Systematic Zoology 27:275
Felsenstein J (1973) Maximum Likelihood and Minimum-Steps Methods for Estimating Evolutionary Trees from Data on Discrete Characters. Systematic Zoology 22:240
Felsenstein J (1978) Cases in Which Parsimony or Compatibility Methods Will Be Positively Misleading. Syst Zool 27:401
Felsenstein J (1979) Alternative Methods of Phylogenetic Inference and Their Interrelationship. Systematic Zoology 28:49
Felsenstein J (1985a) Confidence-Limits on Phylogenies - an Approach Using the Bootstrap. Evolution 39:783
Felsenstein J (1985b) Phylogenies and the Comparative Method. The American Naturalist 125:1
Felsenstein J (1989) Phylip - Phylogeny Inference Package (Version 3.2). Cladistics 5:164
Felsenstein J (2004) Inferring Phylogenies. Sinauer Associates, Sunderland, Mass.
Fiers W, Contreras R, Duerinck F, Haegeman G, Iserentant D, Merregaert J, Min Jou W, Molemans F, Raeymaekers A, Van den Berghe A, Volckaert G, Ysebaert M (1976) Complete Nucleotide Sequence of Bacteriophage Ms2 Rna: Primary and Secondary Structure of the Replicase Gene. Nature 260:500
Fitch WM (1970) Distinguishing Homologous from Analogous Proteins. Syst Zool 19:99
Fitch WM (1971) Toward Defining Course of Evolution - Minimum Change for a Specific Tree Topology. Syst Zool 20:406
Fitch WM (2000) Homology a Personal View on Some of the Problems. Trends Genet 16:227
Fitzpatrick DA, Logue ME, Stajich JE, Butler G (2006) A Fungal Phylogeny Based on 42 Complete Genomes Derived from Supertree and Combined Gene Analysis. BMC Evol Biol 6:99
Florea L, Hartzell G, Zhang Z, Rubin GM, Miller W (1998) A Computer Program for Aligning a Cdna Sequence with a Genomic DNA Sequence. Genome Res 8:967
Fu N, Drinnenberg I, Kelso J, Wu JR, Paabo S, Zeng R, Khaitovich P (2007) Comparison of Protein and Mrna Expression Evolution in Humans and Chimpanzees. PLoS One 2:e216
Garman KS, Acharya CR, Edelman E, Grade M, Gaedcke J, Sud S, Barry W, Diehl AM, Provenzale D, Ginsburg GS, Ghadimi BM, Ried T, Nevins JR, Mukherjee S, Hsu D, Potti A (2009) A Genomic Approach to Colon Cancer Risk Stratification Yields Biologic Insights into Therapeutic Opportunities (Vol 105, 19432, 2008). Proceedings of the National Academy of Sciences of the United States of America 106:6878
Garrett S, Barton WA, Knights R, Jin P, Morgan DO, Fisher RP (2001) Reciprocal Activation by Cyclin-Dependent Kinases 2 and 7 Is Directed by Substrate Specificity Determinants Outside the T Loop. Molecular and Cellular Biology 21:88
Gaschen B, Taylor J, Yusim K, Foley B, Gao F, Lang D, Novitsky V, Haynes B, Hahn BH, Bhattacharya T, Korber B (2002) Aids - Diversity Considerations in Hiv-1 Vaccine Selection. Science 296:2354

Ge XJ, Yamamoto S, Tsutsumi S, Midorikawa Y, Ihara S, Wang SM, Aburatani H (2005) Interpreting Expression Profiles of Cancers by Genome-Wide Survey of Breadth of Expression in Normal Tissues. Genomics 86:127
Gillis B, Gavin IM, Arbieva Z, King ST, Jayaraman S, Prabhakar BS (2007) Identification of Human Cell Responses to Benzene and Benzene Metabolites. Genomics 90:324
Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG (1996) Life with $\mathbf{6 0 0 0}$ Genes. Science 274:546
Goldman N, Anderson JP, Rodrigo AG (2000) Likelihood-Based Tests of Topologies in Phylogenetics. Systematic Biology 49:652
Goloboff PA, Catalano SA, Mirande JM, Szumik CA, Arias JS, Kallersjo M, Farris JS (2009) Phylogenetic Analysis of $\mathbf{7 3} \mathbf{0 6 0}$ Taxa Corroborates Major Eukaryotic Groups. Cladistics 25:211
Gonzalez O, Zimmer R (2008) Assigning Functional Linkages to Proteins Using Phylogenetic Profiles and Continuous Phenotypes. Bioinformatics 24:1257
Grafen A (1989) The Phylogenetic Regression. Philosophical Transactions of the Royal Society of London Series B-Biological Sciences 326:119
Graur D, Shuali Y, Li WH (1989) Deletions in Processed Pseudogenes Accumulate Faster in Rodents Than in Humans. Journal of Molecular Evolution 28:279
Griffiths AJF (2002) Modern Genetic Analysis : Integrating Genes and Genomes. W.H. Freeman and Co., New York
Guindon S, Gascuel O (2003) A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. Syst Biol 52:696
Gygi SP, Rochon Y, Franza BR, Aebersold R (1999) Correlation between Protein and Mrna Abundance in Yeast. Molecular and Cellular Biology 19:1720
Hakes L, Pinney J.W, Lowell S.C, Oliver S.G, Robertson D.L (2007) All Duplicates Are Not Equal: The Difference between Small-Scale and Genome Duplication. Genome Biology 8:R209
Hamming RW (1950) Error Detecting and Error Correcting Codes. Bell System Technical Journal 26:147
Han JDJ, Bertin N, Hao T, Goldberg DS, Berriz GF, Zhang LV, Dupuy D, Walhout AJM, Cusick ME, Roth FP, Vidal M (2004) Evidence for Dynamically Organized Modularity in the Yeast Protein-Protein Interaction Network. Nature 430:88
Harrison CJ, Langdale JA (2006) A Step by Step Guide to Phylogeny Reconstruction. Plant J 45:561
Hart GT, Ramani AK, Marcotte EM (2006) How Complete Are Current Yeast and Human Protein-Interaction Networks? Genome Biol 7:120
Harvey PH and Pagel MD (1991). The Comparative Method in Evolutionary Biology. Oxford: Oxford University Press

Hasegawa H, Holm L (2009) Advances and Pitfalls of Protein Structural Alignment. Curr Opin Struct Biol 19:341
Hasegawa M, Kishino H (1989) Confidence-Limits on the Maximum-Likelihood Estimate of the Hominoid Tree from Mitochondrial-DNA Sequences. Evolution 43:672
Haw R, Hermjakob H, D'Eustachio P, Stein L (2011) Reactome Pathway Analysis to
Enrich Biological Discovery in Proteomics Datasets. Proteomics: 11(18):3598-613.
Hayton K, Su XZ (2004) Genetic and Biochemical Aspects of Drug Resistance in Malaria Parasites. Curr Drug Targets Infect Disord 4:1

He HY, Soncin F, Grammatikakis N, Li YL, Siganou A, Gong JL, Brown SA, Kingston RE, Calderwood SK (2003) Elevated Expression of Heat Shock Factor (Hsf) 2a Stimulates Hsf1-Induced Transcription During Stress. Journal of Biological Chemistry 278:35465
Heath TA, Hedtke SM, Hillis DM (2008) Taxon Sampling and the Accuracy of Phylogenetic Analyses. Journal of Systematics and Evolution 46:239
Henikoff S, Henikoff JG (1992) Amino Acid Substitution Matrices from Protein Blocks. Proc Natl Acad Sci U S A 89:10915
Hershey AD, Chase M (1952) Independent Functions of Viral Protein and Nucleic Acid in Growth of Bacteriophage. J Gen Physiol 36:39
Hert DG, Fredlake CP, Barron AE (2008) Advantages and Limitations of Next-Generation Sequencing Technologies: A Comparison of Electrophoresis and NonElectrophoresis Methods. Electrophoresis 29:4618
Heyer LJ, Kruglyak S, Yooseph S (1999) Exploring Expression Data: Identification and Analysis of Coexpressed Genes. Genome Research 9:1106
Higgins DG, Sharp PM (1988) Clustal: A Package for Performing Multiple Sequence Alignment on a Microcomputer. Gene 73:237
Hill J, Hambley M, Forster T, Mewissen M, Sloan TM, Scharinger F, Trew A, Ghazal P (2008) Sprint: A New Parallel Framework for R. BMC Bioinformatics 9

Hobolth A, Christensen OF, Mailund T, Schierup MH (2007) Genomic Relationships and Speciation Times of Human, Chimpanzee, and Gorilla Inferred from a Coalescent Hidden Markov Model. PLoS Genet 3:e7
Hodges A, Strand AD, Aragaki AK, Kuhn A, Sengstag T, Hughes G, Elliston LA, Hartog C, Goldstein DR, Thu D, Hollingsworth ZR, Collin F, Synek B, Holmans PA, Young AB, Wexler NS, Delorenzi M, Kooperberg C, Augood SJ, Faull RL, Olson JM, Jones L, Luthi-Carter R (2006) Regional and Cellular Gene Expression Changes in Human Huntington's Disease Brain. Hum Mol Genet 15:965
Holder M, Lewis PO (2003) Phylogeny Estimation: Traditional and Bayesian Approaches. Nature Reviews Genetics 4:275
Hruby VJ (1997) Prospects for Peptidomimetic Drug Design. Drug Discovery Today 2:165
Huai Q, Kim HY, Liu YD, Zhao YD, Mondragon A, Liu JO, Ke HM (2002) Crystal Structure of Calcineurin-Cyclophilin-Cyclosporin Shows Common but Distinct Recognition of Immunophilin-Drug Complexes. Proceedings of the National Academy of Sciences of the United States of America 99:12037
Huelsenbeck JP, Bollback JP (2001) Empirical and Hierarchical Bayesian Estimation of Ancestral States. Systematic Biology 50:351
Huelsenbeck JP, Ronquist F, Nielsen R, Bollback JP (2001) Bayesian Inference of Phylogeny and Its Impact on Evolutionary Biology. Science 294:2310
Hughes AL, Friedman R (2005) Poxvirus Genome Evolution by Gene Gain and Loss. Molecular Phylogenetics and Evolution 35:186
Hulsen T, Huynen MA, de Vlieg J, Groenen PM (2006) Benchmarking Ortholog Identification Methods Using Functional Genomics Data. Genome Biol 7:R31
Hurles M (2004) Gene Duplication: The Genomic Trade in Spare Parts. PLoS Biol 2:E206
Ispolatov I, Yuryev A, Mazo I, Maslov S (2005) Binding Properties and Evolution of Homodimers in Protein-Protein Interaction Networks. Nucleic Acids Res 33:3629
Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF, Gerstein M (2003) A Bayesian Networks Approach for Predicting Protein-Protein Interactions from Genomic Data. Science 302:449

Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, Doerks T, Julien P, Roth A, Simonovic M, Bork P, von Mering C (2009) String 8-a Global View on Proteins and Their Functional Interactions in $\mathbf{6 3 0}$ Organisms. Nucleic Acids Research 37:D412
Jeong H, Mason SP, Barabasi AL, Oltvai ZN (2001) Lethality and Centrality in Protein Networks. Nature 411:41
Jessop CE, Chakravarthi S, Garbi N, Hammerling GJ, Lovell S, Bulleid NJ (2007) Erp57 Is Essential for Efficient Folding of Glycoproteins Sharing Common Structural Domains. EMBO J 26:28
Jim K, Parmar K, Singh M, Tavazoie S (2004) A Cross-Genomic Approach for Systematic Mapping of Phenotypic Traits to Genes. Genome Research 14:109
Johnson JM, Castle J, Garrett-Engele P, Kan Z, Loerch PM, Armour CD, Santos R, Schadt EE, Stoughton R, Shoemaker DD (2003) Genome-Wide Survey of Human Alternative Pre-Mrna Splicing with Exon Junction Microarrays. Science 302:2141
Kanehisa M (1997) A Database for Post-Genome Analysis. Trends Genet 13:375
Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M (2006) From Genomics to Chemical Genomics: New Developments in Kegg. Nucleic Acids Res 34:D354
Karlin S, Altschul SF (1990) Methods for Assessing the Statistical Significance of Molecular Sequence Features by Using General Scoring Schemes. Proc Natl Acad Sci U S A 87:2264
Katoh K, Misawa K, Kuma K, Miyata T (2002) Mafft: A Novel Method for Rapid Multiple Sequence Alignment Based on Fast Fourier Transform. Nucleic Acids Res 30:3059
Kawaji H, Hayashizaki Y (2008) Genome Annotation. Methods Mol Biol 452:125
Keane TM, Creevey CJ, Pentony MM, Naughton TJ, Mclnerney JO (2006) Assessment of Methods for Amino Acid Matrix Selection and Their Use on Empirical Data Shows That Ad Hoc Assumptions for Choice of Matrix Are Not Justified. BMC Evolutionary Biology 6
Kensche PR, van Noort V, Dutilh BE, Huynen MA (2008) Practical and Theoretical Advances in Predicting the Function of a Protein by Its Phylogenetic Distribution. J R Soc Interface 5:151
Kersey PJ, Duarte J, Williams A, Karavidopoulou Y, Birney E, Apweiler R (2004) The International Protein Index: An Integrated Database for Proteomics Experiments. Proteomics 4:1985
Kim IY, Shin JH, Seong JK (2010) Mouse Phenogenomics, Toolbox for Functional Annotation of Human Genome. BMB Rep 43:79
Kim MJ, Romero R, Kim CJ, Tarca AL, Chhauy S, LaJeunesse C, Lee DC, Draghici S, Gotsch F, Kusanovic JP, Hassan SS, Kim JS (2009) Villitis of Unknown Etiology Is Associated with a Distinct Pattern of Chemokine up-Regulation in the FetoMaternal and Placental Compartments: Implications for Conjoint Maternal Allograft Rejection and Maternal Anti-Fetal Graft-Versus-Host Disease. Journal of Immunology 182:3919
Knight RD, Landweber LF, Yarus M (2001) How Mitochondria Redefine the Code. J Mol Evol 53:299
Knowles DG, McLysaght A (2009) Recent De Novo Origin of Human Protein-Coding Genes. Genome Res 19:1752
Korf I (2004) Gene Finding in Novel Genomes. BMC Bioinformatics 5:59
Koshi JM, Goldstein RA (1996) Probabilistic Reconstruction of Ancestral ProteinSequences. Journal of Molecular Evolution 42:313Krane DE, Raymer ML (2003) Fundamental Concepts of Bioinformatics. PearsonEducation International, San Francisco
Krylov DM, Wolf YI, Rogozin IB, Koonin EV (2003) Gene Loss, Protein SequenceDivergence, Gene Dispensability, Expression Level, and Interactivity AreCorrelated in Eukaryotic Evolution. Genome Research 13:2229Kuhner MK, Felsenstein J (1994) Simulation Comparison of Phylogeny Algorithms underEqual and Unequal Evolutionary Rates. Mol Biol Evol 11:459
Kummel D, Oeckinghaus A, Wang C, Krappmann D, Heinemann U (2008) DistinctIsocomplexes of the Trapp Trafficking Factor Coexist inside Human Cells. FEBSLett 582:3729Lande J, Gimino V, Berryman T, Hertz MI, King RA (2003) Gene Expression Profiling ofBronchoalveolar Lavage Cells in Acute Lung Rejection. American Journal ofHuman Genetics 73:421
Le SQ, Gascuel O (2008) An Improved General Amino Acid Replacement Matrix. MolBiol Evol 25:1307
Lei PW, Koehly LM (2003) Linear Discriminant Analysis Versus Logistic Regression: AComparison of Classification Errors in the Two-Group Case. Journal ofExperimental Education 72:25
Lequesne WJ (1974) Uniquely Evolved Character Concept and Its Cladistic Application.Systematic Zoology 23:513
Lescure FX, Le Loup G, Freilij H, Develoux M, Paris L, Brutus L, Pialoux G (2010) ChagasDisease: Changes in Knowledge and Management. Lancet Infectious Diseases10:556
Levesque M, Shasha D, Kim W, Surette MG, Benfey PN (2003) Trait-to-Gene: AComputational Method for Predicting the Function of Uncharacterized Genes.Current Biology 13:129
Lewinski MK, Bisgrove D, Shinn P, Chen H, Hoffmann C, Hannenhalli S, Verdin E, BerryCC, Ecker JR, Bushman FD (2005) Genome-Wide Analysis of ChromosomalFeatures Repressing Human Immunodeficiency Virus Transcription. Journal ofVirology 79:6610
Li L, Stoeckert CJ, Jr., Roos DS (2003) Orthomcl: Identification of Ortholog Groups forEukaryotic Genomes. Genome Res 13:2178
Li M, Wang JX, Chen J (2008) A Fast Agglomerate Algorithm for Mining Functional Modules in Protein Interaction Networks. Bmei 2008: Proceedings of the International Conference on Biomedical Engineering and Informatics, Vol 1:3
Linial M (2003) How Incorrect Annotations Evolve - the Case of Short Orfs. Trends in Biotechnology 21:298
Lunter G, Ponting CP, Hein J (2006) Genome-Wide Identification of Human Functional DNA Using a Neutral Indel Model. PLoS Comput Biol 2:2
Macagno A, Molteni M, Rinaldi A, Bertoni F, Lanzavecchia A, Rossetti C, Sallusto F (2006)A Cyanobacterial Lps Antagonist Prevents Endotoxin Shock and BlocksSustained Tlr4 Stimulation Required for Cytokine Expression. Journal ofExperimental Medicine 203:1481
Maddison WP (1990) A Method for Testing the Correlated Evolution of Two BinaryCharacters - Are Gains or Losses Concentrated on Certain Branches of aPhylogenetic Tree. Evolution 44:539
Maddison WP, Maddison DR (2010) Mesquite: A Modular System for EvolutionaryAnalysis. Version 2.73

Maere S, Heymans K, Kuiper M (2005) Bingo: A Cytoscape Plugin to Assess Overrepresentation of Gene Ontology Categories in Biological Networks. Bioinformatics 21:3448
Malcolm BA, Wilson KP, Matthews BW, Kirsch JF, Wilson AC (1990) Ancestral Lysozymes Reconstructed, Neutrality Tested, and Thermostability Linked to Hydrocarbon Packing. Nature 345:86
Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D (1999) Detecting Protein Function and Protein-Protein Interactions from Genome Sequences. Science 285:751
Mardis ER (2008) The Impact of Next-Generation Sequencing Technology on Genetics. Trends Genet 24:133
Martin DM, Berriman M, Barton GJ (2004) Gotcha: A New Method for Prediction of Protein Function Assessed by the Annotation of Seven Genomes. BMC Bioinformatics 5:178
Maston GA, Evans SK, Green MR (2006) Transcriptional Regulatory Elements in the Human Genome. Annual Review of Genomics and Human Genetics 7:29
Maxam AM, Gilbert W (1977) New Method for Sequencing DNA. Proc Natl Acad Sci U S A 74:560
McDowall MD, Scott MS, Barton GJ (2009) Pips: Human Protein-Protein Interaction Prediction Database. Nucleic Acids Res 37:D651
McKernan KJ, Peckham HE, Costa GL, McLaughlin SF, Fu YT, Tsung EF, Clouser CR, Duncan C, Ichikawa JK, Lee CC, Zhang Z, Ranade SS, Dimalanta ET, Hyland FC, Sokolsky TD, Zhang L, Sheridan A, Fu HN, Hendrickson CL, Li B, Kotler L, Stuart JR, Malek JA, Manning JM, Antipova AA, Perez DS, Moore MP, Hayashibara KC, Lyons MR, Beaudoin RE, Coleman BE, Laptewicz MW, Sannicandro AE, Rhodes MD, Gottimukkala RK, Yang S, Bafna V, Bashir A, MacBride A, Alkan C, Kidd JM, Eichler EE, Reese MG, De la Vega FM, Blanchard AP (2009) Sequence and Structural Variation in a Human Genome Uncovered by Short-Read, Massively Parallel Ligation Sequencing Using Two-Base Encoding. Genome Research 19:1527
McLysaght A, Baldi PF, Gaut BS (2003) Extensive Gene Gain Associated with Adaptive Evolution of Poxviruses. Proceedings of the National Academy of Sciences of the United States of America 100:15655
Messler W, Stewart CB (1997) Episodic Adaptive Evolution of Primate Lysozymes. Nature 385:151
Mishra GR, Suresh M, Kumaran K, Kannabiran N, Suresh S, Bala P, Shivakumar K, Anuradha N, Reddy R, Raghavan TM, Menon S, Hanumanthu G, Gupta M, Upendran S, Gupta S, Mahesh M, Jacob B, Mathew P, Chatterjee P, Arun KS, Sharma S, Chandrika KN, Deshpande N, Palvankar K, Raghavnath R, Krishnakanth R, Karathia H, Rekha B, Nayak R, Vishnupriya G, Kumar HG, Nagini M, Kumar GS, Jose R, Deepthi P, Mohan SS, Gandhi TK, Harsha HC, Deshpande KS, Sarker M, Prasad TS, Pandey A (2006) Human Protein Reference Database--2006 Update. Nucleic Acids Res 34:D411
Mohamed TP, Carbonell JG, Ganapathiraju MK (2010) Active Learning for Human Protein-Protein Interaction Prediction. BMC Bioinformatics 11 Suppl 1:S57
Monsalve M, Wu ZD, Adelmant G, Puigserver P, Fan ML, Spiegelman BM (2000) Direct Coupling of Transcription and Mrna Processing through the Thermogenic Coactivator Pge-1. Molecular Cell 6:307
Moore KJ (1999) Utilization of Mouse Models in the Discovery of Human Disease Genes. Drug Discov Today 4:123

Morgenstern B, Frech K, Dress A, Werner T (1998) Dialign: Finding Local Similarities by Multiple Sequence Alignment. Bioinformatics 14:290
Mount DW (2004) Bioinformatics : Sequence and Genome Analysis. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.
Needleman SB, Wunsch CD (1970) A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. J Mol Biol 48:443
Nei M, Kumar S (2000) Molecular Evolution and Phylogenetics. Oxford University Press Nooren IMA, Thornton JM (2003) Structural Characterisation and Functional Significance of Transient Protein-Protein Interactions. Journal of Molecular Biology 325:991
Nuin PA, Wang Z, Tillier ER (2006) The Accuracy of Several Multiple Sequence Alignment Programs for Proteins. BMC Bioinformatics 7:471
Nye TMW, Lio P, Gilks WR (2006) A Novel Algorithm and Web-Based Tool for Comparing Two Alternative Phylogenetic Trees. Bioinformatics 22:117
O'donnell RK, Kupferman M, Wei SJ, Singhal S, Weber R, O'Malley B, Cheng Y, Putt M, Feldman M, Ziober B, Muschel RJ (2005) Gene Expression Signature Predicts Lymphatic Metastasis in Squamous Cell Carcinoma of the Oral Cavity. Oncogene 24:1244
Ohta S, Shiomi Y, Sugimoto K, Obuse C, Tsurimoto T (2002) A Proteomics Approach to Identify Proliferating Cell Nuclear Antigen (Pena)-Binding Proteins in Human Cell Lysates - Identification of the Human Chl12/Rfes2-5 Complex as a Novel Pena-Binding Protein. Journal of Biological Chemistry 277:40362
Ooi SL, Pan X, Peyser BD, Ye P, Meluh PB, Yuan DS, Irizarry RA, Bader JS, Spencer FA, Boeke JD (2006) Global Synthetic-Lethality Analysis and Yeast Functional Profiling. Trends Genet 22:56
Orengo C, Jones D, Thornton JM (2003) Bioinformatics : Genes, Proteins, and Computers. BIOS Scientific ; Distributed in the U.S. by Springer-Verlag, Oxford New York
Orlowski J, Kaczanowski S, Zielenkiewicz P (2007) Overrepresentation of Interactions between Homologous Proteins in Interactomes. Febs Letters 581:52
Page RDM, Holmes EC (1998) Molecular Evolution : A Phylogenetic Approach. Blackwell Science, Oxford
Pagel M (1994) Detecting Correlated Evolution on Phylogenies - a General-Method for the Comparative-Analysis of Discrete Characters. Proceedings of the Royal Society of London Series B-Biological Sciences 255:37
Pagel M (1997) Inferring Evolutionary Processes from Phylogenies. Zoologica Scripta 26:331
Pagel M, Meade A, Barker D (2004a) Bayesian Estimation of Ancestral Character States on Phylogenies. Syst Biol 53:673
Pagel P, Kovac S, Oesterheld M, Brauner B, Dunger-Kaltenbach I, Frishman G, Montrone C, Mark P, Stumpflen V, Mewes HW, Ruepp A, Frishman D (2005) The Mips Mammalian Protein-Protein Interaction Database. Bioinformatics 21:832
Pagel P, Wong P, Frishman D (2004b) A Domain Interaction Map Based on Phylogenetic Profiling. J Mol Biol 344:1331
Parfrey LW, Barbero E, Lasser E, Dunthorn M, Bhattacharya D, Patterson DJ, Katz LA (2006) Evaluating Support for the Current Classification of Eukaryotic Diversity. PLoS Genet 2: 220
Parida L (2008) Pattern Discovery in Bioinformatics : Theory \& Algorithms. Chapman \& Hall/CRC, London

Pazos F, Ranea JAG, Juan D, Sternberg MJE (2005) Assessing Protein Co-Evolution in the Context of the Tree of Life Assists in the Prediction of the Interactome. Journal of Molecular Biology 352:1002
Pazos F, Valencia A (2001) Similarity of Phylogenetic Trees as Indicator of ProteinProtein Interaction. Protein Eng 14:609
Pearson WR, Lipman DJ (1988) Improved Tools for Biological Sequence Comparison. Proc Natl Acad Sci U S A 85:2444
Pellegrini M (2001) Computational Methods for Protein Function Analysis. Curr Opin Chem Biol 5:46
Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO (1999) Assigning Protein Functions by Comparative Genome Analysis: Protein Phylogenetic Profiles. Proc Natl Acad Sci U S A 96:4285
Pesole G (2008) What Is a Gene? An Updated Operational Definition. Gene 417:1
Picardi E, Pesole G (2010) Computational Methods for Ab Initio and Comparative Gene Finding. Methods Mol Biol 609:269
Pickett KM, Randle CP (2005) Strange Bayes Indeed: Uniform Topological Priors Imply Non-Uniform Clade Priors. Molecular Phylogenetics and Evolution 34:203
Pinney JW, Shirley MW, McConkey GA, Westhead DR (2005) Metashark: Software for Automated Metabolic Network Prediction from DNA Sequence and Its Application to the Genomes of Plasmodium Falciparum and Eimeria Tenella. Nucleic Acids Research 33:1399
Posada D, Buckley TR (2004) Model Selection and Model Averaging in Phylogenetics: Advantages of Akaike Information Criterion and Bayesian Approaches over Likelihood Ratio Tests. Syst Biol 53:793
Potter SC, Clarke L, Curwen V, Keenan S, Mongin E, Searle SM, Stabenau A, Storey R, Clamp M (2004) The Ensembl Analysis Pipeline. Genome Res 14:934
Prasad TSK, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, Balakrishnan L, Marimuthu A, Banerjee S, Somanathan DS, Sebastian A, Rani S, Ray S, Kishore CJH, Kanth S, Ahmed M, Kashyap MK, Mohmood R, Ramachandra YL, Krishna V, Rahiman BA, Mohan S, Ranganathan P, Ramabadran S, Chaerkady R, Pandey A (2009) Human Protein Reference Database-2009 Update. Nucleic Acids Research 37:D767
Pressman R (2001) Software Engineering: A Practioners Approach. McGraw-Hill
Pruitt KD, Tatusova T, Maglott DR (2005) Ncbi Reference Sequence (Refseq): A Curated Non-Redundant Sequence Database of Genomes, Transcripts and Proteins. Nucleic Acids Res 33:D501
Qi YJ, Bar-Joseph Z, Klein-Seetharaman J (2006) Evaluation of Different Biological Data and Computational Classification Methods for Use in Protein Interaction Prediction. Proteins-Structure Function and Bioinformatics 63:490
Quackenbush J (2002) Microarray Data Normalization and Transformation. Nat Genet 32 Suppl:496
Raab JR, Kamakaka RT (2010) Opinion Insulators and Promoters: Closer Than We Think. Nature Reviews Genetics 11:439
Radicchi F, Castellano C, Cecconi F, Loreto V, Parisi D (2004) Defining and Identifying Communities in Networks. Proceedings of the National Academy of Sciences of the United States of America 101:2658
Radom-Aizik S, Hayek S, Shahar I, Rechavi G, Kaminski N, Ben-Dov I (2005) Effects of Aerobic Training on Gene Expression in Skeletal Muscle of Elderly Men. Medicine and Science in Sports and Exercise 37:1680

Ralston KS, Kabututu ZP, Melehani JH, Oberholzer M, Hill KL (2009) The Trypanosoma Brucei Flagellum: Moving Parasites in New Directions. Annual Review of Microbiology 63:335
Ramachandran N, Hainsworth E, Bhullar B, Eisenstein S, Rosen B, Lau AY, Walter JC, LaBaer J (2004) Self-Assembling Protein Microarrays. Science 305:86
Ramazzina I, Folli C, Secchi A, Berni R, Percudani R (2006) Completing the Uric Acid Degradation Pathway through Phylogenetic Comparison of Whole Genomes. Nature Chemical Biology 2:144
Ranea JA, Yeats C, Grant A, Orengo CA (2007) Predicting Protein Function with Hierarchical Phylogenetic Profiles: The Gene3d Phylo-Tuner Method Applied to Eukaryotic Genomes. PLoS Comput Biol 3:e237
R Development Core Team (2011) R: A language and environment for statistical computing. R Foundation for Statistical Computing Vienna, Austria
Reghunathan R, Jayapal M, Hsu LY, Chng HH, Tai D, Leung BP, Melendez AJ (2005) Expression Profile of Immune Response Genes in Patients with Severe Acute Respiratory Syndrome. BMC Immunology 6
Remm M, Storm CE, Sonnhammer EL (2001) Automatic Clustering of Orthologs and inParalogs from Pairwise Species Comparisons. J Mol Biol 314:1041
Richmond TJ, Davey CA (2003) The Structure of DNA in the Nucleosome Core. Nature 423:145
Ridley M (1983) The Explanation of Organic Diversity : The Comparative Method and Adaptions for Mating. Clarendon Press, Oxford
Robertson DL, Lovell SC (2009) Evolution in Protein Interaction Networks: CoEvolution, Rewiring and the Role of Duplication. Biochem Soc Trans 37:768
Rodriguez-Ezpeleta N, Brinkmann H, Burey SC, Roure B, Burger G, Loffelhardt W, Bohnert HJ, Philippe H, Lang BF (2005) Monophyly of Primary Photosynthetic Eukaryotes: Green Plants, Red Algae, and Glaucophytes. Curr Biol 15:1325
Rodriguez-Ezpeleta N, Brinkmann H, Burger G, Roger AJ, Gray MW, Philippe H, Lang BF (2007) Toward Resolving the Eukaryotic Tree: The Phylogenetic Positions of Jakobids and Cercozoans. Current Biology 17:1420
Rokas A, Williams BL, King N, Carroll SB (2003) Genome-Scale Approaches to Resolving Incongruence in Molecular Phylogenies. Nature 425:798
Russell SJ, Norvig P, Canny J (2003) Artificial Intelligence : A Modern Approach. Prentice Hall, Upper Saddle River, N.J.
Salemi M, Vandamme A-M (2003) The Phylogenetic Handbook : A Practical Approach to DNA and Protein Phylogeny. Cambridge University Press, Cambridge, U.K. ; New York
Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D (2004) The Database of Interacting Proteins: 2004 Update. Nucleic Acids Res 32:D449
Sanger F, Coulson AR, Friedmann T, Air GM, Barrell BG, Brown NL, Fiddes JC, Hutchison CA, Slocombe PM, Smith M (1978) Nucleotide-Sequence of Bacteriophage-PhiX174. J Mol Biol 125:225
Sanger F, Nicklen S, Coulson AR (1977) DNA Sequencing with Chain-Terminating Inhibitors. Proc Natl Acad Sci U S A 74:5463
Sankoff D (1975) Minimal Mutation Trees of Sequences. Siam Journal on Applied Mathematics 28:35
Sasaoka T, Kobayashi M (2000) The Functional Significance of She in Insulin Signaling as a Substrate of the Insulin Receptor. Endocrine Journal 47:373
Scott MS, Barton GJ (2007) Probabilistic Prediction and Ranking of Human ProteinProtein Interactions. BMC Bioinformatics 8:239

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Research 13:2498
Shortle D, Ackerman MS (2001) Persistence of Native-Like Topology in a Denatured Protein in $\mathbf{8}$ M Urea. Science 293:487
Siddall ME (1998) Success of Parsimony in the Four-Taxon Case: Long-Branch Repulsion by Likelihood in the Farris Zone. Cladistics-the International Journal of the Willi Hennig Society 14:209
Sillentullberg B (1988) Evolution of Gregariousness in Aposematic Butterfly Larvae - a Phylogenetic Analysis. Evolution 42:293
Simmons MP, Ochoterena H, Freudenstein JV (2002) Amino Acid Vs. Nucleotide Characters: Challenging Preconceived Notions. Molecular Phylogenetics and Evolution 24:78
Singh GP, Ganapathi M, Dash D (2007) Role of Intrinsic Disorder in Transient Interactions of Hub Proteins. Proteins-Structure Function and Bioinformatics 66:761
Slater GS, Birney E (2005) Automated Generation of Heuristics for Biological Sequence Comparison. Bmc Bioinformatics 6
Slonim N, Elemento O, Tavazoie S (2006) Ab Initio Genotype-Phenotype Association Reveals Intrinsic Modularity in Genetic Networks. Molecular Systems Biology
Smith TF, Waterman MS (1981) Identification of Common Molecular Subsequences. J Mol Biol 147:195
Sneath PHA, Sokal RR (1973) Numerical Taxonomy : The Principles and Practice of Numerical Classification. W. H. Freeman, San Francisco
Snel B, Bork P, Huynen MA (1999) Genome Phylogeny Based on Gene Content. Nat Genet 21:108
Sokal RR, Rohlf FJ (1995) Biometry : The Principles and Practice of Statistics in Biological Research. W.H. Freeman, New York
Spira A, Beane J, Shah V, Liu G, Schembri F, Yang XM, Palma J, Brody JS (2004) Effects of Cigarette Smoke on the Human Airway Epithelial Cell Transcriptome. Proceedings of the National Academy of Sciences of the United States of America 101:10143
Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M (2006) Biogrid: A General Repository for Interaction Datasets. Nucleic Acids Research 34:D535
Steel M, Penny D (2000) Parsimony, Likelihood, and the Role of Models in Molecular Phylogenetics. Mol Biol Evol 17:839
Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S, Timm J, Mintzlaff S, Abraham C, Bock N, Kietzmann S, Goedde A, Toksoz E, Droege A, Krobitsch S, Korn B, Birchmeier W, Lehrach H, Wanker EE (2005) A Human Protein-Protein Interaction Network: A Resource for Annotating the Proteome. Cell 122:957
Stevens PF, Augier A (1983) Augustin Augier's "Arbre Botanique" (1801), a Remarkable Early Botanical Representation of the Natural System. Taxon 32:203
Stewart CB, Schilling JW, Wilson AC (1987) Adaptive Evolution in the Stomach Lysozymes of Foregut Fermenters. Nature 330:401
Strachan T, Read AP (2004) Human Molecular Genetics. Garland Press, New York Stuart GW, Moffett K, Leader JJ (2002) A Comprehensive Vertebrate Phylogeny Using Vector Representations of Protein Sequences from Whole Genomes. Mol Biol Evol 19:554

Stumpf MP, Thorne T, de Silva E, Stewart R, An HJ, Lappe M, Wiuf C (2008) Estimating the Size of the Human Interactome. Proceedings of the National Academy of Sciences of the United States of America 105:6959
Sundquist A, Ronaghi M, Tang HX, Pevzner P, Batzoglou S (2007) Whole-Genome Sequencing and Assembly with High-Throughput, Short-Read Technologies. PLoS One 2
Swanson KW, Irwin DM, Wilson AC (1991) Stomach Lysozyme Gene of the Langur Monkey - Tests for Convergence and Positive Selection. J Mol Evol 33:418
Swofford DL, Maddison WP (1987) Reconstructing Ancestral Character States under Wagner Parsimony. Mathematical Biosciences 87:199
Swofford DL, Waddell PJ, Huelsenbeck JP, Foster PG, Lewis PO, Rogers JS (2001) Bias in Phylogenetic Estimation and Its Relevance to the Choice between Parsimony and Likelihood Methods. Systematic Biology 50:525
Takatsu H, Futatsumori M, Yoshino K, Yoshida Y, Shin HW, Nakayama K (2001) Similar Subunit Interactions Contribute to Assembly of Clathrin Adaptor Complexes and Copi Complex: Analysis Using Yeast Three-Hybrid System. Biochemical and Biophysical Research Communications 284:1083
Talavera G, Castresana J (2007) Improvement of Phylogenies after Removing Divergent and Ambiguously Aligned Blocks from Protein Sequence Alignments. Syst Biol 56:564
Tanaka R, Yi TM, Doyle J (2005) Some Protein Interaction Data Do Not Exhibit Power Law Statistics. Febs Letters 579:5140
Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA (2003) The Cog Database: An Updated Version Includes Eukaryotes. BMC Bioinformatics 4:41
Telford MJ (2004) Animal Phylogeny: Back to the Coelomata? Curr Biol 14:R274
Tian B, Nowak DE, Jamaluddin M, Wang SF, Brasier AR (2005) Identification of Direct Genomic Targets Downstream of the Nuclear Factor-Kappa B Transcription Factor Mediating Tumor Necrosis Factor Signaling. Journal of Biological Chemistry 280:17435
Tierney EP, Tulac S, Huang STJ, Giudice LC (2003) Activation of the Protein Kinase a Pathway in Human Endometrial Stromal Cells Reveals Sequential Categorical Gene Regulation. Physiological Genomics 16:47
Townsend JP, Lopez-Giraldez F, Friedman R (2008) The Phylogenetic Informativeness of Nucleotide and Amino Acid Sequences for Reconstructing the Vertebrate Tree. J Mol Evol 67:437
Valadkhan S, Jaladat Y (2010) The Spliceosomal Proteome: At the Heart of the Largest Cellular Ribonucleoprotein Machine. Proteomics 10: 4128

Vanacova S, Liston DR, Tachezy J, Johnson PJ (2003) Molecular Biology of the Amitochondriate Parasites, Giardia Intestinalis, Entamoeba Histolytica and Trichomonas Vaginalis. International Journal for Parasitology 33:235
Vanharanta S, Pollard PJ, Lehtonen HJ, Laiho P, Sjoberg J, Leminen A, Aittomaki K, Arola J, Kruhoffer M, Orntoft TF, Tomlinson IP, Kiuru M, Arango D, Aaltonen LA (2006) Distinct Expression Profile in Fumarate-Hydratase-Deficient Uterine Fibroids. Human Molecular Genetics 15:97
Velculescu VE, Zhang L, Zhou W, Polyak K, Basrai M, Bassett D, Hieter P, Vogelstein B, Kinzler KW (1997) Serial Analysis of Gene Expression (Sage). American Journal of Human Genetics 61:A36

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigo R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X (2001) The Sequence of the Human Genome. Science 291:1304
Vert JP (2002) A Tree Kernel to Analyse Phylogenetic Profiles. Bioinformatics 18 Suppl 1:S276
Vidalain PO, Boxem M, Ge H, Li S, Vidal M (2004) Increasing Specificity in HighThroughput Yeast Two-Hybrid Experiments. Methods 32:363
Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E (2009) Ensemblcompara Genetrees: Complete, Duplication-Aware Phylogenetic Trees in Vertebrates. Genome Res 19:327
von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, Foglierini M, Jouffre N, Huynen MA, Bork P (2005) String: Known and Predicted Protein-Protein Associations, Integrated and Transferred across Organisms. Nucleic Acids Res 33:D433
von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P (2002) Comparative Assessment of Large-Scale Data Sets of Protein-Protein Interactions. Nature 417:399
von Mering C, Zdobnov EM, Tsoka S, Ciccarelli FD, Pereira-Leal JB, Ouzounis CA, Bork P (2003) Genome Evolution Reveals Biochemical Networks and Functional Modules. Proc Natl Acad Sci U S A 100:15428
Walhout AJ, Sordella R, Lu X, Hartley JL, Temple GF, Brasch MA, Thierry-Mieg N, Vidal M (2000) Protein Interaction Mapping in C. Elegans Using Proteins Involved in Vulval Development. Science 287:116
Wall DP, Fraser HB, Hirsh AE (2003) Detecting Putative Orthologs. Bioinformatics 19:1710
Wang D, Hsieh M, Li WH (2005a) A General Tendency for Conservation of Protein Length across Eukaryotic Kingdoms. Mol Biol Evol 22:142
Wang H, Xu Z, Gao L, Hao B (2009a) A Fungal Phylogeny Based on 82 Complete Genomes Using the Composition Vector Method. BMC Evol Biol 9:195
Wang J, Xia Q, He X, Dai M, Ruan J, Chen J, Yu G, Yuan H, Hu Y, Li R, Feng T, Ye C, Lu C, Wang J, Li S, Wong GK, Yang H, Wang J, Xiang Z, Zhou Z, Yu J (2005b) Silkdb: A Knowledgebase for Silkworm Biology and Genomics. Nucleic Acids Res 33:D399
Wang Z, Gerstein M, Snyder M (2009b) Rna-Seq: A Revolutionary Tool for Transcriptomics. Nat Rev Genet 10:57
Watson JD, Crick FH (1953) Molecular Structure of Nucleic Acids; a Structure for Deoxyribose Nucleic Acid. Nature 171:737
Watts DJ, Strogatz SH (1998) Collective Dynamics of 'Small-World' Networks. Nature 393:440
Wheeler TJ, Kececioglu JD (2007) Multiple Alignment by Aligning Alignments. Bioinformatics 23:1559
Whelan S, Goldman N (2001) A General Empirical Model of Protein Evolution Derived from Multiple Protein Families Using a Maximum-Likelihood Approach. Molecular Biology and Evolution 18:691
Whitaker JW, McConkey GA, Westhead DR (2009) Prediction of Horizontal Gene Transfers in Eukaryotes: Approaches and Challenges. Biochem Soc Trans 37:792
Wodicka L, Dong H, Mittmann M, Ho MH, Lockhart DJ (1997) Genome-Wide Expression Monitoring in Saccharomyces Cerevisiae. Nat Biotechnol 15:1359
Woese CR, Kandler O, Wheelis ML (1990) Towards a Natural System of Organisms: Proposal for the Domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci U S A 87:4576
Wolf YI, Rogozin IB, Koonin EV (2004) Coelomata and Not Ecdysozoa: Evidence from Genome-Wide Phylogenetic Analysis. Genome Research 14:29
Wootton JC, Federhen S (1993) Statistics of Local Complexity in Amino-Acid-Sequences and Sequence Databases. Computers \& Chemistry 17:149
Wu G, Nie L, Zhang WW (2008) Integrative Analyses of Posttranscriptional Regulation in the Yeast Saccharomyces Cerevisiae Using Transcriptomic and Proteomic Data. Current Microbiology 57:18
Yakovchuk P, Protozanova E, Frank-Kamenetskii MD (2006) Base-Stacking and BasePairing Contributions into Thermal Stability of the DNA Double Helix (Vol 34, Pg 564, 2006). Nucleic Acids Res 34:1082
Yang Z (2006) Computational Molecular Evolution. Oxford University Press, Oxford
Yang Z (2008) Computational Molecular Evolution. Oxford University Press, Oxford
Yang ZH (1994) Maximum-Likelihood Phylogenetic Estimation from DNA-Sequences with Variable Rates over Sites - Approximate Methods. Journal of Molecular Evolution 39:306

Yang ZH, Kumar S, Nei M (1995) A New Method of Inference of Ancestral Nucleotide and Amino-Acid-Sequences. Genetics 141:1641
Yedavalli VSRK, Neuveut C, Chi YH, Kleiman L, Jeang KT (2004) Requirement of Ddx3 Dead Box Rna Helicase for Hiv-1 Rev-Rre Export Function. Cell 119:381
Yu HY, Braun P, Yildirim MA, Lemmens I, Venkatesan K, Sahalie J, Hirozane-Kishikawa T, Gebreab F, Li N, Simonis N, Hao T, Rual JF, Dricot A, Vazquez A, Murray RR, Simon C, Tardivo L, Tam S, Svrzikapa N, Fan CY, de Smet AS, Motyl A, Hudson ME, Park J, Xin XF, Cusick ME, Moore T, Boone C, Snyder M, Roth FP, Barabasi AL, Tavernier J, Hill DE, Vidal M (2008) High-Quality Binary Protein Interaction Map of the Yeast Interactome Network. Science 322:104
Yu HY, Luscombe NM, Lu HX, Zhu XW, Xia Y, Han JDJ, Bertin N, Chung S, Vidal M, Gerstein M (2004a) Annotation Transfer between Genomes: Protein-Protein Interologs and Protein-DNA Regulogs. Genome Research 14:1107
Yu HY, Luscombe NM, Lu HX, Zhu XW, Xia Y, Han JDJ, Bertin N, Chung S, Vidal M, Gerstein M (2004b) Annotation Transfer between Genomes: Protein-Protein Interologs and Protein-DNA Regulogs. Genome Res 14:1107
Zhang JZ (2003) Evolution by Gene Duplication: An Update. Trends in Ecology \& Evolution 18:292
Zheng Q, Wang XJ (2008) Goeast: A Web-Based Software Toolkit for Gene Ontology Enrichment Analysis. Nucleic Acids Research 36:W358
Zhu H, Bilgin M, Bangham R, Hall D, Casamayor A, Bertone P, Lan N, Jansen R, Bidlingmaier S, Houfek T, Mitchell T, Miller P, Dean RA, Gerstein M, Snyder M (2001) Global Analysis of Protein Activities Using Proteome Chips. Science 293:2101
Zhu X, Gerstein M, Snyder M (2007) Getting Connected: Analysis and Principles of Biological Networks. Genes Dev 21:1010

## Appendix A Description of divergence of Java implementation of Inparanoid algorithm from Perl implementation.

In order to examine the differences in output between the novel Java implementation of the Inparanoid algorithm and the version (2.0) distributed by Remm et al. (Remm et al. 2001) the following test was run.

Organism A: Saccharomyces cerevisiae.
Organism B: Encephalitozoon cuniculi.

The algorithm BLASTP (version 2.2.18) was run on with the Fasta formatted file containing all proteins for Saccharomyces cerevisiae as the query and the Fasta formatted file containing all proteins for Encephalitozoon cuniculi as the database. The program formatdb was used to create parsable input for BLASTP.

The substitution matrix BLOSUM62 was used to score alignments. The converse command was also run with Encephalitozoon cuniculi as the query and Saccharomyces cerevisiae as the database. The two organisms were also run against themselves as query and database. The parameters $v$ and $b$ were set to the number of proteins in the database files and the parameter z is set to a theoretical maximum database size by (Remm et al. 2001) to maintain consistent values for relevant statistics such as K and $\lambda$ described below. The exact syntax of the commands is given below:
blastall -i Saccharomyces_cerevisiae -d Saccharomyces_cerevisiae -p blastp -v 5883 -b 5883 -F "m S" -M BLOSUM62 -z 5000000 -V
blastall -i Saccharomyces_cerevisiae -d Encephalitozoon_cuniculi -p blastp -v 1996 -b 1996 F "m S" -M BLOSUM62 -z 5000000 -V
blastall -i Encephalitozoon_cuniculi -d Saccharomyces_cerevisiae -p blastp -v 5883 -b 5883 F "m S" -M BLOSUM62 -z 5000000 -V
blastall -i Encephalitozoon_cuniculi -d Encephalitozoon_cuniculi -p blastp -v 1996 -b 1996 F "m S" -M BLOSUM62 -z 5000000 -V

The output from these commands was fed to the Perl script blast_parser.pl provided in the Inparanoid package, which produces formatted output in the following order:

- Protein Id1.
- Protein Id2.
- Bit Score.
- E value.
- Protein A Length.
- Protein B Length.
- Alignment Length on query sequence.
- Identity percentage.
- Similarity percentage.
- Coordinates of alignment on query sequence.

Blast bit scores are calculated using the formula
$S^{\prime}=\frac{\lambda S-\ln K}{\ln 2}$
$S=$ bit score, $\mathrm{S}=$ raw score, $\mathrm{K}=$ constant associated with search space size and $\lambda=$ constant associated with scoring system(Mount 2004).

Blast bit scores can be affected by composition based score adjustments which were introduced in order to deal with comparisons of proteins with highly biased amino acid compositions (Altschul et al. 2005).

This and variable database sizes (depending on whether the order of database and query is reversed due to search space size variation) can lead to an asymmetry in bit scores for the same pair of sequences. In order to deal with this artefact scores are normalised by both implementations by averaging the A-B and B-A orientations.

Results were filtered for hits containing only single high scoring pairs as the Java implementation was constructed to deal with SSEARCH output, which only returns a single optimal local alignment.

## Reciprocal best hits as marked by Perl implementation but not by Java implementation.

## Protein pair 1

Protein A: NP_015092 (Saccharomyces cerevisiae )
Protein B: NP_586462 (Encephalitozoon cuniculi)
A-B bit score $=56.2$ (Rounded down to 56 by Perl). This is the best score in the A-B direction.

B-A bit score $=55.5$ (Rounded up to 56 by Perl).
Mean bit score $=55.85$ (Rounded to 56 by Perl).
NP_586462 however has another significant score against Saccharomyces cerevisiae, which is NP_013908 with a mean bit score of 56.2 (Rounded down to 56 by Perl).

The Java implementation does not recognise NP_015092 and NP_586462 as reciprocal best hits as $56.2>55.85$.

## Protein pair 2

Protein A: NP_014520 (Saccharomyces cerevisiae ).
Protein B: NP_586468 (Encephalitozoon cuniculi).
A-B Bit score $=61.6$ (Rounded up to 62 by Perl).
B-A Bit score $=61.6$ (Rounded up to 62 by Perl).
Mean bit score $=61.6$ (Rounded up to 62 by Perl).
NP_586468 has another significant score against Saccharomyces cerevisiae, which is NP_014097 with a mean bit score of 62.0.

The Java implementation does not recognise NP_014520 and NP_586468 as reciprocal best hits as $62.0>61.6$.

## Protein pair 3

Protein A: NP_013648 (Saccharomyces cerevisiae ).
Protein B: NP_597364 (Encephalitozoon cuniculi).
A-B Bit score $=111.0$
B-A Bit score $=112.0$
Mean bit score $=111.5$ (Rounded up to 112 by Perl).
NP_597364 has another significant score against Saccharomyces cerevisiae, which is
NP_013546 with a mean bit score of 112.0

The Java implementation does not recognise NP_013648 and NP_597364 as reciprocal best hits as $112.0>111.5$.

## Protein pair 4

Protein A: NP_013182 (Saccharomyces cerevisiae ).
Protein B: NP_586039 (Encephalitozoon cuniculi).
A-B Bit score $=60.5$
B-A Bit score $=60.5$
Mean bit score $=60.5$ (Rounded up to 61 by Perl).
NP_586039 has two other significant scores against Saccharomyces cerevisiae, which are NP_010629 and NP_010630, which both have mean bit scores of 60.85 .
The Java implementation does not recognise NP_013648 and NP_597364 as reciprocal best hits as $60.85>60.5$.

## Protein pair 5

Protein A: NP_010407 (Saccharomyces cerevisiae).
Protein B: NP_597607 (Encephalitozoon cuniculi).
A-B Bit score $=246.0$
B-A Bit score $=245.0$
Mean bit score $=245.5$ (Rounded up to 246 by Perl).
NP_597607 has another significant score against Saccharomyces cerevisiae, which is NP_013197, which has a mean bit score of 246.

The Java implementation does not recognise NP_013648 and NP_597364 as reciprocal best hits as $246>245.5$.

## Differences in Cluster Output

There are a number of groups which differ between the two implementations on this test data. This is due to different scores being stored for various values affecting the criterion for reciprocal bests as well as the criteria for merging and deleting clusters. However the primary purpose for ortholog selection in this project, which is detection of presence and absence of proteins, is achieved, as the number of Saccharomyces cerevisiae proteins found to be present in Encephalitozoon cuniculi was identical.

## Groups, which differ between implementations.

There are 16 groups, which differ between the two implementations. The Java implementation produces 616 groups while the Perl implementation produces 619 .

## Orthologous Group 1

Perl Inparanoid implementation
NP_009501 NP_586181
NP_014887

Java implementation clusters NP_009501 and NP_014887 with a separate protein XP_955683.

## Orthologous Group 2

Perl Inparanoid implementation
NP_011424 NP_586425

## Orthologous Group 3

Perl Inparanoid implementation
NP_012263 NP_597203

Groups 2 and 3 are merged into one group by the Java implementation.

## Orthologous Group 4

Perl Inparanoid implementation
NP_012610. XP_955636
NP_010056.
NP_009928.
NP_010504.

Group 4 does not contain NP_009928 in output from the Java implementation.

## Orthologous group 5

Perl Inparanoid implementation
NP_012710.
NP 597625
NP_014074.

## Orthologous group 6

Perl Inparanoid implementation
NP_014293. NP_597270
NP_014752.
NP_012264.

Groups 5 and 6 are merged into one group by the Java implementation.

## Orthologous group 7

Perl Inparanoid implementation
NP_011573. XP_965975
NP_011975.
NP_013418.

Orthologous group 7 has an additional paralog added in Saccharomyces cerevisiae by the Java implementation NP_009928.

## Orthologous group 8

Perl Inparanoid implementation
NP_011651. NP_597477

Orthologous group 8 has an additional paralog added in Saccharomyces cerevisiae by the Java implementation NP_014604.

## Orthologous Group 9

Perl Inparanoid implementation
NP_013618. NP_597286
NP_014604.

Orthologous group 9 has an additional paralog added in Saccharomyces cerevisiae by the Java implementation NP_015007. This paralog replaces NP_014604.

## Orthologous Group 10

Perl Inparanoid implementation
NP_014045. NP_586473
NP_015007.

Java implementation clusters NP_014045. and NP_015007. with a separate protein NP_597286.

## Orthologous Group 11

Perl Inparanoid implementation

$$
\begin{array}{ll}
\text { NP_012603. } & \text { NP_584802 } \\
& \text { NP_597429 }
\end{array}
$$

Group 11 does not contain NP_597429 in output from the Java implementation.

## Orthologous Group 12

Perl Inparanoid implementation
NP_010144.
NP_597320
NP_010089.
NP_586125
NP_014737.

## Orthologous Group 13

Perl Inparanoid implementation
NP_009723. NP_597558
NP_015274.

Orthologous groups 12 and 13 are merged into one group by the Java implementation.

## Orthologous Group 14

Perl Inparanoid implementation

NP_009800
NP_010629
NP_010630
NP_013182
NP_011960
NP_012316
NP_014486
NP_010632
NP_011962
NP_012321
NP_011964
NP_013724
NP_116644
NP_010845
NP_014470
NP_010036
NP_012692
NP_011411
NP_014081
NP_010087
NP_010143
NP_010825
NP_014538
NP_010785
NP_010675
NP_116613
NP_011805
NP_010034
NP_012694
NP_009857

NP_010082

Orthologous group 14 has an additional paralog added in Saccharomyces cerevisiae by the Perl implementation NP_010082.

## Orthologous Group 15

Perl Inparanoid implementation
NP_012710. NP 597625
NP_014074.

## Orthologous Group 16

Perl Inparanoid implementation
NP_014293. NP_597270
NP_014752.
NP_012264.

Orthologous groups 15 and 16 are merged into one group by the Java implementation.

Appendix B Individual Gene trees for genes in super matrix utilised in construction of
Phylogeny


Gene RPL23: 60S ribosomal protein L23.


Gene RPS8: 40S ribosomal protein S8.


Gene SRP54:signal recognition particle 54 kDa protein.


Gene ERCC3: TFIIH basal transcription factor complex helicase XPB.


Gene KARS: lysyl-tRNA synthetase.


Gene METAP2: methionine aminopeptidase 2


Gene ATP6V1D: V-type proton ATPase subunit D.


Gene PSMC1: 26S protease regulatory subunit 4.


Gene NFS1:cysteine desulfurase, mitochondrial precursor.


Gene GARS: glycyl-tRNA syntheta

## Appendix C: Predictions made by constrained ML

| Protein 1 | Description |  |  |
| :--- | :--- | :--- | :--- |
| Protein 2 | Description |  |  |
|  |  |  |  |
|  | apolipoprotein A-I binding protein |  | prar to adaptor-related protein complex 1 |


|  | transcription elongation factor A protein |  | PREDICTED: similar to large subunit |
| :---: | :---: | :---: | :---: |
| 4507385 | 2 isoform a | 113427529 | ribosomal protein L36a |
| 4557896 | myotubularin | 44680154 | myotubularin-related protein 2 isoform 1 |
|  | transcription elongation factor A protein |  | small nuclear ribonucleoprotein polypeptide |
| 4507385 | 2 isoform a | 4507129 | E |
|  | transcription elongation factor A protein |  | RNA, U3 small nucleolar interacting protein |
| 4507385 | 2 isoform a | 4759276 | 2 |
|  | transcription elongation factor A protein |  |  |
| 4507385 | 2 isoform a | 116812591 | RER1 retention in endoplasmic reticulum 1 |
|  | transcription elongation factor A protein |  |  |
| 4507385 | 2 isoform a | 4557719 | DNA ligase I |
|  | transcription elongation factor A protein |  |  |
| 4507385 | 2 isoform a | 56549681 | small CTD phosphatase 3 isoform 2 |
| 4557896 | myotubularin | 18491016 | exonuclease 1 isoform b |
|  | transcription elongation factor A protein |  | phosphatidylinositol glycan anchor |
| 4507385 | 2 isoform a | 4758922 | biosynthesis, class L |
|  | transcription elongation factor A protein |  |  |
| 4507385 | 2 isoform a | 4506233 | proteasome 26 S non-ATPase subunit 8 |
|  | transcription elongation factor A protein |  |  |
| 4507385 | 2 isoform a | 7019319 | activator of basal transcription 1 |
|  | transcription elongation factor A protein |  | DNA directed RNA polymerase II |
| 4507385 | 2 isoform a | 10863925 | polypeptide L |
|  | transcription elongation factor A protein |  | dehydrodolichyl diphosphate synthase |
| 4507385 | 2 isoform a | 45580738 | isoform b |
|  | transcription elongation factor A protein |  |  |
| 4507385 | 2 isoform a | 150170706 | anaphase promoting complex subunit 10 |
| 4557896 | myotubularin | 19923424 | myotubularin-related protein 9 |
|  | transcription elongation factor A protein |  |  |
| 4507385 | 2 isoform a | 8923942 | nucleolar protein family A, member 3 |
| 4507385 | transcription elongation factor A protein | 40254869 | pre-mRNA processing factor 31 homolog |


| 2 isoform a |  |  |  |
| :---: | :---: | :---: | :---: |
| 4507385 | transcription elongation factor A protein |  |  |
|  | 2 isoform a | 41327715 | p53-related protein kinase |
| transcription elongation factor A protein |  |  |  |
| 4507385 | 2 isoform a | 4507311 | suppressor of Ty 4 homolog 1 |
| 4507385 | transcription elongation factor A protein |  | PREDICTED: similar to large subunit |
|  | 2 isoform a | 113427044 | ribosomal protein L36a |
| transcription elongation factor A protein |  |  |  |
| 4507385 | 2 isoform a | 4506651 | ribosomal protein L36a-like protein |
|  |  |  | PREDICTED: similar to peptidylprolyl |
| 4557896 | myotubularin | 113429091 | isomerase A isoform 1 |
| 4557896 | myotubularin | 113414586 | PREDICTED: similar to CG17293-PA |
|  | transcription elongation factor A protein |  | DNA directed RNA polymerase II |
| 4507385 | 2 isoform a | 4505947 | polypeptide G |
|  | phosphoribosyl pyrophosphate | 28557709 | phosphoribosyl pyrophosphate synthetase 1- |
| 4506127 | synthetase 1 |  | like 1 |
|  |  |  | PREDICTED: similar to adaptor-related |
| 153791910 | hypothetical protein LOC79868 | 89042891 | protein complex 1 sigma 2 subunit |
| 4506541 | retinaldehyde binding protein 1 | 4557719 | DNA ligase I |
| 38348232 | dual specificity phosphatase 7 | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
|  | pre-mRNA processing factor 31 |  |  |
| 40254869 | homolog | 113414586 | PREDICTED: similar to CG17293-PA |
|  | pre-mRNA processing factor 31 |  | PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein) |
| 40254869 | homolog | 113418826 | (SIG-20) |
|  | pre-mRNA processing factor 31 |  | PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein) |
| 40254869 | homolog | 113431146 | (SIG-20) |
| 40254869 |  | 113429091 |  |
|  | pre-mRNA processing factor 31 |  | PREDICTED: similar to peptidylprolyl |



| 7662482 | transmembrane protein 15 | 4557719 | DNA ligase I |
| :---: | :---: | :---: | :---: |
| 4826675 | cyclin-dependent kinase 5 | 38201680 | meiotic recombination protein SPO11 isoform b |
| 4826675 | cyclin-dependent kinase 5 | 113429091 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 |
| 4826675 | cyclin-dependent kinase 5 | 4557719 | DNA ligase I |
| 4507213 | signal recognition particle 19 kDa | 113414586 | PREDICTED: similar to CG17293-PA |
| 5453954 | delta isoform of regulatory subunit B56, protein phosphatase 2 A isoform 1 | 113414586 | PREDICTED: similar to CG17293-PA |
| 4826675 | cyclin-dependent kinase 5 | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 4826675 | cyclin-dependent kinase 5 | 89041736 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 35493987 | ubiquitin-conjugating enzyme E2I | 113429091 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 |
| 44917606 | N -ethylmaleimide-sensitive factor attachment protein, beta | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 44917606 | N -ethylmaleimide-sensitive factor attachment protein, beta | 4557719 | DNA ligase I |
| 16945972 | kelch domain containing 3 | 89041736 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 44917606 | N -ethylmaleimide-sensitive factor attachment protein, beta | 4505331 | N -ethylmaleimide-sensitive factor attachment protein, gamma |
| 35493987 | ubiquitin-conjugating enzyme E2I | 113414586 | PREDICTED: similar to CG17293-PA |
| 44917606 | N -ethylmaleimide-sensitive factor attachment protein, beta | 89041736 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 16945972 | kelch domain containing 3 | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 35493987 | ubiquitin-conjugating enzyme E2I | 38201680 | meiotic recombination protein SPO11 isoform b |


| 16945972 | kelch domain containing 3 | 7019405 | host cell factor C 2 |
| :---: | :---: | :---: | :---: |
| 133925811 | transportin 1 isoform 1 | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 133925811 | transportin 1 isoform 1 | 23510381 | transportin 1 isoform 2 |
|  | ras homolog gene family, member C |  | meiotic recombination protein SPO11 |
| 111494248 | precursor | 38201680 | isoform b |
|  | ras homolog gene family, member C |  | meiotic recombination protein SPO11 |
| 111494251 | precursor | 38201680 | isoform b |
|  | ras homolog gene family, member C |  | PREDICTED: similar to adaptor-related |
| 111494251 | precursor | 89042891 | protein complex 1 sigma 2 subunit |
|  | ras homolog gene family, member C |  |  |
| 111494251 | precursor | 4557719 | DNA ligase I |
|  | ras homolog gene family, member C |  |  |
| 111494251 | precursor | 118600973 | RNA binding motif protein, X-linked 2 |
|  | ras homolog gene family, member C |  |  |
| 111494248 | precursor | 118600973 | RNA binding motif protein, X-linked 2 |
|  | ras homolog gene family, member C |  | PREDICTED: similar to adaptor-related |
| 111494251 | precursor | 89041736 | protein complex 1 sigma 2 subunit |
|  |  |  | PREDICTED: similar to peptidylprolyl |
| 119943098 | dihydropyrimidine dehydrogenase | 113429091 | isomerase A isoform 1 |
|  | ras homolog gene family, member C |  | ras homolog gene family, member C |
| 111494251 | precursor | 111494248 | precursor |
|  | ras homolog gene family, member C |  |  |
| 111494251 | precursor | 4506717 | ribosomal protein S29 isoform 1 |
|  | ras homolog gene family, member C |  |  |
| 111494248 | precursor | 71772583 | ribosomal protein S29 isoform 2 |
|  | ras homolog gene family, member C |  |  |
| 111494251 | precursor | 71772583 | ribosomal protein S29 isoform 2 |
|  | ras homolog gene family, member C |  |  |
| 111494251 | precursor | 47717139 | leucine-zipper-like transcription regulator |


| 111494248 | precursor | 47717139 | leucine-zipper-like transcription regulator 1 |
| :---: | :---: | :---: | :---: |
|  | ras homolog gene family, member C |  | dehydrodolichyl diphosphate synthase |
| 111494248 | precursor | 45580738 | isoform b |
| 111494251 | ras homolog gene family, member C |  | dehydrodolichyl diphosphate synthase |
|  | precursor | 45580738 | isoform b |
|  | ras homolog gene family, member C |  |  |
| 111494251 | precursor | 38201710 | DEAD box polypeptide 17 isoform 1 |
|  | ras homolog gene family, member C |  |  |
| 111494248 | precursor | 38201710 | DEAD box polypeptide 17 isoform 1 |
|  | ras homolog gene family, member C |  |  |
| 111494248 | precursor | 56549681 | small CTD phosphatase 3 isoform 2 |
|  | ras homolog gene family, member C |  |  |
| 111494251 | precursor | 56549681 | small CTD phosphatase 3 isoform 2 |
|  | ras homolog gene family, member C |  |  |
| 111494248 | precursor | 14249398 | PHD-finger 5A |
|  | ras homolog gene family, member C |  |  |
| 111494251 | precursor | 14249398 | PHD-finger 5A |
|  | ras homolog gene family, member C |  | DNA directed RNA polymerase II |
| 111494251 | precursor | 10863925 | polypeptide L |
|  | ras homolog gene family, member C |  | PREDICTED: similar to adaptor-related |
| 111494248 | precursor | 89042891 | protein complex 1 sigma 2 subunit |
|  | ras homolog gene family, member C |  |  |
| 111494248 | precursor | 4557719 | DNA ligase I |
|  | ras homolog gene family, member C |  | PREDICTED: similar to adaptor-related |
| 111494248 | precursor | 89041736 | protein complex 1 sigma 2 subunit |
|  | ras homolog gene family, member C |  |  |
| 111494248 | precursor | 4502859 | CDC28 protein kinase 2 |
|  | ras homolog gene family, member C |  |  |
| 111494248 | precursor | 4506717 | ribosomal protein S29 isoform 1 |


|  | glucose-6-phosphate dehydrogenase |  | PREDICTED: similar to adaptor-related |
| :--- | :--- | :--- | :--- |
| 109389365 | isoform a | 89042891 | protein complex 1 sigma 2 subunit |


| 35493996 | ubiquitin-conjugating enzyme E2I | 113429091 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 |
| :---: | :---: | :---: | :---: |
| 35493996 | ubiquitin-conjugating enzyme E2I | 38201680 | meiotic recombination protein SPO11 isoform b |
| 35494003 | ubiquitin-conjugating enzyme E2I | 38201680 | meiotic recombination protein SPO11 isoform b |
| 31581534 | tRNA isopentenyltransferase 1 | 38201680 | meiotic recombination protein SPO11 isoform b |
|  | minor histocompatibility antigen 13 |  |  |
| 30581111 | isoform 3 | 4557719 | DNA ligase I |
| 31543831 | tubulin, gamma 1 | 6996005 | dynamin 1-like protein isoform 1 |
|  | minor histocompatibility antigen 13 |  |  |
| 30581111 | isoform 3 | 6996005 | dynamin 1-like protein isoform 1 |
| 31543831 | tubulin, gamma 1 | 4557719 | DNA ligase I |
| 31581534 | tRNA isopentenyltransferase 1 | 113429091 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 |
| 29826282 | protein phosphatase 1G | 4505999 | protein phosphatase 1G |
| 31581534 | tRNA isopentenyltransferase 1 | 113414586 | PREDICTED: similar to CG17293-PA |
| 30581111 | minor histocompatibility antigen 13 isoform 3 | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
|  | minor histocompatibility antigen 13 |  | PREDICTED: similar to adaptor-related |
| 30581111 | isoform 3 | 89041736 | protein complex 1 sigma 2 subunit |
| 19913408 | DNA topoisomerase II, beta isozyme | 113414586 | PREDICTED: similar to CG17293-PA |
| 19913408 | DNA topoisomerase II, beta isozyme | 113429091 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 |
| 19913408 | DNA topoisomerase II, beta isozyme | 38201680 | meiotic recombination protein SPO11 isoform b |
| 13236516 | Der1-like domain family, member 1 | 89041736 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |


| 13236516 | Der1-like domain family, member 1 | 113429091 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 |
| :---: | :---: | :---: | :---: |
| 13236516 | Der1-like domain family, member 1 | 113414586 | PREDICTED: similar to CG17293-PA |
| 13236516 | Der1-like domain family, member 1 | 38201680 | meiotic recombination protein SPO11 isoform b |
| 13236516 | Der1-like domain family, member 1 | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 10835049 | ras homolog gene family, member A | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 9955970 | ATP-binding cassette, sub-family C, member 3 | 113429091 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 |
| 10092619 | nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor, alpha | 89041736 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 5729877 | heat shock 70 kDa protein 8 isoform 1 | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
|  | proteasome 26S ATPase subunit 4 |  |  |
| 5729991 | isoform 1 | 4557719 | DNA ligase I |
| 5729991 | proteasome 26S ATPase subunit 4 isoform 1 | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 5729991 | proteasome 26S ATPase subunit 4 isoform 1 | 113429091 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 |
|  | proteasome 26S ATPase subunit 4 |  |  |
| 5729991 | isoform 1 | 71772583 | ribosomal protein S29 isoform 2 |
|  | proteasome 26S ATPase subunit 4 |  |  |
| 5729991 | isoform 1 | 4506717 | ribosomal protein S29 isoform 1 |
|  | proteasome 26S ATPase subunit 4 |  | DNA directed RNA polymerase II |
| 5729991 | isoform 1 | 10863925 |  |
| 5729877 | heat shock 70 kDa protein 8 isoform 1 | 89041736 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 5729991 | proteasome 26S ATPase subunit 4 | 8923942 | nucleolar protein family A, member 3 |

isoform 1

| 6005764 | GABA(A) receptor-associated protein | 09042891 | protein complex 1 sigma 2 subunit |
| :--- | :--- | ---: | :--- |
|  | proteasome 26S ATPase subunit 4 |  | PREDICTED: similar to adaptor-related |
| 5729991 | isoform 1 | 89041736 | protein complex 1 sigma 2 subunit |


| 153252132 | ribosomal protein L31 isoform 3 | 113414586 | PREDICTED: similar to CG17293-PA |
| :---: | :---: | :---: | :---: |
| 33286434 | p47 protein isoform c | 116256336 | SEC31 homolog A isoform 4 |
| 33286434 | p47 protein isoform c | 6996005 | dynamin 1-like protein isoform 1 |
| 30520314 | hypothetical protein LOC118812 | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 7657339 | molybdenum cofactor synthesis 3 | 113414586 | PREDICTED: similar to CG17293-PA |
| 151101386 | coenzyme Q10 homolog A isoform b | 151101384 | coenzyme Q10 homolog A isoform a |
| 73622130 | BolA-like protein 2 | 85797673 | bolA-like protein 2B |
| 7662010 | zinc finger protein 516 | 10190686 | zinc finger protein 286 |
|  | H2A histone family, member V isoform |  |  |
| 41406067 | 3 | 19718751 | uracil-DNA glycosylase isoform UNG2 |
|  |  |  |  |
| 149944735 | hypothetical protein LOC728937 | 89041601 | protein S26 isoform 1 |
| 149944735 | hypothetical protein LOC728937 | 15011936 | ribosomal protein S26 |
| 149944735 | hypothetical protein LOC728937 | 88980535 | PREDICTED: similar to 40S ribosomal protein S26 |
| 149944735 | hypothetical protein LOC728937 | 88982349 | PREDICTED: similar to 40S ribosomal protein S26 |
| 149944735 | hypothetical protein LOC728937 | 113420084 | PREDICTED: similar to 40S ribosomal protein S26 |
| 149944735 | hypothetical protein LOC728937 | 89025350 | PREDICTED: similar to 40S ribosomal protein S26 isoform 2 |
| 149944735 | hypothetical protein LOC728937 | 113430282 | PREDICTED: similar to 40S ribosomal protein S26 |
| 149944735 | hypothetical protein LOC728937 | 88987217 | PREDICTED: similar to 40S ribosomal protein S26 |
| 149944735 | hypothetical protein LOC728937 | 113429703 | PREDICTED: similar to 40S ribosomal protein S26 |
| 150010661 | SEC14-like 5 | 89042891 | similar to adaptorrelated |


|  |  |  | protein complex 1 sigma 2 subunit |
| :---: | :---: | :---: | :---: |
| 150170706 | anaphase promoting complex subunit 10 | 113429091 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 |
|  |  |  | meiotic recombination protein SPO11 |
| 38201710 | DEAD box polypeptide 17 isoform 1 | 38201680 | isoform b |
| 28626498 | kinesin family member C1 | 4557719 | DNA ligase I |
| 47778943 | syntaxin 16 isoform a | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
|  |  |  | PREDICTED: similar to adaptor-related |
| 38201710 | DEAD box polypeptide 17 isoform 1 | 89041736 | protein complex 1 sigma 2 subunit |
| 38201710 | DEAD box polypeptide 17 isoform 1 | 4758496 | H2A histone family, member Y isoform 2 |
| 38201710 | DEAD box polypeptide 17 isoform 1 | 113414586 | PREDICTED: similar to CG17293-PA |
|  |  |  | PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein) |
| 38201710 | DEAD box polypeptide 17 isoform 1 | 113418826 | (SIG-20) |
|  |  |  | PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein) |
| 38201710 | DEAD box polypeptide 17 isoform 1 | 113431146 | (SIG-20) |
| 38201710 | DEAD box polypeptide 17 isoform 1 | 113429091 | PREDICTED: similar to peptidylprolyl |
|  |  |  | isomerase A isoform 1 |
|  | excision repair cross-complementing |  |  |
| 15834617 | rodent repair deficiency, | 45580738 | dehydrodolichyl diphosphate synthase |
|  | complementation group 2 protein |  | isoform b |
|  | excision repair cross-complementing |  |  |
| 15834617 | rodent repair deficiency, |  |  |
|  | complementation group 2 protein | 8923942 | nucleolar protein family A, member 3 |
|  | excision repair cross-complementing |  |  |
|  | rodent repair deficiency, |  |  |
| 15834617 | complementation group 2 protein | 47717139 | leucine-zipper-like transcription regulator 1 |
| 15834617 | excision repair cross-complementing rodent repair deficiency, | 4557719 | DNA ligase I |



145275210 RNA processing factor 1
145275210 RNA processing factor 1
126723390 ankyrin repeat domain 24
124256496 heat shock 70kDa protein 1-like

126723390 ankyrin repeat domain 24
dual specificity phosphatase 27
122937243 (putative)

121582655 ankyrin repeat domain 35

118600973 RNA binding motif protein, X-linked 2

| 118600973 | RNA binding motif protein, X-linked 2 |
| :--- | :--- |
|  | trafficking protein particle complex 6B |
| 118600991 | isoform 1 |
|  | trafficking protein particle complex 6B |
| 118600991 | isoform 1 |
| 118498359 | ribosomal L1 domain containing 1 |

113431146

113414586

113429091

113429091

38201680

121582655

34419635

89041736

89042891

89041736

|  | PREDICTED: similar to 60S ribosomal |
| :--- | :--- |
| protein L26 (Silica-induced gene 20 protein) |  | (SIG-20)

PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein)

113431146 (SIG-20)

13129120 trafficking protein particle complex 6A meiotic recombination protein SPO11 isoform b

PREDICTED: similar to CG17293-PA

|  |  |  | PREDICTED: similar to peptidylprolyl |
| :--- | :--- | ---: | :--- |
| 118600973 | RNA binding motif protein, X-linked 2 | 113429091 | isomerase A isoform 1 |


|  | FGR | protein complex 1 sigma 2 subunit |  |
| :--- | :--- | :--- | :--- |
|  |  |  | PREDICTED: similar to Ubiquitin- <br> conjugating enzyme E2S (Ubiquitin- |
|  |  |  | conjugating enzyme E2-24 kDa) (Ubiquitin- <br> protein ligase) (Ubiquitin carrier protein) |
| 112382377 | ubiquitin-conjugating enzyme E2S | 113430896 | (E2-EPF5) |
|  | SEC24 (S. cerevisiae) homolog B |  | PREDICTED: similar to adaptor-related |


|  |  |  | protein complex 1 sigma 2 subunit |
| :---: | :---: | :---: | :---: |
| 109452595 | zinc finger protein 205 | 109452593 | zinc finger protein 205 |
| 109255245 | serine/threonine kinase 17a | 113414586 | PREDICTED: similar to CG17293-PA |
| 109255245 | serine/threonine kinase 17a | 6996005 | dynamin 1-like protein isoform 1 |
| 108773782 | ATP-binding cassette, sub-family E, member 1 | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 108773784 | ATP-binding cassette, sub-family E, member 1 | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 95147356 | mitogen-activated protein kinase 15 | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 95147356 | mitogen-activated protein kinase 15 | 89041736 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 94721250 | vesicle-associated membrane proteinassociated protein A isoform 1 | 89041736 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 93004102 | nuclear LIM interactor-interacting factor 2 | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
|  | nuclear LIM interactor-interacting factor |  | PREDICTED: similar to adaptor-related |
| 93004102 | 2 | 89041736 | protein complex 1 sigma 2 subunit |
| 93141204 | methyltransferase like 2B | 113414586 | PREDICTED: similar to CG17293-PA |
| 89145417 | methyltransferase like 7A | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
|  | eukaryotic translation initiation factor |  | PREDICTED: similar to peptidylprolyl |
| 84043963 | 5B | 113429091 | isomerase A isoform 1 |
| 77812674 | nucleolar protein family A, member 2 isoform b | 8923444 | nucleolar protein family A, member 2 isoform a |
| 77812670 | exosome component 9 isoform 2 | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 77812670 | exosome component 9 isoform 2 | 89041736 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |


|  | myosin head domain containing 1 |  | PREDICTED: similar to peptidylprolyl |
| :--- | :--- | ---: | :--- |
| 75812980 | isoform 3 | 113429091 | isomerase A isoform 1 |
|  | myosin head domain containing 1 |  | PREDICTED: similar to adaptor-related |
| 75812980 | isoform 3 | 89042891 | protein complex 1 sigma 2 subunit |


| 71772583 | ribosomal protein S29 isoform 2 | 4502743 | cyclin-dependent kinase 7 |
| :---: | :---: | :---: | :---: |
| 71772583 | ribosomal protein S29 isoform 2 | 4557719 | DNA ligase I |
|  |  |  | meiotic recombination protein SPO11 |
| 71772583 | ribosomal protein S29 isoform 2 | 38201680 | isoform b |
|  |  |  | PREDICTED: similar to peptidylprolyl |
| 71772583 | ribosomal protein S29 isoform 2 | 113429091 | isomerase A isoform 1 |
| 68509270 | transcriptional adaptor 2-like isoform a | 113414586 | PREDICTED: similar to CG17293-PA |
|  |  |  | PREDICTED: similar to adaptor-related |
| 68303635 | mutS homolog 3 | 89042891 | protein complex 1 sigma 2 subunit |
| 68226422 | Yip1 domain family, member 5 | 32401427 | Yip1 domain family, member 5 |
|  |  |  | PREDICTED: similar to adaptor-related |
| 62955833 | DNA-damage inducible protein 2 | 89042891 | protein complex 1 sigma 2 subunit |
|  | serologically defined colon cancer |  | PREDICTED: similar to adaptor-related |
| 64276486 | antigen 10 | 89042891 | protein complex 1 sigma 2 subunit |
|  | serologically defined colon cancer |  | PREDICTED: similar to adaptor-related |
| 64276486 | antigen 10 | 89041736 | protein complex 1 sigma 2 subunit |
| 62955833 | DNA-damage inducible protein 2 | 48717485 | DDI1, DNA-damage inducible 1, homolog 1 |
|  |  |  | PREDICTED: similar to peptidylprolyl |
| 62865890 | dual specificity phosphatase 5 | 113429091 | isomerase A isoform 1 |
|  |  |  | PREDICTED: similar to adaptor-related |
| 62460637 | importin 4 | 89041736 | protein complex 1 sigma 2 subunit |
|  |  |  | dehydrodolichyl diphosphate synthase |
| 62865890 | dual specificity phosphatase 5 | 45580738 | isoform b |
| 62865890 | dual specificity phosphatase 5 | 4557719 | DNA ligase I |
|  |  |  | ubiquitin-conjugating enzyme E2D 4 |
| 62865890 | dual specificity phosphatase 5 | 8393719 | (putative) |
|  |  |  | PREDICTED: similar to adaptor-related |
| 62865890 | dual specificity phosphatase 5 | 89042891 | protein complex 1 sigma 2 subunit |
| 62865890 | dual specificity phosphatase 5 | 8923942 | nucleolar protein family A, member 3 |


| 62865890 | dual specificity phosphatase 5 | 89041736 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| :---: | :---: | :---: | :---: |
| 62865890 | dual specificity phosphatase 5 | 56549681 | small CTD phosphatase 3 isoform 2 |
| 62240994 | cysteinyl-tRNA synthetase isoform d | 62240992 | cysteinyl-tRNA synthetase isoform c |
| 62234438 | Notchless gene homolog isoform b | 41350318 | myotubularin-related protein 2 isoform 2 |
| 62234438 | Notchless gene homolog isoform b | 44680154 | myotubularin-related protein 2 isoform 1 |
| 62234461 | Notchless gene homolog isoform a | 41350318 | myotubularin-related protein 2 isoform 2 |
| 62234461 | Notchless gene homolog isoform a | 62234438 | Notchless gene homolog isoform b |
| 62234438 | Notchless gene homolog isoform b | 4502703 | cell division cycle 6 protein |
| 62234438 | Notchless gene homolog isoform b | 21536371 | telomerase-associated protein 1 |
| 62234461 | Notchless gene homolog isoform a | 44680154 | myotubularin-related protein 2 isoform 1 |
| 58533179 | trafficking protein particle complex 2 | 7657548 | trafficking protein particle complex 2 |
| 60279265 | Sec61 gamma subunit | 38201680 | meiotic recombination protein SPO11 isoform b |
|  |  |  | meiotic recombination protein SPO11 |
| 58533179 | trafficking protein particle complex 2 | 38201680 | isoform b |
| 60279265 | Sec61 gamma subunit | 7657546 | Sec61 gamma subunit |
| 60279265 | Sec61 gamma subunit | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 58533179 | trafficking protein particle complex 2 | 113429091 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 |
| 58533179 | trafficking protein particle complex 2 | 113431146 | PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein) (SIG-20) |
|  |  |  | PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein) |
| 58533179 | trafficking protein particle complex 2 | 113418826 | (SIG-20) |
| 58533179 | trafficking protein particle complex 2 | 113414586 | PREDICTED: similar to CG17293-PA |


| 57165436 | serine/threonine kinase 16 | 57165434 | serine/threonine kinase 16 |
| :---: | :---: | :---: | :---: |
| 56549681 | small CTD phosphatase 3 isoform 2 | 88943062 | PREDICTED: similar to peptidylprolyl isomerase A (cyclophilin A)-like 4 |
| 56549681 | small CTD phosphatase 3 isoform 2 | 113423887 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 |
| 56549681 | small CTD phosphatase 3 isoform 2 | 89041736 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 56549681 | small CTD phosphatase 3 isoform 2 | 31543091 | RNA binding motif protein 13 |
| 56549681 | small CTD phosphatase 3 isoform 2 | 22035624 | phosphatidate cytidylyltransferase 1 |
| 56549683 | small CTD phosphatase 3 isoform 1 | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 56549681 | small CTD phosphatase 3 isoform 2 | 4502743 | cyclin-dependent kinase 7 |
| 56549681 | small CTD phosphatase 3 isoform 2 | 113418826 | PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein) (SIG-20) |
|  |  |  | PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein) |
| 56549681 | small CTD phosphatase 3 isoform 2 | 113431146 | (SIG-20) |
| 56549681 | small CTD phosphatase 3 isoform 2 | 113422777 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 |
| 56549681 | small CTD phosphatase 3 isoform 2 | 89042897 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 |
| 56549681 | small CTD phosphatase 3 isoform 2 | 38016127 | RNA binding motif protein 34 |
|  |  |  | meiotic recombination protein SPO 11 |
| 56549681 | small CTD phosphatase 3 isoform 2 | 38201680 | isoform b |
| 56549681 | small CTD phosphatase 3 isoform 2 | 5729840 | tubulin, gamma complex associated protein 2 |
| 56549681 | small CTD phosphatase 3 isoform 2 | 8393719 | ubiquitin-conjugating enzyme E2D 4 (putative) |
| 56549681 | small CTD phosphatase 3 isoform 2 | 88943041 | PREDICTED: similar to peptidylprolyl isomerase A (cyclophilin A)-like 4 |


| 56549681 | small CTD phosphatase 3 isoform 2 | 88953813 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 |
| :---: | :---: | :---: | :---: |
|  |  |  | PREDICTED: similar to TBC1 domain family member 3 (Rab GTPase-activating protein PRC17) (Prostate cancer gene 17 |
| 56549681 | small CTD phosphatase 3 isoform 2 | 113426831 | protein) (TRE17 alpha protein) isoform 1 |
| 56549681 | small CTD phosphatase 3 isoform 2 | 4557719 | DNA ligase I |
| 56550059 | CCR4-NOT transcription complex, subunit 4 isoform b | 56550057 | CCR4-NOT transcription complex, subunit isoform a |
| 56699411 | solute carrier family 35, member E2 | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 56549683 | small CTD phosphatase 3 isoform 1 | 89041736 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 56549681 | small CTD phosphatase 3 isoform 2 | 113429091 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 |
| 56549681 | small CTD phosphatase 3 isoform 2 | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 56549681 | small CTD phosphatase 3 isoform 2 | 4758496 | H2A histone family, member Y isoform 2 |
| 56549681 | small CTD phosphatase 3 isoform 2 | 6912680 | meiotic recombination protein SPO11 isoform a |
| 56699411 | solute carrier family 35, member E2 | 89041736 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 56549681 | small CTD phosphatase 3 isoform 2 | 62909985 | hypothetical protein LOC140711 |
| 55956895 | CGI-01 protein isoform 3 | 89041736 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 56549113 | debranching enzyme homolog 1 | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 56118223 | choline/ethanolaminephosphotransferase | 5174415 | choline/ethanolaminephosphotransferase |
| 56549113 | debranching enzyme homolog 1 | 89041736 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |


|  |  |
| :---: | :---: |
| 54792071 | 2 isoform b precursor |
|  | oxoglutarate (alpha-ketoglutarate) |
|  | dehydrogenase (lipoamide) isoform 1 |
| 51873036 | precursor |
| 51944950 | phosducin-like 2 |
| 50593537 | DnaJ (Hsp40) homolog, subfamily B, member 12 |
|  | phosphatidylinositol (4,5) bisphosphate |
| 50726960 | 5-phosphatase, A isoform 2 |
| 50658065 | SMC4 structural maintenance of chromosomes 4-like 1 |
| 50658063 | SMC4 structural maintenance of chromosomes 4-like 1 |
|  | phosphatidylinositol (4,5) bisphosphate |
| 50726960 | 5-phosphatase, A isoform 2 |
|  | phosphatidylinositol (4,5) bisphosphate |
| 50726960 | 5-phosphatase, A isoform 2 |
| 50658063 | SMC4 structural maintenance of chromosomes 4-like 1 |
| 50658065 | SMC4 structural maintenance of chromosomes 4-like 1 |
| 50409789 | APC11 anaphase promoting complex subunit 11 isoform 2 |
| 50409796 | APC11 anaphase promoting complex subunit 11 isoform 2 |
| 50409789 | APC11 anaphase promoting complex subunit 11 isoform 2 |
| 50409796 | APC11 anaphase promoting complex subunit 11 isoform 2 |

SMT3 suppressor of mif two 3 homolog 2 54792069 isoform a precursor

PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit PREDICTED: similar to CG17293-PA DnaJ (Hsp40) homolog, subfamily B, member 12

PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit skeletal muscle and kidney enriched inositol phosphatase isoform 2 PREDICTED: similar to peptidylprolyl 113429091 isomerase A isoform 1 PREDICTED: similar to peptidylprolyl 113429091 isomerase A isoform 1

APC11 anaphase promoting complex subunit 5040978111 isoform 2

APC11 anaphase promoting complex subunit $18777675 \quad 11$ isoform 2

APC11 anaphase promoting complex subunit 11 isoform 2

APC11 anaphase promoting complex subunit
$50409750 \quad 11$ isoform 2

APC11 anaphase promoting complex
subunit 11 isoform 2

APC11 anaphase promoting complex
subunit 11 isoform 2

APC11 anaphase promoting complex 50409804
subunit 11 isoform 2

APC11 anaphase promoting complex subunit 11 isoform 2

APC11 anaphase promoting complex
50409804
subunit 11 isoform 2

APC11 anaphase promoting complex
subunit 11 isoform 2

APC11 anaphase promoting complex 50409789
subunit 11 isoform 2

APC11 anaphase promoting complex
50409796 subunit 11 isoform 2
small nuclear ribonucleoprotein
50593002
polypeptide A'

NAD(P)H:quinone oxidoreductase type 49574502

3, polypeptide A2

50083277
ATPase class I type 8B member 4

APC11 anaphase promoting complex
50409781
subunit 11 isoform 2

APC11 anaphase promoting complex
50409781 subunit 11 isoform 2

APC11 anaphase promoting complex
50409750 subunit 11 isoform 2

NAD(P)H:quinone oxidoreductase type
3, polypeptide A2

APC11 anaphase promoting complex subunit
11 isoform 2

APC11 anaphase promoting complex subunit 11 isoform 2

APC11 anaphase promoting complex subunit 11 isoform 2

APC11 anaphase promoting complex subunit 11 isoform 2

APC11 anaphase promoting complex subunit
$50409781 \quad 11$ isoform 2

APC11 anaphase promoting complex subunit 11 isoform 2

APC11 anaphase promoting complex subunit 11 isoform 2

APC11 anaphase promoting complex subunit 11 isoform 2

PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit

PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit PREDICTED: similar to adaptor-related

89042891 protein complex 1 sigma 2 subunit

APC11 anaphase promoting complex subunit 11 isoform 2

APC11 anaphase promoting complex subunit 877767511 isoform 2

APC11 anaphase promoting complex subunit 11 isoform 2

PREDICTED: similar to adaptor-related
89041736 protein complex 1 sigma 2 subunit

| 48717485 | DDI1, DNA-damage inducible 1, homolog 1 |
| :---: | :---: |
| 48717485 | DDI1, DNA-damage inducible 1, homolog 1 |
| 47717139 | leucine-zipper-like transcription regulator 1 |
| 47717139 | leucine-zipper-like transcription regulator 1 |
| 47717139 | leucine-zipper-like transcription regulator 1 |
| 47717139 | leucine-zipper-like transcription regulator 1 |
|  | solute carrier family 25 member 3 |
| 47132595 | isoform b precursor |
|  | dehydrodolichyl diphosphate synthase |
| 45580738 | isoform b |
|  | dehydrodolichyl diphosphate synthase |
| 45580738 | isoform b |
|  | dehydrodolichyl diphosphate synthase |
| 45580738 | isoform b |
|  | dehydrodolichyl diphosphate synthase |
| 45580742 | isoform a |
|  | dehydrodolichyl diphosphate synthase |
| 45580742 | isoform a |
|  | dehydrodolichyl diphosphate synthase |
| 45580738 | isoform b |
| 45580738 | dehydrodolichyl diphosphate synthase isoform b |
| 45580738 | dehydrodolichyl diphosphate synthase isoform b |

89042891

89041736

6996005

38201680

4758496

4557719

45580738

113414586 45580738 isoform b

21536371 telomerase-associated protein 1

23397458 kinesin family member 19 dehydrodolichyl diphosphate synthase

113414586 PREDICTED: similar to CG17293-PA PREDICTED: similar to peptidylprolyl isomerase A isoform 1 PREDICTED: similar to peptidylprolyl 88943062 isomerase A (cyclophilin A)-like 4 PREDICTED: similar to peptidylprolyl 89042897 isomerase A isoform 1

PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit dynamin 1-like protein isoform 1 meiotic recombination protein SPO11 isoform b H2A histone family, member Y isoform 2

DNA ligase I
dehydrodolichyl diphosphate synthase isoform b

PREDICTED: similar to CG17293-PA

| 45580738 | dehydrodolichyl diphosphate synthase isoform b |
| :---: | :---: |
| 45580738 | dehydrodolichyl diphosphate synthase isoform b |
| 45580738 | dehydrodolichyl diphosphate synthase isoform b |
| 45580738 | dehydrodolichyl diphosphate synthase isoform b |
| 45580738 | dehydrodolichyl diphosphate synthase isoform b |
| 45580738 | dehydrodolichyl diphosphate synthase isoform b |
| 45580738 | dehydrodolichyl diphosphate synthase isoform b |
| 45580738 | dehydrodolichyl diphosphate synthase isoform b |
| 45580738 | dehydrodolichyl diphosphate synthase isoform b |
| 45580742 | dehydrodolichyl diphosphate synthase isoform a |
| 45580738 | dehydrodolichyl diphosphate synthase isoform b |
|  | dehydrodolichyl diphosphate synthase |
| 45580738 | isoform b |
|  | dehydrodolichyl diphosphate synthase |
| 45580738 | isoform b |
| 45580738 | dehydrodolichyl diphosphate synthase isoform b |
| 45580738 | dehydrodolichyl dipho |


|  | isoform b |  | protein PRC17) (Prostate cancer gene 17 <br> protein) (TRE17 alpha protein) isoform 1 |
| :--- | :--- | :--- | :--- |
|  | protein phosphatase 1, catalytic subunit, |  | protein phosphatase 1, catalytic subunit, beta |

\(\left.\begin{array}{llll} \& isoform b \& \& <br>
\& dehydrodolichyl diphosphate synthase \& \& PREDICTED: similar to peptidylprolyl <br>

45580738 \& isoform b \& 88953813 \& isomerase A isoform 1\end{array}\right]\)| PREDICTED: similar to peptidylprolyl |
| :--- |

|  |  |  | PREDICTED: similar to adaptor-related |
| :--- | :--- | :--- | :--- | :--- |
| 42516563 | UDP-glucuronate decarboxylase 1 | 89041736 | protein complex 1 sigma 2 subunit |
| 41872631 | fatty acid synthase |  | PREDICTED: similar to adaptor-related |


|  |  |  | protein complex 1 sigma 2 subunit |
| :---: | :---: | :---: | :---: |
| 41349441 | SEC31 homolog A isoform 2 | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 41327715 | p53-related protein kinase | 113414586 | PREDICTED: similar to CG17293-PA |
| 41349441 | SEC31 homolog A isoform 2 | 89041736 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
|  | ubiquitin-conjugating enzyme E2 |  | PREDICTED: similar to peptidylprolyl |
| 40806167 | variant 1 isoform a | 113429091 | isomerase A isoform 1 |
|  | ubiquitin-conjugating enzyme E2 |  |  |
| 40806167 | variant 1 isoform a | 113414586 | PREDICTED: similar to CG17293-PA |
|  | ubiquitin-conjugating enzyme E2 |  | meiotic recombination protein SPO11 |
| 40806167 | variant 1 isoform a | 38201680 | isoform b |
|  | transmembrane emp24 protein transport |  |  |
| 39725636 | domain containing 9 | 113414586 | PREDICTED: similar to CG17293-PA |
|  | transmembrane emp24 protein transport |  | PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein) |
| 39725636 | domain containing 9 | 113418826 | (SIG-20) |
|  | transmembrane emp24 protein transport |  | PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein) |
| 39725636 | domain containing 9 | 113431146 | (SIG-20) |
|  | transmembrane emp24 protein transport |  | PREDICTED: similar to peptidylprolyl |
| 39725636 | domain containing 9 | 113429091 | isomerase A isoform 1 |
| 38708309 | hypothetical protein LOC51029 | 113428755 | PREDICTED: similar to CG7222-PA |
| 38327644 | hypothetical protein LOC57707 | 89041736 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 38327644 | hypothetical protein LOC57707 | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 38327644 | hypothetical protein LOC57707 | 62909985 | hypothetical protein LOC140711 |
|  | meiotic recombination protein SPO11 |  | PREDICTED: similar to ribosomal protein |
| 38201680 | isoform b | 29742309 | L31 |



| 38201680 | meiotic recombination protein SPO11 isoform b | 11321585 | guanine nucleotide-binding protein, beta-1 subunit |
| :---: | :---: | :---: | :---: |
| 38201680 | meiotic recombination protein SPO11 isoform b | 21071060 | SWI/SNF-related matrix-associated actindependent regulator of chromatin a-like 1 |
| 38201680 | meiotic recombination protein SPO11 isoform b | 4506631 | ribosomal protein L30 |
| 38201680 | meiotic recombination protein SPO11 isoform b | 15150809 | SEC11-like 3 |
| 38201680 | meiotic recombination protein SPO11 isoform b | 7657546 | Sec61 gamma subunit |
| 38201680 | meiotic recombination protein SPO11 isoform b | 8923475 | thioredoxin-like 4B |
| 38201680 | meiotic recombination protein SPO11 isoform b | 113418682 | PREDICTED: similar to postmeiotic segregation increased 2-like 2 |
| 38201680 | meiotic recombination protein SPO11 isoform b | 4758384 | FK506 binding protein 5 |
| 38201680 | meiotic recombination protein SPO11 isoform b | 8922905 | RIO kinase 2 |
| 38201680 | meiotic recombination protein SPO 11 isoform b | 4506701 | ribosomal protein S23 |
| 38201680 | meiotic recombination protein SPO11 isoform b | 18105063 | vacuolar protein sorting 45A |
| 38201680 | meiotic recombination protein SPO11 isoform b | 113418826 | PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein) (SIG-20) |
| 38201680 | meiotic recombination protein SPO11 isoform b | 113431146 | PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein) (SIG-20) |
| 38201680 | meiotic recombination protein SPO11 isoform b | 8923942 | nucleolar protein family A, member 3 |
| 38201680 | meiotic recombination protein SPO11 | 4502643 | chaperonin containing TCP1, subunit 6A |


|  | isoform b |  | isoform a |
| :---: | :---: | :---: | :---: |
|  | meiotic recombination protein SPO11 |  | phosphatidylinositol glycan anchor |
| 38201680 | isoform b | 4758922 | biosynthesis, class L |
| 38201680 | meiotic recombination protein SPO11 |  |  |
|  | isoform b | 15431295 | ribosomal protein L13 |
|  | meiotic recombination protein SPO11 |  |  |
| 38201680 | isoform b | 15431297 | ribosomal protein L13 |
| 38201680 | meiotic recombination protein SPO11 |  | PREDICTED: similar to peptidylprolyl |
|  | isoform b | 113429091 | isomerase A isoform 1 |
|  | meiotic recombination protein SPO11 |  |  |
| 38201680 | isoform b | 7706667 | trafficking protein particle complex 4 |
|  | meiotic recombination protein SPO11 |  |  |
| 38201680 | isoform b | 4507311 | suppressor of Ty 4 homolog 1 |
|  |  |  | PREDICTED: similar to 60S ribosomal |
|  | meiotic recombination protein SPO11 |  | protein L29 (Cell surface heparin-binding |
| 38201680 | isoform b | 113428574 | protein HIP) |
|  | meiotic recombination protein SPO11 |  |  |
| 38201680 | isoform b | 32189369 | DNA polymerase epsilon subunit 2 |
|  | meiotic recombination protein SPO11 |  |  |
| 38201680 | isoform b | 4503729 | FK506-binding protein 4 |
|  | meiotic recombination protein SPO11 |  | PREDICTED: similar to 40S ribosomal |
| 38201680 | isoform b | 89035017 | protein S28 isoform 2 |
|  | meiotic recombination protein SPO11 |  |  |
| 38201680 | isoform b | 4507873 | von Hippel-Lindau binding protein 1 |
|  | meiotic recombination protein SPO11 |  | small nuclear ribonucleoprotein polypeptide |
| 38201680 | isoform b | 4507129 | E |
|  | meiotic recombination protein SPO11 |  | solute carrier family 2 (facilitated glucose |
| 38201680 | isoform b | 8923733 | transporter), member 6 |
|  | meiotic recombination protein SPO11 |  |  |
| 38201680 | isoform b | 113414586 | PREDICTED: similar to CG17293-PA |


|  | meiotic recombination protein SPO11 |  |  |
| :--- | :--- | ---: | :--- |
| 38201680 | isoform b | 4502859 | CDC28 protein kinase 2 |
|  | meiotic recombination protein SPO11 |  |  |
| 38201680 | isoform b | 4506609 | ribosomal protein L19 |
|  |  |  | CTD (carboxy-terminal domain, RNA |



|  | antigen 1 |  | polypeptide L |
| :--- | :--- | :--- | :--- |
| 32130516 | antigen 1 |  | skeletal muscle and kidney enriched inositol |
|  |  | 18765707 | phosphatase isoform 2 |


|  | phosphorylation regulated kinase 4 |  | protein complex 1 sigma 2 subunit |
| :---: | :---: | :---: | :---: |
| 28872761 | myotubularin-related protein 1 | 113414586 | PREDICTED: similar to CG17293-PA |
| 24762236 | PRP38 pre-mRNA processing factor 38 (yeast) domain containing A | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
|  | zinc finger, DHHC-type containing 14 |  | zinc finger, DHHC-type containing 14 |
| 24371272 | isoform 2 | 24371241 | isoform 1 |
| 24430186 | phosphatidylinositol glycan, class C | 4505795 | phosphatidylinositol glycan, class C |
|  |  |  | PREDICTED: similar to adaptor-related |
| 24430186 | phosphatidylinositol glycan, class C | 89042891 | protein complex 1 sigma 2 subunit |
| 23510381 | transportin 1 isoform 2 | 8923942 | nucleolar protein family A, member 3 |
| 23397458 | kinesin family member 19 | 8923942 | nucleolar protein family A, member 3 |
|  |  |  | DNA directed RNA polymerase II |
| 23397458 | kinesin family member 19 | 10863925 | polypeptide L |
| 23510381 | transportin 1 isoform 2 | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
|  |  |  | PREDICTED: similar to adaptor-related |
| 23397458 | kinesin family member 19 | 89042891 | protein complex 1 sigma 2 subunit |
| 23199991 | casein kinase 1 epsilon | 4503093 | casein kinase 1 epsilon |
|  |  |  | PREDICTED: similar to peptidylprolyl |
| 22202633 | prefoldin subunit 5 isoform alpha | 113429091 | isomerase A isoform 1 |
| 22035624 | phosphatidate cytidylyltransferase 1 | 113429091 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 |
|  |  |  | PREDICTED: similar to peptidylprolyl |
| 21624654 | spermatogenesis associated 5 | 113429091 | isomerase A isoform 1 |
| 21362110 | thiamin pyrophosphokinase 1 isoform a | 89041736 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 21624654 | spermatogenesis associated 5 | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 21362110 | thiamin pyrophosphokinase 1 isoform a | 89042891 | EDICTED: similar to adaptor-related |



|  |  |  | PREDICTED: similar to peptidylprolyl |
| :--- | :--- | ---: | :--- |
| 19913428 | vacuolar H+ATPase B2 | 113429091 | isomerase A isoform 1 |
| 19718751 | uracil-DNA glycosylase isoform UNG2 | 30795252 | autophagy-related cysteine endopeptidase 2 |
| isoform a |  |  |  |


| 15011936 | ribosomal protein S26 |  | PREDICTED: similar to 40S ribosomal |
| :--- | :--- | :--- | :--- |
|  |  | 88982349 | protein S26 |
| 15011936 | ribosomal protein S26 |  | PREDICTED: similar to 40S ribosomal |
|  |  | 113420084 | protein S26 |
| 15150809 | SEC11-like 3 |  | PREDICTED: similar to peptidylprolyl |
|  |  | 113429091 | isomerase A isoform 1 |


|  | isoform 1 |  | isomerase A isoform 1 |
| :--- | :--- | ---: | :--- |
|  | DNA directed RNA polymerase II |  | PREDICTED: similar to adaptor-related |
| 10863925 | polypeptide L | 89042891 | protein complex 1 sigma 2 subunit |
|  |  |  | PREDICTED: similar to peptidylprolyl |
| 10864021 | trafficking protein particle complex 1 | 113429091 | isomerase A isoform 1 |

polypeptide L

DNA directed RNA polymerase II
10863925
8923942 nucleolar protein family A, member 3
8923942 nucleolar protein family A, member 3
8923942 nucleolar protein family A, member 3
8923942 nucleolar protein family A, member 3
8923942 nucleolar protein family A, member 3

8923942 nucleolar protein family A, member 3

8923942 nucleolar protein family A, member 3

8923942 nucleolar protein family A, member 3

8923942 nucleolar protein family A, member 3

8923942 nucleolar protein family A, member 3

8923942 nucleolar protein family A, member 3 uncharacterized hypothalamus protein

8923712 HARP11

8923942 nucleolar protein family A, member 3 uncharacterized hypothalamus protein

8923712 HARP11

113414586 PREDICTED: similar to CG17293-PA
protein phosphatase 1, catalytic subunit, beta
isoform 1
cyclin-dependent kinase 7

DNA ligase I

H2A histone family, member Y isoform 2

RNA binding motif protein 34

PREDICTED: similar to TBC1 domain family member 3 (Rab GTPase-activating protein PRC17) (Prostate cancer gene 17 protein) (TRE17 alpha protein) isoform 1 PREDICTED: similar to CG17293-PA PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein) (SIG-20)

PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein) (SIG-20)

PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit PREDICTED: similar to peptidylprolyl isomerase A isoform 1

PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit

|  |  |  | PREDICTED: similar to adaptor-related |
| :--- | :--- | :--- | :--- |
| 7706326 | splicing factor 3B, 14 kDa subunit | 89042891 | protein complex 1 sigma 2 subunit |


| 7657546 | Sec61 gamma subunit | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| :---: | :---: | :---: | :---: |
| 7657548 | trafficking protein particle complex 2 | 113431146 | PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein) (SIG-20) |
|  |  |  | PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein) |
| 7657548 | trafficking protein particle complex 2 | 113418826 | (SIG-20) |
| 7657548 | trafficking protein particle complex 2 | 113414586 | PREDICTED: similar to CG17293-PA |
| 6912680 | meiotic recombination protein SPO11 isoform a | 113429091 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 |
|  | meiotic recombination protein SPO11 |  |  |
| 6912680 | isoform a | 4557719 | DNA ligase I |
| 6912680 | meiotic recombination protein SPO11 isoform a | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 6005701 | ATP-binding cassette, sub-family A member 8 | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 5902002 | dual specificity phosphatase 14 | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 6006001 | plasma glutathione peroxidase 3 precursor | 31542539 | DnaJ (Hsp40) homolog, subfamily A, member 3 |
| 6005701 | ATP-binding cassette, sub-family A member 8 | 89041736 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
|  | tubulin, gamma complex associated |  | PREDICTED: similar to peptidylprolyl |
| 5729840 | protein 2 | 113429091 |  |
|  | tubulin, gamma complex associated |  |  |
| 5729840 | protein 2 | 113414586 | PREDICTED: similar to CG17293-PA |
|  | tubulin, gamma complex associated |  |  |
| 5729840 | protein 2 | 6996005 | dynamin 1-like protein isoform 1 |
|  |  |  | PREDICTED: similar to Ubiquitin-63E |
| 5454144 | ubiquitin D | 113423966 | CG11624-PA, isoform A |



| 4758922 | phosphatidylinositol glycan anchor biosynthesis, class L | 113429091 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 |
| :---: | :---: | :---: | :---: |
|  | phosphatidylinositol glycan anchor |  | PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein) |
| 4758922 | biosynthesis, class L | 113431146 | (SIG-20) |
| 4758922 | phosphatidylinositol glycan anchor biosynthesis, class L | 113418826 | PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein) (SIG-20) |
| 4507873 | von Hippel-Lindau binding protein 1 | 113431146 | PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein) (SIG-20) |
|  |  |  | PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein) |
| 4507873 | von Hippel-Lindau binding protein 1 | 113418826 | (SIG-20) |
|  | excision repair cross-complementing rodent repair deficiency, |  |  |
| 4557563 |  | 9910180 | ACN9 homolog |
|  |  |  | PREDICTED: similar to peptidylprolyl |
| 4507947 | tyrosyl-tRNA synthetase | 113429091 | isomerase A isoform 1 |
| 4507873 | von Hippel-Lindau binding protein 1 | 113414586 | PREDICTED: similar to CG17293-PA |
|  | ubiquitin-conjugating enzyme E2 |  | PREDICTED: similar to peptidylprolyl |
| 4507797 | variant 2 | 113429091 | isomerase A isoform 1 |
|  |  |  | PREDICTED: similar to peptidylprolyl |
| 4507873 | von Hippel-Lindau binding protein 1 | 113429091 | isomerase A isoform 1 |
|  | ubiquitin-conjugating enzyme E2 |  |  |
| 4507797 | variant 2 | 113414586 | PREDICTED: similar to CG17293-PA |
| 4506701 | ribosomal protein S23 | 113429091 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 |
|  |  |  | PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein) |
| 4506699 | ribosomal protein S21 | 113418826 | (SIG-20) |
| 4506699 | ribosomal protein S21 | 113431146 |  |


|  |  |  | protein L26 (Silica-induced gene 20 protein) (SIG-20) |
| :---: | :---: | :---: | :---: |
| 4506717 | ribosomal protein S29 isoform 1 | 4758496 | H2A histone family, member Y isoform 2 |
| 4507047 | solute carrier family 7 (cationic amino acid transporter, $\mathrm{y}+$ system), member 1 | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 4506717 | ribosomal protein S29 isoform 1 | 113418826 | PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein) (SIG-20) |
|  |  |  | PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein) |
| 4506717 | ribosomal protein S29 isoform 1 | 113431146 | (SIG-20) |
| 4506699 | ribosomal protein S21 | 4758496 | H2A histone family, member Y isoform 2 |
| 4506717 | ribosomal protein S29 isoform 1 | 62909985 | hypothetical protein LOC140711 |
| 4506717 | ribosomal protein S29 isoform 1 | 113414586 | PREDICTED: similar to CG17293-PA |
| 4506699 | ribosomal protein S21 | 4502743 | cyclin-dependent kinase 7 |
| 4506717 | ribosomal protein S29 isoform 1 | 8393719 | ubiquitin-conjugating enzyme E2D 4 (putative) |
| 4506699 | ribosomal protein S21 | 113429091 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 |
|  | solute carrier family 7 (cationic amino |  | PREDICTED: similar to adaptor-related |
| 4507047 | acid transporter, $\mathrm{y}+$ system), member 1 | 89041736 | protein complex 1 sigma 2 subunit |
| 4506717 | ribosomal protein S29 isoform 1 | 38016127 | RNA binding motif protein 34 |
| 4506717 | ribosomal protein S29 isoform 1 | 4502743 | cyclin-dependent kinase 7 |
| 4506701 | ribosomal protein S23 | 113431146 | PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein) (SIG-20) |
|  |  |  | PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein) |
| 4506701 | ribosomal protein S23 | 113418826 | (SIG-20) |
| 4506701 | ribosomal protein S23 | 113414586 | PREDICTED: similar to CG17293-PA |


| 4506699 | ribosomal protein S21 | 113414586 | PREDICTED: similar to CG17293-PA |
| :--- | :--- | ---: | :--- |
|  | small nuclear ribonucleoprotein |  | PREDICTED: similar to adaptor-related |
| 4507123 | polypeptide B" | 89042891 | protein complex 1 sigma 2 subunit |
| 4506717 | ribosomal protein S29 isoform 1 | 4557719 | DNA ligase I |
|  |  | 113422526 | protein S28 isoform 1 |


| 4506005 | beta isoform 1 | 4557719 | DNA ligase I |
| :---: | :---: | :---: | :---: |
| 4506233 | proteasome 26 S non-ATPase subunit 8 | 113429091 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 |
| 4505621 | prostatic binding protein | 22165364 | mitochondrial ribosomal protein L38 |
| 4505795 | phosphatidylinositol glycan, class C | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 4504511 | DnaJ (Hsp40) homolog, subfamily A, member 1 | 89041736 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 4504511 | DnaJ (Hsp40) homolog, subfamily A, member 1 | 113429091 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 |
| 4504221 | guanylate kinase 1 | 113414586 | PREDICTED: similar to CG17293-PA |
| 4504007 | glycerol kinase isoform b | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 4504007 | glycerol kinase isoform b | 89041736 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
|  |  |  | PREDICTED: similar to peptidylprolyl |
| 4504221 | guanylate kinase 1 | 113429091 | isomerase A isoform 1 |
| 4502703 | cell division cycle 6 protein | 21536371 | telomerase-associated protein 1 |
| 4503301 | 2,4-dienoyl CoA reductase 1 precursor | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
|  | chaperonin containing TCP1, subunit |  | PREDICTED: similar to adaptor-related |
| 4502643 | 6A isoform a | 89042891 | protein complex 1 sigma 2 subunit |
| 63029935 | H2A histone family, member B3 | 89041736 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
|  | phosphoribosyl pyrophosphate |  | PREDICTED: similar to adaptor-related |
| 28557709 | synthetase 1-like 1 | 89042891 | protein complex 1 sigma 2 subunit |
| 63029935 | H2A histone family, member B3 | 63029943 | H2A histone family, member B2 |
| 28557709 | phosphoribosyl pyrophosphate synthetase 1-like 1 | 89041736 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |


| 66912162 | histone 2, H2bf | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| :---: | :---: | :---: | :---: |
| 148747574 | hypothetical protein LOC51030 | 21945058 | hypothetical protein LOC201158 |
| 63029935 | H2A histone family, member B3 | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 118402582 | cell division cycle 20 | 113429091 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 |
| 38016127 | RNA binding motif protein 34 | 89041736 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 38016127 | RNA binding motif protein 34 | 51467029 | PREDICTED: similar to 40S ribosomal protein S26 |
| 38016127 | RNA binding motif protein 34 | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 38016127 | RNA binding motif protein 34 | 4557719 | DNA ligase I |
| 37595752 | lamin B receptor | 37595750 | lamin B receptor |
| 38016127 | RNA binding motif protein 34 | 113429091 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 |
| 38348260 | ankyrin repeat domain 47 | 89041736 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 38016127 | RNA binding motif protein 34 | 6996005 | dynamin 1-like protein isoform 1 |
| 32189369 | DNA polymerase epsilon subunit 2 | 113414586 | PREDICTED: similar to CG17293-PA |
| 32189369 | DNA polymerase epsilon subunit 2 | 113429091 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 |
| 32189369 | DNA polymerase epsilon subunit 2 | 113431146 | PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein) (SIG-20) |
|  |  |  | PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein) |
| 32189369 | DNA polymerase epsilon subunit 2 | 113418826 | (SIG-20) |
| 32484973 | adenosine kinase isoform a | 113429091 | PREDICTED: similar to peptidylprolyl |



| 28559085 | cytidine triphosphate synthase II | 28559083 | cytidine triphosphate synthase II |
| :---: | :---: | :---: | :---: |
|  | autophagy-related cysteine |  |  |
| 30795252 | endopeptidase 2 isoform a | 113414586 | PREDICTED: similar to CG17293-PA |
|  | autophagy-related cysteine |  |  |
| 30795248 | endopeptidase 2 isoform b | 113414586 | PREDICTED: similar to CG17293-PA |
|  | autophagy-related cysteine |  | PREDICTED: similar to peptidylprolyl |
| 30795248 | endopeptidase 2 isoform b | 113429091 | isomerase A isoform 1 |
|  | autophagy-related cysteine |  | PREDICTED: similar to peptidylprolyl |
| 30795252 | endopeptidase 2 isoform a | 113429091 | isomerase A isoform 1 |
| 24586679 | testis-specific histone H2B | 89041736 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 24586675 | slingshot homolog 3 | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 24586679 | testis-specific histone H 2 B | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
|  | potassium voltage-gated channel, shaker-related subfamily, beta member |  | potassium voltage-gated channel, shaker- |
| 27436969 | 2 isoform 2 | 4504825 | related subfamily, beta member 2 isoform 1 |
| 22538446 | tumor protein p53 inducible protein 3 | 22538444 | tumor protein p53 inducible protein 3 |
| 22001417 | gemin 5 | 21536371 | telomerase-associated protein 1 |
|  | adaptor-related protein complex 1 sigma |  | PREDICTED: similar to adaptor-related |
| 22027655 | 2 subunit | 89041736 | protein complex 1 sigma 2 subunit |
|  | adaptor-related protein complex 1 sigma |  | PREDICTED: similar to adaptor-related |
| 22027655 | 2 subunit | 89042891 | protein complex 1 sigma 2 subunit |
| 21362084 | TBC1 domain family, member 15 | 89041736 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 21362084 | TBC1 domain family, member 15 | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
|  | IMP1 inner mitochondrial membrane |  |  |
| 21450679 | peptidase-like | 113414586 | PREDICTED: similar to CG17293-PA |


| 21314720 | Smad nuclear interacting protein | 113429091 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 |
| :---: | :---: | :---: | :---: |
| 20911035 | peptidylprolyl isomerase-like 4 | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 19923315 | serine hydroxymethyltransferase 2 (mitochondrial) | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 18860884 | WW domain-containing oxidoreductase isoform 2 | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 18860884 | WW domain-containing oxidoreductase isoform 2 | 7706523 | WW domain-containing oxidoreductase isoform 1 |
| 19923315 | serine hydroxymethyltransferase 2 <br> (mitochondrial) | 4557719 | DNA ligase I |
| 18860884 | WW domain-containing oxidoreductase isoform 2 | 89041736 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 15431297 | ribosomal protein L13 | 113429091 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 |
| 15431297 | ribosomal protein L13 | 113414586 | PREDICTED: similar to CG17293-PA |
| 16306568 | poly(A) polymerase gamma | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 16306566 | histone H2B | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 15431297 | ribosomal protein L13 | 113418826 | PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein) (SIG-20) |
| 15431297 | ribosomal protein L13 | 113431146 | PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein) (SIG-20) |
| 14251212 | DEAD (Asp-Glu-Ala-Asp) box polypeptide 20 | 113429091 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 |
| 13376747 | nucleotide binding protein-like | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |


| 14043026 | vesicle-associated membrane protein 8 | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| :---: | :---: | :---: | :---: |
| 14043026 | vesicle-associated membrane protein 8 | 113429091 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 |
| 13430872 | nucleolar protein 10 | 113414586 | PREDICTED: similar to CG17293-PA |
| 13430872 | nucleolar protein 10 | 113429091 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 |
| 13129120 | trafficking protein particle complex 6A | 113414586 | PREDICTED: similar to CG17293-PA |
| 11321585 | guanine nucleotide-binding protein, beta-1 subunit | 89041736 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 11321585 | guanine nucleotide-binding protein, beta-1 subunit | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 11056006 | kelch-like 12 | 89041736 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 11056006 | kelch-like 12 | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 12758125 | hypothetical protein LOC23378 | 89041736 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 12758125 | hypothetical protein LOC23378 | 113429091 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 |
| 12758125 | hypothetical protein LOC23378 | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 11386163 | ELAV-like 4 | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 8922905 | RIO kinase 2 | 113429091 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 |
|  | solute carrier family 2 (facilitated |  |  |
| 8923733 | glucose transporter), member 6 | 113414586 | PREDICTED: similar to CG17293-PA |
| 10190686 | zinc finger protein 286 | 10190696 | zinc finger protein 304 |
| 10190686 | zinc finger protein 286 | 18765707 | skeletal muscle and kidney enriched inos |


|  |  |  | phosphatase isoform 2 |
| :---: | :---: | :---: | :---: |
| 8923733 | solute carrier family 2 (facilitated glucose transporter), member 6 | 62909985 | hypothetical protein LOC140711 |
|  |  |  | PREDICTED: similar to zinc finger protein |
| 10190686 | zinc finger protein 286 | 113413881 | 114 |
| 8923733 | solute carrier family 2 (facilitated glucose transporter), member 6 | 113429091 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 |
| 8923733 | solute carrier family 2 (facilitated glucose transporter), member 6 | 113431146 | PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein) (SIG-20) |
|  | solute carrier family 2 (facilitated |  | PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein) |
| 8923733 | glucose transporter), member 6 | 113418826 | (SIG-20) |
| 8922905 | RIO kinase 2 | 113414586 | PREDICTED: similar to CG17293-PA |
| 10190686 | zinc finger protein 286 | 6996005 | dynamin 1-like protein isoform 1 |
| 7706343 | hypothetical protein LOC51647 | 113414586 | PREDICTED: similar to CG17293-PA |
| 7705477 | hypothetical protein LOC51504 | 113414586 | PREDICTED: similar to CG17293-PA |
| 7705748 | TNNI3 interacting kinase | 113429091 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 |
|  | U6 snRNA-associated Sm-like protein |  |  |
| 7706423 | LSm7 | 113414586 | PREDICTED: similar to CG17293-PA |
|  |  |  | PREDICTED: similar to adaptor-related |
| 7705748 | TNNI3 interacting kinase | 89041736 |  |
| 7706497 | cytidylate kinase | 113414586 | PREDICTED: similar to CG17293-PA |
| 7705369 | coatomer protein complex, subunit beta | 113414586 | PREDICTED: similar to CG17293-PA |
| 7705369 | coatomer protein complex, subunit beta | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 7706523 | WW domain-containing oxidoreductase isoform 1 | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |


| 7706343 | hypothetical protein LOC51647 | 113429091 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 |
| :---: | :---: | :---: | :---: |
| 7705477 | hypothetical protein LOC51504 | 4557719 | DNA ligase I |
| 7706497 | cytidylate kinase | 113429091 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 |
| 8922388 | RNA binding motif protein 28 | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 7657508 | ring-box 1 | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 7705477 | hypothetical protein LOC51504 | 113431146 | PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein) (SIG-20) |
|  |  |  | PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein) |
| 7705477 | hypothetical protein LOC51504 | 113418826 | (SIG-20) |
| 7706523 | WW domain-containing oxidoreductase isoform 1 | 89041736 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 7705477 | hypothetical protein LOC51504 | 113429091 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 |
|  | U6 snRNA-associated Sm-like protein |  | PREDICTED: similar to peptidylprolyl |
| 7706423 | LSm7 | 113429091 | isomerase A isoform 1 |
| 7705748 | TNNI3 interacting kinase | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 7019405 | host cell factor C2 | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 7019405 | host cell factor C2 | 89041736 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 6912280 | AHA1, activator of heat shock 90 kDa protein ATPase homolog 1 | 113429091 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 |
| 7019319 | activator of basal transcription 1 | 113414586 | PREDICTED: similar to CG17293-PA |


| 7657315 | Lsm3 protein | 113414586 | PREDICTED: similar to CG17293-PA |
| :---: | :---: | :---: | :---: |
| 6996005 | dynamin 1-like protein isoform 1 | 10190696 | zinc finger protein 304 |
|  | AHA1, activator of heat shock 90 kDa |  |  |
| 6912280 | protein ATPase homolog 1 | 113414586 | PREDICTED: similar to CG17293-PA |
| 7019319 | activator of basal transcription 1 | 113429091 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 |
|  |  |  | PREDICTED: similar to adaptor-related |
| 5729953 | nuclear distribution gene C homolog | 89042891 | protein complex 1 sigma 2 subunit |
| 5902034 | periodic tryptophan protein 1 | 113414586 | PREDICTED: similar to CG17293-PA |
| 4557719 | DNA ligase I | 113429091 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 |
|  | H2A histone family, member Y isoform |  |  |
| 4758496 | 2 | 4503729 | FK506-binding protein 4 |
|  | small nuclear ribonucleoprotein |  |  |
| 4759156 | polypeptide A | 113414586 | PREDICTED: similar to CG17293-PA |
|  |  |  | PREDICTED: similar to adaptor-related |
| 4759224 | programmed cell death 5 | 89041736 | protein complex 1 sigma 2 subunit |
| 4758384 | FK506 binding protein 5 | 4503729 | FK506-binding protein 4 |
|  | RNA, U3 small nucleolar interacting |  |  |
| 4759276 | protein 2 | 113414586 | PREDICTED: similar to CG17293-PA |
| 4759224 | programmed cell death 5 | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
|  | small nuclear ribonucleoprotein |  | PREDICTED: similar to peptidylprolyl |
| 4759156 | polypeptide A | 113429091 | isomerase A isoform 1 |
| 4758384 | FK506 binding protein 5 | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
|  | H2A histone family, member Y isoform |  |  |
| 4758496 | 2 | 4557719 | DNA ligase I |
| 4557719 | DNA ligase I | 62909985 | hypothetical protein LOC140711 |

H2A histone family, member Y isoform $4758496 \quad 2$

H2A histone family, member Y isoform $4758496 \quad 2$

## 2

4557719 DNA ligase I

H2A histone family, member Y isoform 47584962

4557719 DNA ligase I

RNA, U3 small nucleolar interacting 4759276

4557719 DNA ligase I

H2A histone family, member Y isoform $4758496 \quad 2$

4557719 DNA ligase I

4557719 DNA ligase I

4557719 DNA ligase I

4557719 DNA ligase I

H2A histone family, member Y isoform 4758496

4557719 DNA ligase I

4507311 suppressor of Ty 4 homolog 1

PREDICTED: similar to large subunit
113427044

4506651 ribosomal protein L36a-like protein PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein) (SIG-20)

PREDICTED: similar to peptidylprolyl isomerase A isoform 1 skeletal muscle and kidney enriched inositol

18765707 phosphatase isoform 2

PREDICTED: similar to peptidylprolyl isomerase A isoform 1
protein phosphatase 1, catalytic subunit,
gamma isoform

PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit glucosamine-fructose-6-phosphate

4503981 aminotransferase

4502743 cyclin-dependent kinase 7

N -ethylmaleimide-sensitive factor
4505331 attachment protein, gamma
ubiquitin-conjugating enzyme E2D 4
8393719 (putative)

PREDICTED: similar to large subunit
ribosomal protein L36a

PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit

PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein) (SIG-20)

|  |  |  | PREDICTED: similar to 60S ribosomal |
| :--- | :--- | :--- | :--- |
| 4507311 | suppressor of Ty 4 homolog 1 | protein L26 (Silica-induced gene 20 protein) |  |
|  |  |  |  |


| 4557719 | DNA ligase I |
| :--- | :--- |
| 4506631 | ribosomal protein L30 |
| 4506629 | ribosomal protein L29 |
|  | heat shock 10kDa protein 1 (chaperonin  <br> 4504523 10) <br>  heat shock 10kDa protein 1 (chaperonin <br> 4504523 $10)$ |

4503729 FK506-binding protein 4

4503729 FK506-binding protein 4

4505235 mannose-6- phosphate isomerase

4505773 prohibitin
alpha isoform of regulatory subunit
4506019 B55, protein phosphatase 2

N -ethylmaleimide-sensitive factor
4505331 attachment protein, gamma

4504257 H2B histone family, member A

4504261 H2B histone family, member D

4504269 H2B histone family, member J

89041736

113414586

113428574

113414586

113429091

113431146

PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein)
113418826 (SIG-20)

PREDICTED: similar to peptidylprolyl isomerase A isoform 1

PREDICTED: similar to peptidylprolyl isomerase A isoform 1

PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit

| 4506019 | alpha isoform of regulatory subunit B55, protein phosphatase 2 |
| :---: | :---: |
|  | alpha isoform of regulatory subunit |
| 4506019 | B55, protein phosphatase 2 |
| 4504263 | H2B histone family, member E |
| 4505331 | N -ethylmaleimide-sensitive factor attachment protein, gamma |
|  | heat shock 10 kDa protein 1 (chaperonin |
| 4504523 | 10) |
|  | heat shock 10 kDa protein 1 (chaperonin |

4504523 10)

4505997 protein phosphatase 1D
protein phosphatase 1, catalytic subunit,
4506007 gamma isoform

4502743 cyclin-dependent kinase 7

4502743 cyclin-dependent kinase 7

148222882 hypothetical protein LOC644820

S-phase kinase-associated protein 1A
25777713 isoform b

S-phase kinase-associated protein 1A
25777711 isoform a

21166389 H2B histone family, member L

PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein)

113418826

113431146

89042891

89041736

113431146

113418826

89042891

89042891

113414586

89041736

113429091

113414586

113429091

89042891
(SIG-20)

PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein) (SIG-20)

PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein) (SIG-20)

PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein) (SIG-20)

PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit

PREDICTED: similar to CG17293-PA PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit PREDICTED: similar to peptidylprolyl isomerase A isoform 1

PREDICTED: similar to CG17293-PA

PREDICTED: similar to peptidylprolyl isomerase A isoform 1

PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit

| 4502743 | cyclin-dependent kinase 7 | 113431146 | PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein) (SIG-20) |
| :---: | :---: | :---: | :---: |
|  |  |  | PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein) |
| 4502743 | cyclin-dependent kinase 7 | 113418826 | (SIG-20) |
| 25777713 | S-phase kinase-associated protein 1A isoform b | 113429091 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 |
| 23592238 | glucose transporter 14 | 113414586 | PREDICTED: similar to CG17293-PA |
| 25777711 | S-phase kinase-associated protein 1A isoform a | 113414586 | PREDICTED: similar to CG17293-PA |
| 23592238 | glucose transporter 14 | 113429091 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 |
| 4502859 | CDC28 protein kinase 2 | 113429091 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 |
| 25777713 | S-phase kinase-associated protein 1A isoform b | 25777711 | S-phase kinase-associated protein 1A isoform a |
| 4502743 | cyclin-dependent kinase 7 | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 63029943 | H2A histone family, member B2 | 89041736 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 4758650 | kinesin family member 5C | 113413289 | PREDICTED: similar to Kinesin heavy chain isoform 5C (Kinesin heavy chain neuronspecific 2) |
| 58615669 | cytochrome c oxidase subunit III | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 58615665 | cytochrome c oxidase subunit I | 17981855 | cytochrome c oxidase subunit I |
| 58615673 | NADH dehydrogenase subunit 5 | 17981853 | NADH dehydrogenase subunit 1 |
| 58615673 | NADH dehydrogenase subunit 5 | 58615663 | NADH dehydrogenase subunit 1 |
| 58615673 | NADH dehydrogenase subunit 5 | 17981862 | NADH dehydrogenase subunit 4 |


| 58615673 | NADH dehydrogenase subunit 5 | 58615672 | NADH dehydrogenase subunit 4 |
| :---: | :---: | :---: | :---: |
| 13128862 | histone deacetylase 3 | 113414586 | PREDICTED: similar to CG17293-PA |
| 62909985 | hypothetical protein LOC140711 | 8923475 | thioredoxin-like 4B |
|  | protein phosphatase 1 (formerly 2C)- |  |  |
| 63003905 | like | 113414586 | PREDICTED: similar to CG17293-PA |
|  | dual specificity phosphatase and pro |  | PREDICTED: similar to adaptor-related |
| 51491914 | isomerase domain containing 1 | 89042891 | protein complex 1 sigma 2 subunit |
| 58615672 | NADH dehydrogenase subunit 4 | 17981862 | NADH dehydrogenase subunit 4 |
|  |  |  | PREDICTED: similar to peptidylprolyl |
| 62909985 | hypothetical protein LOC140711 | 113429091 | isomerase A isoform 1 |
| 58615666 | cytochrome c oxidase subunit II | 17981859 | cytochrome c oxidase subunit III |
| 58615669 | cytochrome c oxidase subunit III | 58615666 | cytochrome c oxidase subunit II |
| 58615669 | cytochrome c oxidase subunit III | 17981856 | cytochrome c oxidase subunit II |
| 58615669 | cytochrome c oxidase subunit III | 17981859 | cytochrome c oxidase subunit III |
|  | solute carrier family 2 (facilitated |  |  |
| 4557851 | glucose transporter), member 2 | 113414586 | PREDICTED: similar to CG17293-PA |
| 58615666 | cytochrome c oxidase subunit II | 17981856 | cytochrome c oxidase subunit II |
| 13128862 | histone deacetylase 3 | 113429091 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 |
|  |  |  | PREDICTED: similar to adaptor-related |
| 63029943 | H2A histone family, member B2 | 89042891 | protein complex 1 sigma 2 subunit |
| 58615663 | NADH dehydrogenase subunit 1 | 17981853 | NADH dehydrogenase subunit 1 |
|  | CTD (carboxy-terminal domain, RNA polymerase II, polypeptide A) small |  |  |
| 32813443 | phosphatase 1 isoform 2 | 113414586 | PREDICTED: similar to CG17293-PA |
| 17981859 | cytochrome c oxidase subunit III | 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 32813443 | CTD (carboxy-terminal domain, RNA polymerase II, polypeptide A) small | 113429091 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 |

phosphatase 1 isoform 2

|  | small nuclear ribonucleoprotein |
| :--- | :--- |
| 4507129 | polypeptide E |
| 17981859 | cytochrome c oxidase subunit III |
|  | protein phosphatase 2A, regulatory <br> subunit B' isoform b |
| 30725611 | protein phosphatase 2A, regulatory <br> subunit B' isoform b |
| 450065643 | DNA directed RNA polymerase II |
| 450597 | polypeptide G |
| 4507129 | small nuclear ribonucleoprotein |


| 113418826 | PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein) (SIG-20) |
| :---: | :---: |
|  | PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein) |
| 113431146 | (SIG-20) |
| 89041736 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 89041736 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 113429091 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 |
| 113429091 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 |
| 113431146 | PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein) (SIG-20) |
| 113418826 | PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein) (SIG-20) |
| 113429091 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 |
| 17981856 | cytochrome c oxidase subunit II |
| 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 113429091 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 |
| 113414586 | PREDICTED: similar to CG17293-PA |

polypeptide E

RNA pseudouridylate synthase domain 27734887 31542539

DnaJ (Hsp40) homolog, subfamily A, member 3
ubiquitin-conjugating enzyme E2A
32967280
isoform 1

10190696

8923475

15431295

| 8923475 | thioredoxin-like 4B |
| :--- | :--- |
| 21361394 | adaptor-related protein complex 4, <br> sigma 1 subunit |
|  | general transcription factor IIH, |
| 19923732 | polypeptide 3, 34kDa |
| 19923732 | general transcription factor IIH, <br> polypeptide $3,34 \mathrm{kDa}$ |

8923475 thioredoxin-like 4B

8923475
thioredoxin-like 4B
ribosomal protein L13

AFG3 ATPase family gene 3-like 2

14249470

113429091

32967276

113413881

113414586

113429091

113429091

89041736

113431146

13418826

113418826 (SIG-20)

113431146 (SIG-20)
(SIG-20)

PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein)

PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein)

113414586 PREDICTED: similar to CG17293-PA

113414586 PREDICTED: similar to CG17293-PA
RNA pseudouridylate synthase domain containing 4

PREDICTED: similar to peptidylprolyl
isomerase A isoform 1 PREDICTED: similar to zinc finger protein 114

PREDICTED: similar to CG17293-PA PREDICTED: similar to peptidylprolyl isomerase A isoform 1

PREDICTED: similar to peptidylprolyl isomerase A isoform 1

PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein) (SIG-20)

PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein) (SIG-20)

|  |  |  | PREDICTED: similar to adaptor-related |
| :--- | :--- | :--- | :--- |
| 21396484 | H2B histone family, member H | 89042891 | protein complex 1 sigma 2 subunit |
|  | ubiquitin-conjugating enzyme E2D 4 |  | PREDICTED: similar to peptidylprolyl |
| 8393719 | (putative) | 113429091 | isomerase A isoform 1 |


|  | protein S26 isoform 1 |
| :---: | :---: |
| 113420393 | PREDICTED: similar to 40S ribosomal protein S26 |
| 113420393 | PREDICTED: similar to 40S ribosomal protein S26 |
| 113429091 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 |
| 29742309 | PREDICTED: similar to ribosomal protein L31 |
| 113427093 | PREDICTED: similar to ribosomal protein L31 |
| 4758754 | napsin A preproprotein |
| 113427613 | PREDICTED: similar to ribosomal protein L31 |
| 113422777 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 |
| 113431146 | PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein) (SIG-20) |
| 4504265 | H2B histone family, member G |
| 4504271 | H2B histone family, member K |
| 113420393 | PREDICTED: similar to 40S ribosomal protein S26 |
| 113430282 | PREDICTED: similar to 40S ribosomal protein S26 |
| 113429091 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 |


|  | protein S26 |
| :---: | :---: |
| 113420084 | PREDICTED: similar to 40S ribosomal protein S26 |
| 88982349 | PREDICTED: similar to 40S ribosomal protein S26 |
| 113413585 | PREDICTED: similar to APG4 autophagy 4 homolog B isoform a |
| 113414586 | PREDICTED: similar to CG17293-PA |
| 113414586 | PREDICTED: similar to CG17293-PA |
| 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 113414586 | PREDICTED: similar to CG17293-PA |
| 89042897 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 |
|  | PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein) |
| 113418826 | (SIG-20) |
| 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 88987217 | PREDICTED: similar to 40S ribosomal protein S26 |
| 88987217 | PREDICTED: similar to 40S ribosomal protein S26 |
|  | PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein) |
| 113418826 | (SIG-20) |



| 113429091 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 |
| :---: | :---: |
| 88980535 | PREDICTED: similar to 40S ribosomal protein S26 |
| 88980535 | PREDICTED: similar to 40S ribosomal protein S26 |
| 89041601 | PREDICTED: similar to 40S ribosomal protein S26 isoform 1 |
| 88982349 | PREDICTED: similar to 40S ribosomal protein S26 |
| 113414586 | PREDICTED: similar to CG17293-PA |
| 113414586 | PREDICTED: similar to CG17293-PA |
| 51467029 | PREDICTED: similar to 40S ribosomal protein S26 |
| 89025350 | PREDICTED: similar to 40S ribosomal protein S26 isoform 2 |
| 113418084 | PREDICTED: similar to DNA primase large subunit, 58 kDa |
| 113418086 | PREDICTED: similar to DNA primase large subunit, 58 kDa |
| 113418826 | PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein) (SIG-20) |
| 51467029 | PREDICTED: similar to 40S ribosomal protein S26 |
| 89025350 | PREDICTED: similar to 40S ribosomal protein S26 isoform 2 |


| 89025350 | PREDICTED: similar to 40S ribosomal protein S26 isoform 2 |
| :---: | :---: |
|  | PREDICTED: similar to TBC1 domain family member 3 (Rab GTPaseactivating protein PRC17) (Prostate cancer gene 17 protein) (TRE17 alpha |
| 113426831 | protein) isoform 1 |
| 113431146 | PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein) (SIG-20) |
| 89041736 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 113429091 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 |
| 113427044 | PREDICTED: similar to large subunit ribosomal protein L36a |
| 4506651 | ribosomal protein L36a-like protein |
| 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 4505289 | diphosphomevalonate decarboxylase |
| 113427529 | PREDICTED: similar to large subunit ribosomal protein L36a |
|  | PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 |
| 113431146 | protein) (SIG-20) |
| 113428574 | PREDICTED: similar to 60S ribosomal protein L29 (Cell surface heparin- |


| 88980535 | PREDICTED: similar to 40S ribosomal protein S26 |
| :---: | :---: |
| 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 89041736 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 113418826 | PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein) (SIG-20) |
| 113419590 | PREDICTED: similar to 40S ribosomal protein S28 |
| 113414586 | PREDICTED: similar to CG17293-PA |
| 113414586 | PREDICTED: similar to CG17293-PA |
|  | PREDICTED: similar to ribosomal protein |
| 89042328 | S18 isoform 4 |
|  | PREDICTED: similar to ribosomal protein |
| 41150652 | S18 isoform 1 |
| 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 113418826 | PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein) (SIG-20) |
| 113427529 | PREDICTED: similar to large subunit ribosomal protein L36a |
| 27482992 | PREDICTED: similar to 60S ribosomal protein L29 (Cell surface heparin-binding |


|  | binding protein HIP) |  | protein HIP) |
| :---: | :---: | :---: | :---: |
| 113430282 | PREDICTED: similar to 40S ribosomal protein S26 | 113420393 | PREDICTED: similar to 40S ribosomal protein S26 |
| 113429091 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 | 89035017 | PREDICTED: similar to 40S ribosomal protein S28 isoform 2 |
| 113418086 | PREDICTED: similar to DNA primase large subunit, 58 kDa | 113414586 | PREDICTED: similar to CG17293-PA |
| 113418084 | PREDICTED: similar to DNA primase large subunit, 58 kDa | 113414586 | PREDICTED: similar to CG17293-PA |
| 4506651 | ribosomal protein L36a-like protein | 113418826 | PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein) (SIG-20) |
| 113431146 | PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein) (SIG-20) | 113427044 | PREDICTED: similar to large subunit ribosomal protein L36a |
| 113427044 | PREDICTED: similar to large subunit ribosomal protein L36a | 113418826 | PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein) (SIG-20) |
| 4506651 | ribosomal protein L36a-like protein | 113431146 | PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein) (SIG-20) |
| 113429091 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 | 113427044 | PREDICTED: similar to large subunit ribosomal protein L36a |
| 4506651 | ribosomal protein L36a-like protein | 113429091 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 |
| 113429091 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 | 113418682 | PREDICTED: similar to postmeiotic segregation increased 2-like 2 |
| 89035017 | PREDICTED: similar to 40S ribosomal protein S28 isoform 2 | 113414586 | PREDICTED: similar to CG17293-PA |
| 113429091 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 | 113414586 | PREDICTED: similar to CG17293-PA |


| 88953813 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 |
| :---: | :---: |
| 113418682 | PREDICTED: similar to postmeiotic segregation increased 2-like 2 |
| 113430282 | PREDICTED: similar to 40S ribosomal protein S26 |
| 113430282 | PREDICTED: similar to 40S ribosomal protein S26 |
| 4506651 | ribosomal protein L36a-like protein |
| 113427529 | PREDICTED: similar to large subunit ribosomal protein L36a |
| 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 113429091 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 |
|  | PREDICTED: similar to TBC1 domain family member 3 (Rab GTPaseactivating protein PRC17) (Prostate cancer gene 17 protein) (TRE17 alpha |
| 113426831 | protein) isoform 1 |
| 89041601 | PREDICTED: similar to 40S ribosomal protein S26 isoform 1 |
| 88987217 | PREDICTED: similar to 40S ribosomal protein S26 |
| 1134302 | PREDICTED: similar to 40S ribosomal protein S26 |


| 4758754 | napsin A preproprotein |
| ---: | :--- |
|  | PREDICTED: similar to 40S ribosomal |
| 113429703 | protein S26 |

88943041

PREDICTED: similar to peptidylprolyl isomerase A (cyclophilin A)-like 4

PREDICTED: similar to CG17293-PA PREDICTED: similar to 40S ribosomal protein S26 isoform 1

PREDICTED: similar to 40S ribosomal protein S26

PREDICTED: similar to large subunit ribosomal protein L36a

PREDICTED: similar to large subunit ribosomal protein L36a

PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit PREDICTED: similar to kidney-specific protein (KS)

PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit PREDICTED: similar to 40S ribosomal protein S26

PREDICTED: similar to 40S ribosomal protein S26

PREDICTED: similar to 40S ribosomal protein S26

PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit

PREDICTED: similar to 40S ribosomal protein S26

| 113429703 | PREDICTED: similar to 40S ribosomal protein S26 |
| :---: | :---: |
| 113420084 | PREDICTED: similar to 40S ribosomal protein S26 |
| 89025350 | PREDICTED: similar to 40S ribosomal protein S26 isoform 2 |
| 113429091 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 |
| 113422526 | PREDICTED: similar to 40S ribosomal protein S28 isoform 1 |
| 113422526 | PREDICTED: similar to 40S ribosomal protein S28 isoform 1 |
| 89034184 | PREDICTED: similar to 40S ribosomal protein S28 isoform 2 |
| 113423050 | PREDICTED: similar to 40S ribosomal protein S28 isoform 1 |
| 113423050 | PREDICTED: similar to 40S ribosomal protein S28 isoform 1 |
| 113422526 | PREDICTED: similar to 40S ribosomal protein S28 isoform 1 |
| 89034184 | PREDICTED: similar to 40S ribosomal protein S28 isoform 2 |
| 88959151 | PREDICTED: similar to 40S ribosomal protein S28 |
| 113423050 | PREDICTED: similar to 40S ribosomal protein S28 isoform 1 |
| 113423050 | PREDICTED: similar to 40S ribosomal protein S28 isoform 1 |
| 113420393 | PREDICTED: similar to 40S ribosomal protein S26 |


| 113420084 | PREDICTED: similar to 40S ribosomal protein S26 |
| :---: | :---: |
| 88982349 | PREDICTED: similar to 40S ribosomal protein S26 |
| 88987217 | PREDICTED: similar to 40S ribosomal protein S26 |
|  | PREDICTED: similar to ribosomal protein |
| 113427613 | L31 |
| 88953906 | PREDICTED: similar to 40S ribosomal protein S28 |
| 88959151 | PREDICTED: similar to 40S ribosomal protein S28 |
| 88953906 | PREDICTED: similar to 40S ribosomal protein S28 |
| 89034184 | PREDICTED: similar to 40S ribosomal protein S28 isoform 2 |
| 88959151 | PREDICTED: similar to 40S ribosomal protein S28 |
| 89034184 | PREDICTED: similar to 40S ribosomal protein S28 isoform 2 |
| 88959151 | PREDICTED: similar to 40S ribosomal protein S28 |
| 88953906 | PREDICTED: similar to 40S ribosomal protein S28 |
| 88953906 | PREDICTED: similar to 40S ribosomal protein S28 |
| 113422526 | PREDICTED: similar to 40S ribosomal protein S28 isoform 1 |
| 89025350 | PREDICTED: similar to 40S ribosomal protein S26 isoform 2 |


| 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| :---: | :---: |
| 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 113431146 | PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein) (SIG-20) |
| 113431146 | PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein) (SIG-20) |
| 113427613 | PREDICTED: similar to ribosomal protein L31 |
| 4506651 | ribosomal protein L36a-like protein |
| 113429091 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 |
| 113429703 | PREDICTED: similar to 40S ribosomal protein S26 |
| 113429091 | PREDICTED: similar to peptidylprolyl isomerase A isoform 1 |
| 113429703 | PREDICTED: similar to 40S ribosomal protein S26 |
| 113429703 | PREDICTED: similar to 40S ribosomal protein S26 |


| 113414263 | PREDICTED: similar to aortic preferentially expressed gene 1 |
| :---: | :---: |
| 113418826 | PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein) (SIG-20) |
| 89042891 | PREDICTED: similar to adaptor-related protein complex 1 sigma 2 subunit |
| 113427613 | PREDICTED: similar to ribosomal protein L31 |
| 113418826 | PREDICTED: similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein) (SIG-20) |
| 113427044 | PREDICTED: similar to large subunit ribosomal protein L36a |
| 27482992 | PREDICTED: similar to 60S ribosomal protein L29 (Cell surface heparin-binding protein HIP) |
| 89025350 | PREDICTED: similar to 40S ribosomal protein S26 isoform 2 |
| 113428574 | PREDICTED: similar to 60S ribosomal protein L29 (Cell surface heparin-binding protein HIP) |
| 89041601 | PREDICTED: similar to 40S ribosomal protein S26 isoform 1 |
| 88980535 | PREDICTED: similar to 40S ribosomal protein S26 |

## Appendix D Concatenated Filtered Alignment

Alignment Follows on Subsequent page.

Anopheles_gambiae/1-3220
Arabidopsis_thaliana/1-3286
Ashbya_gossypii/1-3281
A_fumigatus/1-3241
Aspergillus_niger/1-3300 Bombyx_mori/1-2389
Bos_taurus/1-3273
_briggsae/1-3231
Caenorhabditis_elegans/1-3289
Candida_albicans/1-3266 Candida_glabrata/1-3282
Canis_familiaris/1-3286
iona intestinalis/1-2712
Cryptococcus_neoformans/1-3301 Cryptosporidium_hominis/1-3020 C_parvum/1-3281
Danio rerio/1-3286
Debaryomyces hansenii/1-3166
Dictyostelium_discoideum/1-3275 D_melanogaster/1-3287
Drosophila_pseudoobscura/1-3286 Encephalitozoon_cuniculi/1-3199 Entamoeba_histolytica/1-3271 Gallus_gallus/1-3286
Homo_sapiens/1-3286
Kluyveromyces_lactis/1-3281
Leishmania_major/1-3286
Macaca mulatta/1-3182
Magnaporthe_grisea/1-3286
Methanosarcina_acetivorans/1-3043
Monodelphis_domestica/1-3196
Mus musculus/1-3286
Neurospora_crassa/1-3264
Oryza_sativa/1-3218
Ostreococcus_lucimarinus/1-3287
Pan troglodyte/1-3286
Paramecium tetraurelia/1-3282
Pichia_stipitis/1-3290
_falciparum/1-3285
P_knowlesi/1-3284
Plasmodium_yoelii/1-3285
Populus_trichocarpa/1-3286
attus_norvegicus/1-3286
Saccharomyces_cerevisiae/1-3284
Schizosaccharomyces_pombe/1-3284
Strongylocentrotus purpuratus/1-3252
Takifugu_rubripes/1-2930
Tetrahymena_thermophila/1-3270
T_annulata/1-3248
Theileria parva/1-3273
Trichomonas vaginalis/1-3198
T_brucei/1-3264
rypanosoma_cruzi/1-3266
stilago maydis/1-3291
Yarrowia_lipolytica/1-3278

RARMEDLLKRRFFYDQSFAIY-
RKAVVNTLERRLFYIPSFKIY
RENLESVLKRRFFFAPAFELY
RTVLDSMLRRRLFYYPSFDIY
RTLFESLLKRRLFYTESFEIYR
RAKMEDLIKRRFFYDQSFAIY
RAKMEDTLKRRFFYDQAFAIY
RLKLEDLLKRRFFYDQSFAIY
RLKLEDLLKRRFFYDQSFAIY
RDSLEQTLKRRFFFFAPSFEIY
REKLESVLRGRFFYAPAFDLY
RAKMEDTLKRRFFYDQAFAIY
RQQMEDTLKRRFFYGQAFELY
KSTLDALIARRFFFAPSFEIY
----SGVAGLFDYGPPGCAIKSNVLSFWROHFI LEENMLEVDCPCVTPEVVVLKASGVVD ------GGV SGLYDYGPPGCA FQANIVDVWRKHFI LE EDMLEVDCTMLTPY EV LKT SGHVD -----GGV SGLYDYGPPGTALLNNIVDLWRKHFVLEEDMLEVDCTMLTPHEVLKTSGHVD SGNLTCDSRGLYDYGPPGCALQSNIVDLWRKHFVLQEDMLELDCTI LTPEEVFKTSCHVD ------GGI TGQFDFG PMGCALK SNMI HLWKKFFI LQEQMLEVECSI LTP EPVLKASGHVE ------GGV SGLYDFGPVGCALKNNIIQTWRQHFIQEEQILEIDCTMLTPEPVLKTSGHVD ------GGVTG LYDFG PMGCALKANMLQQWRKHFI LEEGMLEVDCT SLTPEPV LKASGHVD ------GGVTGLYDFGPMGCSLKANMLQEWRKHFILEEGMLEVDCT SLTPEPVLKASGHVD ------GGVAGLFDFGPPGCAFQNNVIDAWRKHFI LEEDMLEVEATMLT PHDVLKT SGHVD ------GGVSGLYDYGPPGCSFQANVVDQWRKHFI LE EDMLEVDCTMLTPY EV LKT SGHVD -----GGVGGLYDFGPVGCALKNNIIQTWR QHFIQEEQI LEIDCTMLTPEPVLKTSGHVD ------GGVSGLYDFGPVGCMLKNNIISEWKQHFI LHDQMLEI ECTMLTPEPVLRASGHIE -GGVAGLYDYGPTGSALQANI LDAWRKHYII EEDMLELDTTIMT LSDVLKT SGHVD -MLEI SATCLT PYNPLKASGHVD

## REALENLLKRRFFIAPSFEIY

 TKMEDT LKRRFFYDQAFAI ResLeqV Lkrrfffapafely RAGLEDLMKRRFFITQSFSIY RAKMEDLLKRRFFYDQSFAIY RAKMEDLLKRRFFYDQSFAIY QQQI EQI LKKRFFITQSAYI KAK LDEI LKQRNMVIQSYEI RVKMEDT LKRRFFYDQAFSI RAKMEDT LKRRFFYDQAFAI RESLESVLKRRFFYAPAFELY RTEFEDTCRRRFFYGLAFDP RAKMEDT LKRRFFYDQAFAIY RGALDTI LRRRMFYTPSFEIY Y EKVFELAKRRGFLWNSFELY RVKMEDT LKRRFFYDQAFAI RAKMEDTLKRRFFYDQAFAI KGALESMLRRRMFFAPSFDI RQAVVNTLERKLFYI PSFKIY RAKLGQLLEGRLFYI PSFKIY RAKMEDT LKRRFFYDQAFAI REALEQVIKRRFIYQPAFSLY RESLEQVLKRRFFFAPAFDI RTKLENLVKRKFFYTNSFEIY TKI DNLAKRKLFYTNSFEIY RSKLESLIKRRLFYTNSFEIY RQAVVNTLERRLFFI PSFKIY RAKMEDT LKRRFFYDQAFAIY RDK LESTLRRRFFYTPSFEIY RTQFEELMKKRFFFSPSFQI RAKMEDT LKRRFYYDQSYAI RTKMEDT LKRRFFYDQAFAI RKY FEDLIKRRYFFNQGFEIY KVDCENLLRRRFFYTNSFEIY KVDCENLLRRRFFYANSFEIY RATAEDLEVSGFFWVPSFEIY RSEFEDTCRRRFFFGLAFDPYRAEFEDTCRRRFFFGLAFDPY
SQLEV LMTKRFFYI QSFEI
RETLDAVLKRRFFYAPAFEI
-----GGVAGLFDYGPPGCALKSEVESFWRRHFVLAEDMLEI SATCLTPYNPLKASGHVD ------GGVSGLYDYGPPGCALQANI MDTWRKHFI LE EDMLEVDCTMLTPHEV LKTSGHVD -----GGQAGLYDYGPPGCAVKANLINLWRQHFVLNEDMSEVDCVSVTPEQV LKA SGHVA -----GGI TGQYDFGPMGCALK SNI LALWRQY FALE EQMLEVDC SI LTPEPV LKASCHVE ------GGITGQYDG PMGCALKSNI LSLWRQY FALE EQMLEVDCSI LTPEPV LKASGHVE ------GGVSGLYDLGPPGLSIKTNI LSLWRKHFVLE EDMLEI ETTTMLPHDVLKASGHVD ------GGIAGLYDMGPLGCALKQNI LQFWRKHFTTYENFFEVEGPI LTPKCVLAASGHTA ---GGV SGLYDFGPVGCALKNNI I QAWRQHFIQEEQI LEIDCTMLTPEPVLKTSGHVD ------GGV SGLYDFG PVGCALKNNI QTWRQHFIQEEQI LEIDCTMLTPEPV LKT SGHVD ------GGVSGGYDYGPPGCSFQANI VDVWRKHFVLEEDMLEVDCTMLTPY EVLKT SGHVD ------GGTAGLYDLGPTMCAMK SNMLHFWR QHFVI EESMC EVDTTCLTPEEV FKASGHVT -----GGVSGLYDFGPVGCALKNNII QTWRQHFIQEEQI LEIDCTMLTPEPV LKT SGHVD ------GGVSGLYDYGPPGCALQANIIDAWRKHFVLEDDMLEVDCSVLTPADVLKT SGHVD ------GGSRGFYDYGPLGST LKRRI EQVWREFYVI QEGHMEIECPTIGI EEVFIASGHVG ------GGECE----GPGLGSLAPWAV SSDRSVLRLPQSLAGRRCSLGWPE
.
------GGVSGLYDFG PVGCALKNNI I QAWR QHFI QEEQI LEI DCTMLTPEPVLKT SGHVD -----GGVAGLYDYGPPGCALQANIIDI WRKHFVLEEDMLEVDCTALTPHDVLKTSGHVD -------RGVAGLYDYGPPGCAVKANVLAFWR QHFVLEENMLEVDCPCVTP EVVLKASGHVE ------GGVAGLYDYGPPGCAVK SNVQQFWR QHFVLEESMLEV EC PAVTPEPVLRASGHVE ------GGVSGLYDFG PVGCALKNNIIQTWR QHFIQEEQI LEIDCTMLTPEPVLKTSGHVD ------GGVAGLYDYGPVGCAIKTNI EQY WR EHFII EEDLFEIAATI LTPEPVLKASGHVD ------GGVSGLYDYGPPGCAFQANVVDTWRKHFVLEEDMLEVDCTMLTPHDVLKTSGHVD ------GGASGLFDYGPSGCLLKSELENLWRCHFIYYDEMLEI SGSCVTPYQVLKTSGHVD ------GGSSGLIDYGPSGCLLKSELENLWRYHFI FYDEMLEI SATCITPYTVLKT SGHVD ------GGVSGLIDYGPSGCLLKYELEK LWRNHFV FYDEMLEIKGTCITPY SVLKT SGHVD ------RGVAGLYDYGPPGCAVKSNV LA FWR QHFVLEENMLEVDCPCVTP EVVLKASGHVD ------GGV SGLYDFGPVGCALKNNIIQTWRQHFIQEEQI LEIDCTMLTPEPV LKT SGHVD ------GGV SGLFDLGPPGCQLQNNLIRLWREHFIMEENMLQVDG PMLT PYDVLKT SGHVD ------GGI SGLYDYGPPGSALQSNLVDI WRKHFVI EESMLEVDC SMLTPHEVLKT SGHVD ------GGVSGLYDFGPTGCAMKANFINIWRNHFII EEGMLEVDSAI LAPENV FKASGHVE -----GGVSGLYDFG PMGCALKNNI LQVWRQHFIQEEQI LEIDCTMLTPEPVLKTSGHVD ------GGVAGLYDYGPPGCAI KNNLLK LWR EHFI LEEDMLEI SSTCITPYPVFKASGHVD ------GGSAGLFDFGPPGCALK SELER LWR EHFVVFDEMLEV SCTCITPHPVLKSSGHVD ------GGSAGLFDFGPPGCALKSELERLWR EHFIVFDEMLEV SCSCITPHPVLKSSGHVD ------GSVAGIYDLGPTGCAI ERNFLQKWRDHFVLEDDMLEVRCSALTPR PV LDASGHTE ------GGSAGLYDMG PP LCAMK ANLLAHWR QHFVLAESMC EVDTTCLTPQEVFVTSGHVT ------GGSAGLYDLGPPLCAMKANLLSYWR QHFVLEENMC EVDTT SLTPEEV FKASGHVV ------GGVGG LYDYGPTGAA LQANII NQWRNHFIIEEEMLELDTTIMTLSDV LKT SGHVD ------ - DGV SGLYDYGPPGCALQTRII DTWRDHFVLEDDMLEVDTTMLTPHEVLKT SGHVD

Anopheles_gambiae/1-3220 Arabidopsis_thaliana/1-3286
Ashbya_gossypii/1-3281
A_fumigatus/1-3241
Aspergillus_niger/1-3300 Bombyx_mori/1-2389
Bos_taurus/1-3273
briggsae/1-3231
Caenorhabditis elegans/1-3289
Candida_albicans/1-3266
Candida_glabrata/1-3282
Canis_familiaris/1-3286
iona intestinalis/1-2712
Cryptococcus_neoformans/1-3301 Cryptosporidium_hominis/1-3020 C_parvum/1-3281
anio rerio/1-3286
Debaryomyces hansenii/1-3166
Dictyostelium_discoideum /1-3275 D_melanogaster/1-3287
Drosophila_pseudoobscura/1-3286 Encephalitozoon cuniculi/1-3199 Entamoeba_histolytica/1-3271 Gallus_gallus/1-3286
Homo_sapiens/1-3286
Kluyveromyces_lactis/1-3281 Leishmania_major/1-3286 Macaca_mulatta/1-3182 Magnaporthe_grisea/1-3286 Methanosarcina_acetivorans/1-3043 Monodelphis_domestica/1-3196
Mus musculus/1-3286
Neurospora_crassa/1-3264
Oryza_sativa/1-3218
Ostreococcus_lucimarinus/1-3287
Pan troglodyte/1-3286
Paramecium tetraurelia/1-3282
Pichia_stipitis/1-3290
_falciparum/1-3285
P_knowlesi/1-3284
Plasmodium_yoelii/1-3285
Populus_trichocarpa/1-3286
attus_norvegicus/1-3286
Saccharomyces_cerevisiae/1-3284
Schizosaccharomyces_pombe/1-3284
Strongylocentrotus_purpuratus/1-3252
Takifugu_rubripes/1-2930
Tetrahymena_thermophila/1-3270
_annulata/1-3248
heileria parva/1-3273
Trichomonas vaginalis/1-3198
__brucei/1-3264
rypanosoma_cruzi/1-3266
stilago_maydis/1-3291
Yarrowia_lipolytica/1-3278

R FADLMTKDVKNGECFRLDPITGNDLTEPIEFNLMFGTQIGPLRPETAQGIFVNFKRLLEF-NQGRLPFAAAQIGNSFRNEIS KFTD LMVKDEKTGTCYRADPDTKNPLSDPYPFNLMFQTSIG PMR PETAQGIFVNFKDLYYY-NGKKLPFAAAQIGQAFRNEI S

KFSDWMCQDPK SGEI FRADPVTGETLEPPKAFNLMFETAIGPLRPETAQGQFLNFNKLLEF K FADWMCKDPKTGEI FRADPTTDGNLLPPVAFNLMFQTSIGPLRPETAQGQFLNFQKLLEF K FEDWMCKDFKKGDF LRADPDGDAPVSSPVPFNLMFKTTVGPLRPETAQGQF LNFKKLLDY FFADLMTKDIKTGECFRLDPI SGNDLTPPIEFNLMFNTQIGPLRPETAQGIFVNFKRLLEF K FADFMVKDLKNGECFRADPTTGNDLSPPVPFNLMFKTFIGPLRPETAQGIFLNFKRLLEF R FADWMVKDTKNGECFRADPITGNDLTEPIAFNLMFPTQIGPLRPETAQGI FVNFKRLLEF R FADWMVKDMKNGECFRADPITGNDLTEPIAFNLMFPTQIGPLRPETAQGIFVNFKRLLEF RFSDWMCKDLKTGEI FRADPSTGGKLEPPVEFNLMFDTAIGPLRPETAQGQFLNFNKLLEF KFSDWMCRDLKTGEIFRADPVTGEPLEPPMAFNLMFETAIGPLRPETAQGQFLNFNKLLEF K FADFMVKDVKNGECFRADPNTGNDLSPPVSFNLMFKTFIGPLRPETAQGIFLNFKRLLEF R FADLMVKDEKTGACFRADPVTGNEI SDPMDFNLMFQTTIGPLRPETAQGI FLNFKRLLEF K FADWMVKDVKNGEI YRADPTTGNEVSEPVEFNLMFESNIGPLRPETAQGHFVNFARLLEF RFTDSMI TDIKTNEYYRADP-SGGEWSEPYPFNLMFRTKIGPMRPETAQGI FVNFKRLYEY R FTDSMITDIKTNEYYRADP-SGGEWSEPYPFNLMFRTKIGPMRPETAQGI FVNFKRLYEY KFADYMVKDVKNGECFRADPSTGNDLTPPI SFNLMFQTSIGPLRPETAQGIFLNFKRLLEF KFADWMCRDLKTGEI FRADPATDGPLELPIEFNLMFETAIGPLRPETAQGQFLNFSKLLDC KFADFMVKDEVTKAFFRADPETGNALTEPYPFNLMFQTQIGPLRPETAQGIFTNFGKLYEY R FAD LMVKDVKTGECFRLDPLTGNDLTEPIEFNLMFATQIGPLRPETAQGIFVNFKRLLEF R FADLMVKDVKTGECFRLDPLTGNDLTEPI EFNLMFATQIGPLRPETAQGIFVNFKRLLEF KFCDI LVFDEVSGDCFRADT - LGNK LSK SQQFNLMFGTQI GY LR PETAQGQF LNFKK LC EY KFSDYMVKDLKNGCCYRADPDTGNDLSEPLAFNLMFATDIGPLRPETAQGIFTMFKRNLEF KFADFMVKDMKNGECFRADPITGNDLSPPVSFNLMFKTSIGPLRPETAQGIFLNFKRLLEF KFADFMVKDVKNGECFRADPITGNDLSPPVSFNLMFKTFIGPLRPETAQGIFLNFKRLLEF K F SDWMCKDPKTGEI FRADPVSGDK LEPPRAFNLMFETAIGPLRPETAQGQF LNFNKLLEF R FNDVMVRDTVTGECIRADP-KGNPFSDPFPFNLMFATHIGPMRPELAQGII LNFKR LMDSG KFADFMVKDVKNGECFRADPITGNDLSPPVSFNLMFKTFIGPLRPETAQGIFLNFKRLLEF KFADWMCKDPKTGDI FRADPATGLLPTPPVSFNLMFSTSIGPLRPETAQGQFLNFAKLLEY GFSDPLC ECMNCK EAFRADPECGGEFEDAY EFNLMFKTTIGPLRPETAQGMFVDFQRLSRF NNGKTPFASASIGKSFRNEI S QQ SMPFA SA SI GK SFRNEI NQN SMPFASASIGKSFRNEI S NQGRLPFAAAQI GNSFRNEI NQGKLPFAAAQI GNSFRNEI QQGKLPFAAAOI GLGFRNEI QGGKLPFAAAQIGLGFRNEI S NNDKMPFASASIGKSFRNEIA NNGKTPFASASI GKSFRNEI S NQGKLPFAAAQIGNSFRNEI S QGKLPFSAVOIGMSFRNEI S FRNEI S NGKVPFASAQI GKSFRNEIA NGKKLPFSVAQIGLGFRNEIA NGKKLPFSVAQIGLGFRNEIA NQGKLPFAAAQI GNSFRNEI S NEKMPFASASIGKSFRNEI S NEMMPASASIGKSFRNEI S GKKLPFAAAQI GNAFRNEIA NQGKLPFAVAQI GNSFRNEI S NQGKLPFAVAQI GNSFRNEI S NNDKLPFASASIGKAYRNEI S NGGKVPFGVTQIGNVFRNEIA QGKLPFAAAQI GNSFRNEIS NQGKLPFAAAQI GNSFRNEI S NNGKTPFASASIGKSFRNEI S AQRMPFAGACVGTAFRNEIA NQGKLPFAAAQI GNSFRNEIS NQQMPFASASIGKSYRNEI S YRDKLPFGAVQI GK SYRNEIA K FAD FMVKDVKNGEC FRADPTTGNDLSPPVPFNLMFQTFIGPLRPETAQGI FLNFKRLLEF-NQGKLPFAAAQIGNSFRNEIS KFADWMCKDPKNGDI LRADPATGVQPEPPVAFNLMFQTAIGPMRPETAQGQFLNFAKLLEY-NAGNMPFASASIGKSYRNEIA KFTDLMVKDEKTGTCYRADPDTKNPLSDPYPFNLMFQTSIGPMRPETAQGIFVNFKDLYYY K FTD LMVNDVVTKDCFRADPVTGNDLSEPYPFNLMFPTQIGPLRPETAQGI FVNFRDLLYY K FADFMVKDVKNGECFRADPITGNDLSPPVSFNLMFKTFIGPLRPETAQGI FLNFKRLLEF RFTDLLVCDSKTGTGYRADPETGNDLTDPTPFNLMLPTIIGPLRPETAQGMFLNFARLLEQ KFADWMCKDLKTGEI FRADPATGGKLEPPVEFNLMFETAIGPLRPETAQGQFLNFAKLLEF R FTDLMI RDVVTNDCYRADP - LKNDLSEPFPFNLMFQTKIGPLRPETAQGI FVNFKKLLEY R FTDLMIRDAVTGDFYRADP-GKNDFVGPFPFNLMFQTRIGPLRPETAQGI FVNFKKLLEY R FTDLMIKDIVTKDCYRADP - EKNDLSDPFPFNLMFQTKIGPLRPETAQGI FVNFKKLLEY K FTD LMVKDEKTGTCYRADPDTKNPLSDPYPFNLMFQTSIGPMRPETAQGIFVNFKDLYYY KFADFMVKDVKNGECFRADPTTGNDLSPPVPFNLMFQTFIGPLRPETAQGIFLNFKRLLEF K FTDWMCRNPKTGEYYRADPVTNDVLDALTSFNLMFETKIGALRPETAQGQFLNFNKLLEI K FADWMCKDPATGEI FRADPATNGELETPRQFNLMFETQIGPLRPETAQGQFLNFSRLLEF RFADFMVKDGKTGECFRADPTTNNDLSDPMEFNLMFATAIGPLRPETAQGIFVNFKRLLEF KFADYMVKDVKNGECFRADPTTGNDLTPPISFNLMFQTSIGPLRPETAQGIFLNFKRLLEF R FTD LMVKDVKNGAGHRADPDTGNELGFPEPFNLMFGTPIGPLRPETAQGMFVNFNRLNEF R FTDLMVKNLSNGDCYRADP - EGDEFSKPFPFNLMFSTSIGPLRPETAQGI FVNFNRLLEF R FTDLMVKNLSNGDCYRADP - EGNDFSKPFPFNLMFSTSIGPLRPETAQGI FVNFTRLLEF KFND LMLTDMTTKALYRADP-EGNEFSEPAPFNLMFNTRVGPLRPETAQGI FVNFTRLLNA-NRGSLPFAAAQVGAGYRNEIS R FNDVMVRDTVTGECIRADP-KGNALSDPFPFNLMFSTSIGPMRPELAQGII LNFKRLLDTGNAQRMPFACASIGTAFRNEIA R FNDAMVRDTVTGECIRADP-KGNALSEPFPFNLMFSTSIGPMRPELAQGII LNFKRLLDSGNAQRMPFAGACIGTAFRNEIA K FADWMCKDTKTGEI FRADPESGNEVSEPVEFNLMFESYIGPLRPETAQGHFVNFQRLLEF-NNGRVPFASAQIGKSFRNEIS KFADWMCRDLASGEI FRADPVTGGPLEK PMEFNLMFETAIGPLRPETAQGQF LNFNKLLDC

Anopheles_gambiae/1-3220
Arabidopsis_thaliana/1-3286
Ashbya_gossypii/1-3281
A fumigatus/1-3241
Aspergillus_niger/1-3300 Bombyx_mori/1-2389
Bos_taurus/1-3273
C_briggsae/1-3231
Caenorhabditis elegans/1-3289
Candida_albicans/1-3266
Candida_glabrata/1-3282
Canis_familiaris/1-3286
Ciona intestinalis/1-2712
Cryptococcus_neoformans/1-3301 Cryptosporidium_hominis/1-3020 C_parvum/1-3281
Danio rerio/1-3286
Debaryomyces hansenii/1-3166
Dictyostelium_discoideum/1-3275 D_melanogaster/1-3287
Drosophila_pseudoobscura/1-3286 Encephalitozoon cuniculi/1-3199 Entamoeba_histolytica/1-3271 Gallus_gallus/1-3286
Homo_sapiens/1-3286
Kluyveromyces_lactis/1-3281 Leishmania_major/1-3286 Macaca_mulatta/1-3182 Magnaporthe_grisea/1-3286 Methanosarcina_acetivorans/1-3043 Monodelphis_domestica/1-3196
Mus musculus/1-3286
Neurospora_crassa/1-3264
Oryza_sativa/1-3218
Ostreococcus_lucimarinus/1-3287
Pan troglodyte/1-3286
Paramecium tetraurelia/1-3282
Pichia_stipitis/1-3290
_falciparum/1-3285
P_knowlesi/1-3284
Plasmodium_yoelii/1-3285
Populus_trichocarpa/1-3286
Rattus_norvegicus/1-3286
Saccharomyces_cerevisiae/1-3284
Schizosaccharomyces_pombe/1-3284
Strongylocentrotus_purpuratus/1-3252
Takifugu_rubripes/1-2930
etrahymena_thermophila/1-3270
T_annulata/1-3248
Theileria parva/1-3273
Trichomonas vaginalis/1-3198
T_brucei/1-3264
Trypanosoma_cruzi/1-3266
stilago_maydis/1-3291
Yarrowia_lipolytica/1-3278

PR SGLIRVREFTMAEI EHFVDPS-EKDHPK PRGGLLRCREFQMAEI EY FVDPTEKSTFKK ( PRNG PRNG LRVREFQMAEI EY FVNPK-KKNHEKYY LFKY LMLPLYPRDNQAV--EKNI IANEALAY FLARTYLFLLKCGINKDGIR PRNGLLRVREFEMGEI EYFFNPE-K SKHEKYDLYKHLVLPLYPRTNQAV--NNGIICNEALAYFLARTYLFLLKCGIKKDGIR PRQGLLRVREFTLAEI EHFVDPE-DK SHPKYSEVADLEFLMFPREQQAV--SKGIVNNETLGYFIGRVYLFLTHLGIDKDRLR PR SGLIRVREFTMAEI EHFVDPT - EKDHPKFPSVADLY LY LY SAKAQAV--EQGVINNSVLGYFIGRIYLYLTKVGISPDKLR PR SGLLRVREFLMAEI EHFVDPL-NK SHAK FNEVLNEEI PLLSRRLQAV--NSGMVENETLGYFMARVHQFLLNIGINKDKFR PR SGLLRVREFLMAEVEHFVDPK -NK EHDR PR SGLLRVREFTMCEI EHFIDPT -NKDHPK PR PRNGIRVREFIMAEI EHFVDPK-EKVHQKFANVADLEI LLYSSKAQAV--EQGVINNSVLGYFIGRIYLYLIKVGVAKDKLR PRNGYRVREFDMAEI EHFFDPK - R P EHPK FKYVKDLK LPLLTAK SQAV--K SGTV SNETHAYFIGRTFLFLVEAGVNQNNIR PRNGLRVREFPMAEI EY FVNPK - FKTHEK PRNG LRVREFPMAEI EYFVNPK-FKTHE PRNGLVRCREFQMAEI EHFADPEQLNNFP PRANLIRVREFTLAEI EHFVNPN-DKTHEK PRANLIRVREFTLAEI EHFVNPN-DK SHEK PRAGLLRVREFTMAEI EHFVDPE-DKNHDR

FPEFKNTVLPLLTRDQQAV--SSGIVGNEALAY FLARTFLFLKRVGINEAGLR FPEFSHVVLPLVTRDQQAV--S SGMVGNEALAYFLARTFLFLKRVGINEAGLR FETVKNLKVKLFPASIQAI - -AQHVVSHKTLGYYIGRVYLFLCEIGI QPDTIR FA LVKDVEI WMWARKQQAV--AQK I I DNET LAYFI ARTAQFLEAVGA--RYVR FESVRGTE FWAWSR ELQAV--AKKI I DNET LGYFIARTVLFLEAVGL--RFLR FDEVKHINVPLLAKDVQAV--SAGII DNQT LGYFIGRIYLFLVKIGIDATRLR FDEVKDVKLR FLAKDVQAV - - ETGLVDNKT LGY FLARIYLFLIKIGVNPDR

Anopheles_gambiae/1-3220
Arabidopsis_thaliana/1-3286
Ashbya_gossypii/1-3281
A fumigatus/1-3241
Aspergillus_niger/1-3300
Bombyx_mori/1-2389
Bos_taurus/1-3273
C_briggsae/1-3231
aenorhabditis_elegans/1-3289
Candida_albicans/1-3266
Candida_glabrata/1-3282
Canis_familiaris/1-3286
iona intestinalis/1-2712
Cryptococcus_neoformans/1-3301 Cryptosporidium_hominis/1-3020 C_parvum/1-3281
Danio rerio/1-3286
Debaryomyces hansenii/1-3166
Dictyostelium_discoideum/1-3275 D_melanogaster/1-3287
Drosophila_pseudoobscura/1-3286 Encephalitozoon cuniculi/1-3199 Entamoeba_histolytica/1-3271 Gallus_gallus/1-3286
Homo_sapiens/1-3286
Kluyveromyces_lactis/1-3281
Leishmania_major/1-3286
Macaca mulatta/1-3182
Magnaporthe_grisea/1-3286
Methanosarcina_acetivorans/1-3043
Monodelphis_domestica/1-3196
Mus musculus/1-3286
Neurospora_crassa/1-3264
Oryza_sativa/1-3218
Ostreococcus_lucimarinus/1-3287
Pan troglodyte/1-3286
Paramecium tetraurelia/1-3282
Pichia_stipitis/1-3290
_falciparum/1-3285
P_knowlesi/1-3284
Plasmodium_yoelii/1-3285
Populus_trichocarpa/1-3286
attus_norvegicus/1-3286
Saccharomyces_cerevisiae/1-3284
Schizosaccharomyces_pombe/1-3284
Strongylocentrotus_purpuratus/1-3252
Takifugu_rubripes/1-2930
Tetrahymena_thermophila/1-3270
_annulata/1-3248
Theileria parva/1-3273
Trichomonas vaginalis/1-3198
T_brucei/1-3264
rypanosoma_cruzi/1-3266
Ustilago maydis/1-3291
Yarrowia_lipolytica/1-3278

FR QHMGN EMAHYACDCWDA ECL FRQH LAN EMAHYAADCWDAEIE FRQHMSN EMAHYATDCWDA E LK FRQHMAN EMAHYAADCWDAELL FR QHMAN EMAHYATDCWDA ELQ FR QHMGN EMAHYACDCWDA ECL FRQHMEN EMAHYACDCWDAESK FR QH LSN EMAHYAQDCWDAEI L FRQH LSNEMAHYAQDCWDAEI L FR QHMSN EMAHYA SDCWDA ELE FR QHMAN EMAHYAADCWDA E LK FR QHMEN EMAHYACDCWDA ESK FR QHMQN EMAHYACDCWDA ECK CR QHMAN EMAHYATDCWDFEIQ FRQHLST EMAHYA SDCWDA EV L FRQH LST EMAHYA SDCWDA EV L FR QHMDN EMAHY ACDCWDA ETK FRQHMSN EMAHYAADCWDA E LH FRQHQK N EMAHYAQDCWDAEI L FR QHMSN EMAHYACDCWDAEI LT FRQHMSN EMAHYACDCWDAEI L FR QHKK D EMA HYAKGCWDA EI YT FRQHLKTEMAHYAKDCWDAEIRL FRQHMEN EMAHYACDCWDA ESKT FRQHMEN EMAHYACDCWDA ESK FR QHMAN EMAHYAADCWDA ELQ FRQHQSTEMAHYAQDCWDAELL FR QHMEN EMAHYACDCWDA ESK FR QHMAN EMAHYACDCWDAELL FRQH LTDEMAHYAI DCWDAEI ET FR QHMDN EMAHYACDCWDA EAR FRQHMEN EMAHYACDCWDA ESK FR QHMGN EMA HYACDCWDA ELLT FRQH LPNEMAHYAADCWDAEI EC FR QH LKHEMAHYAADCWDAEI QC FR QHMEN EMAHYACDCWDA ESKT FRQHQAD EMAHY S SDCWDA EI EM FRQHMGN EMA HY A SDCWDA E LQT FRQHLKTEMAHYANDCWDAEI LT FRQH LPT EMAHYANDCWDAEI L YRQHLEKEMAHYANDCWDAEI LT FRQH LAN EMAHYAADCWDAEI ES FR QHMEN EMAHYACDCWDA ESK FRQH LKNEMAHYATDCWDGEI LT FR QHMSN EMAHYACDCWDAEI QC FRQHMFN EMAHYATDCWDA ETKT FR QHMDN EMAHYACDCWDA ETKT FRQHMSN EMAHYACDCWDAEI EF FR QHMAN EMAHYA SDCWDA E I LT FRQHTANEMAHYA SDCWDA EI LT FR MHR KN EMA HYAR ECWDA EI YT FR QH LR N EMA HY AQDCWDA E L LT FRQHQR DEMAHY AQDCWDA ELLT FR QHMSN EMAHYA SDCWDA EI HT FR OHMSN EMAHY ATDCWDA ELHT

SYGWI ECVGCADR SAYDLTQH'TNAT--...--GVKLVAEKK L'PAPKÁAIGKÁFKKEÁKA SYGWI ECVGI ADR SAYDLRAHSDK S--. ----GTPLVAE EK FAEPKK ELGLAFKGNQKN SYGWI ECVGCADR SAYDLTVHKNKT - . - - - - -GAPLVVREPRAEPKKKFGPRFKKDGKA TYGWI ECVGCADR SAYDLTVHSRKT------KEPLVVREPRREPKPKLGPLFKKNAKA SYGWI ECVGCADR SAYDLTQHTKAT-------GIR LAAEKKLPAPKAAI GKAFKKDSQA SYGWI EIVGCADR SCYDLSCHARAT-------KVPLVAEKPLKEPKGAIGKAYKKDAKL YGWI ECVGNADRACYDLQQHYKAT-------NVK LVAEKKLPEPMALLGKKYKK EAKK YGWI ECVGNADRACYDLQQHYKA YGWI ECVGCADR SAYDLSVHSAR FGWI ECVGCADR SAYDLTVHANK YGWI EI VGCADR SCYDLSCHARA
YGWV ECVGCADR SCYD LKCH SQAA YGWI ECVGCADR SAYDLTVHSVR YGWI ECAGHADR SCYDLLQHSKA GWI ECAGHADR SCYDLLQHSKA YGWI EI VGCADR SCYDLLCHARA YGWI ECVGCADR SAYDLSVHSAR YGWV ECVGHADR SCYDLKVHAT E YGWV ECVGCADR SAYD LGQHTAA YGWV ECVGCADR SAYD LGQHTAA YGWI ECVGI ADRACYDLSCHEDG YGWV ECVGHADRGDFDLSNHARC YGWI EI VGCADR SCYDLSCHARA YGWI EIVGCADR SCYDLSCHARA YGWI ECVGCADR SAYDLTVHSNK YGWI ECVGIADR SAYDLTQHSNA YGWI EIVGCADR SCYDLSCHARA SYGWI ECVGCADR SAYDLSVHAKK DRFGWVEIVGIADRTDYDLKAHARV SYGWI EI VGCADR SCYDLTCHSRA YGWI EIVGCADR SCYDLSCHARA SGWV ECVGCADR SAYDLTVHAKK YGWI ECVGI ADR SAYDLRAHSDK YGWI ECVG LADR SAFDLKAHSDK YGWI EIVGCADR SCYDLSCHARA SGWV ECVGLADR SAYDLNAHSEA YGWI ECVGCADR SAYDLSVHAAR SFGFI EVVGHADR SAYD LQHHMKY YYGFI EVVGHADR SAYD LKNHMKV SYGY I ECVGHADR SAYDLKHHMNA YGWI ECVGIADR SAYDLRAHTDK YGWI EIVGCADR SCYDLSCHARA YGWI ECVGCADRAAFDLTVHSKK YGWI ECVGCADR SAYDLSVHSKA YGWV ECVGNADR SCYDLTCHAKH YGWI EIVGCADR SCYDLACHARV HG FK ECVG LANR SAFDLESHTKG YGWV EVVGHADR MAYD LMCHSK S YGWV EVVGHADR MAYD LMCHSK S LGWL ECVGIADRQSWD LSRHAKY YGWV ECVGVADR SAYDLTQHSGA YGWV ECVGI ADR SAYDLTQH SAA YGWI ECVGCADR SAYDLTVHSKR YGWI ECVGCADR SAYDLSVHEAR
------NVKLVAEKKLPEPMALLGK SFKKDAKK ---GEKLVARQT LAEPKKKFG PKFRKDAGT ------K EK LVVR QK LeTPKK LFG PK FRKDAPK --KVPLVAEKPLKEPKGAI GKAYKKDAKL ---KVNLSAERPLPEPKQAVGKAFKKDAKK ---K QPLRV VQR LDQPAKA FGMK FKKDATM ------KTDLFASEKYDEPKPLIGKT FKQEASL ------KTDLFASEKYDEPKPLIGKTFKQEASL --KVPLVAEKPLKEPKGAIGKAYKKDAKF ------NEK LVVR QPLPEPKKK FG PK FRKDAGT ------KSNLSAYEEFKEPPGAI SKKHRAAVSP ---GVRLVAEKR LPAPKQALGKTFKKEAKN ---GVR LVAEKR LPAPKQALGKT FKKEAKT ---kVDLRCKRR LAEPKK EWGAK LRDRFSV ---KVDQSVFIAYDEPKGVMGKKYKKDSQK ---KVPLIAEKLLKEPKGAIGKAYKKDAKV ----KVPLVAEKPLKEPKGAIGKAYKKDAKL ---K EK LVVREALETPKKLFGPK FRKDAPK ------KKDLCAREEYDEPKGLIGKTFGKKAGE ------KVPLVAEKPLKEPKGAIGKAYKKDAKL ------NAPLIVR QR LPEPKKK FG PK FKKDAKA -------KTDLYVYV EYDEPMGK LGPLFKGKAKA ------KV PLVAEK LLREPKAAI GRTYKKDAR ------KVPLVAEKPLKEPKGAVGKAYKKDAKL ------GAPLVVRETLETPSKKFGPT FRKDAKT ---gVPLVAHEKFSK PKKDLGLAFKGNQKM --KVDLVAY ER FDK PKKV LGKA FKKDAK P --KVPLVAEKPLKEPKGAIGKAYKKDAKL --GQK LQAARK FKV PK QK I GK ELKKDGMA --NA SLVVR QPLPEPKKK FG PK FKKDGGA --GANLYAC EKYNEPKAKIGHTFKSEQNK -GAN LYAC EK YDT PKAK MGMK FK SQQNV -GSNLYGCOKYDK PKAKIGMK FK SDQNK ---gV PLVAHEKFSEPKKELGLSFKGNQKK ---KVPLVAEKPLKEPKGAVGKAYKKDAKL ---Gr SLTVK QK LDT PKK FFGSK FK QKAKL --KTPLVVQEALPEPRKKFG PR FKRDAKA ------KVAMVAEKKLPEPKSVMGKAFKKEAKV ------KVPLVAEKPLKEPKGALGKAYKKDAKI
------GVKLLAARR LPEPKKEI FKA LKGDGNE ------NSQLVAHHRYDNPKPEIGKAFKSDQK --NSQLVAHHRYDTPKPEIGKAFKADQKI KKGdaEs Sply lsapldt PK SAIGKI frkdak ------KKDLCAR EEFAEPKGAI GKAFGKNAGD ------KKDLFAREEFAEPKGAIGKAFGKNAGE ------KKDLVVQKAHKEPKKNLG PK FKKDAKF ------KKDLVVQKAHKEPKKNLG PK FKKDAK

Anopheles_gambiae/1-3220 Arabidopsis_thaliana/1-3286
Ashbya_gossypii/1-3281
A fumigatus/1-3241
Aspergillus_niger/1-3300 Bombyx_mori/1-2389
Bos_taurus/1-3273
C_briggsae/1-3231
Caenorhabditis_elegans/1-3289
Candida_albicans/1-3266 Candida_glabrata/1-3282
Canis_familiaris/1-3286
Ciona intestinalis/1-2712
Cryptococcus_neoformans/1-3301 Cryptosporidium_hominis/1-3020 C_parvum/1-3281
Danio rerio/1-3286
Debaryomyces hansenii/1-3166
Dictyostelium_discoideum/1-3275 D_melanogaster/1-3287
Drosophila_pseudoobscura/1-3286 Encephalitozoon cuniculi/1-3199 Entamoeba_histolytica/1-3271 Gallus_gallus/1-3286
Homo_sapiens/1-3286
Kluyveromyces_lactis/1-3281 Leishmania_major/1-3286 Macaca_mulatta/1-3182 Magnaporthe_grisea/1-3286 Methanosarcina_acetivorans/1-3043 Monodelphis_domestica/1-3196
Mus musculus/1-3286
Neurospora_crassa/1-3264
Oryza_sativa/1-3218
Ostreococcus_lucimarinus/1-3287
Pan_troglodyte/1-3286
Paramecium tetraurelia/1-3282
Pichia_stipitis/1-3290
_falciparum/1-3285
P_knowlesi/1-3284
Plasmodium_yoelii/1-3285
Populus_trichocarpa/1-3286 Rattus_norvegicus/1-3286 Saccharomyces_cerevisiae/1-3284 Schizosaccharomyces_pombe/1-3284 Strongylocentrotus purpuratus/1-3252 Takifugu_rubripes/1-2930 etrahymena_thermophila/1-3270 T_annulata/1-3248
Theileria parva/1-3273 Trichomonas vaginalis/1-3198 T_brucei/1-3264
rypanosoma_cruzi/1-3266
stilago_maydis/1-3291
Yarrowia_lipolytica/1-3278

ITEALPSVIEPSFGIGRIMYSLLEHSFRMRRSYFSLPPVVAPLKCSVLPLSNNAEFAPFVKKISSALTSVDVSHVVDSSGSI VVESLPSVIEPSFGIGRIIYCLYEHCFSTRLNLFRFPPLVAPIKCTVFPLVQNQQFEEVAKVISKELASVGISHKIDITGTSI VENYLPNVIEPSFGIGRIIYAIFEHSFWSRRSVLSFPPLVAPTKVLLVPLSNNADLAEVVTEVSRVLRKEQI PFKVDDSGVSI VAAAVPNVIEPSFGIGRILYSMI EHVYWARRGVLSFPPAIAPTKVLIVPLSTHASFRPLLQQLMTKLRRMGI SNRVDDSSASI VEAALPNVIEPSFGFGRI FY SLLEHVYWHRRGVLSLPI SVAPTKVLIVPLSTHQDFVPITKRITEDLRELGI SCRADESSASI INDTL--------------------------------APMKCVVLPLSGNAEFQPFVRDLSQELITVDVSHKVDDSSGSI VMEYLPSVIEPSFGLGRIMYTVFEHTFQVRRTFFSFPAVVAPFKCSVLPLSQNQEFMPFVKELSEALTRHGI SHKVDDSSGSI IQTALPSVIEPSYGIGRIMYALLEHSFRQRRTFLAFKPLVAPIKCSILPI SANDTLVPVMDAVKEELSHYELSYKVDDSSGTI I QTSLPSVIEPSYGIGRIMYALLEHSFRQRRTFLAFKPLVAPIKCSVLPI SANDTLI PVMDAVKEELSRFEMSYKVDDSSGTI VEKWLPNVIEPSFGIGRILYSIFEHQFWCRRGVLSLPPIVAPTKVLLVPLSNNSELQPIVKKVSQALRKEKIPFKVDDSSASI VEAYLPNVIEPSFGIGRIIYSIFEHSFWSRRAVLSFPPLVAPTKVLLVPLSNHKDLAPVTAQVSKILRKEQIAFRVDDSGVSI VLEY LPNVIEPSFGLGRIMYTVFEHTFHVRRTFFSFPAVVAPFKCSVLPLSQNQEFMPFVKELSEALTRNGVSHKVDDSSGSI VQAHLPSVIEPSFGFGRLLYSTLEHNFKIRRTFFSLPAVIAPYKCSLLPLSNKPDFNPFITSLSLALKKLGISHKVDTSSGSI IKETLPNVIEPSFGIGRILYCVLEHTYWARRGVLSLPALVAPIKCLIVSI SQDAQLRSKIHEVSREMRKRGIASRVDDSSATI VTEALPGVIEPSFGIGRIIYCLLEHSFKIRRSYLSLPALIAPVKCSILPISSNAIFNDLINLLHKSFINHGISCKVDTSSASI VTEALPGVIEPSFGIGRIIYCLLEHSFKIRRSYLSLPALIAPVKCSILPISSNAI FNDLINLLHKSFINHGISCKVDTSSASI AMEYLPNVIEPSFGIGRIMYSIFEHTFRIRRTYFSFPATVAPYKCSVLPLSQNQEFMPFVRELSEALTRNGVSHKVDDSSGSI VENWLPNVIEPSFGIGRILYSIFEHQFWARRTVLSLPPLVAPTKVLLVPLSSNAELQPIVKKI SAFLRKEQVPFKVDDSSASI IKKYLPHVIEPSFGLGRIIYSILEQNYYTRRGVLSLPAIIAPVKASILPLTSSDRIAPFVQTISKALKEANISTKVDDTGNAI ITDALPSVVEPSFGIGRIMYSLLEHSFQCRRCYFTLPPLVAPIKCSI LPLSNNTDFQPYTQKLSSALTKAELSHKVDDSSGSI ITEALPSVVEPSFGIGRIMYALLEHSFQCRRCYFTLPPLVAPLKCSI LPLSNNAEFQPYTQKLSSSLTKAELSHKVDDSSGSI LMETVPDVIEPSFGIGRILYALIEHSFY LRRPVFRFKPAIAPVQCAIGYLIHFDEFNEHILNIKRFLTDNGLVVHVNERSCSI LFAYAPNVIEPSFGVGRVLTAVLEHSFWVRKSVLSIPASIAPVKVGLFPLLTKLEFNNKIAEIEKICKNGFLSFKSNTTAVAI VMEYLPNVIEPSFGIGRIMYTVFEHTFRIRRTFFSFPAVVAPFKCSVLPLSQNQEFMPFVKELSEALTRNGISHKVDSSGSI VMEYLPNVIEPSFGLGRIMYTVFEHTFHVRRTFFSFPAVVAPFKCSVLPLSQNQEFMPFVKELSEALTRHGVSHKVDDSSGSI VEARLPNVIEPSFGIGRIIYSIFEHSFWSRRAVLSFPPLVAPTKVLLVPLLNNPELSKITAQVSQI LRKEQIPFKVDESGVSI VMAYLPSVIEPSFGIGRILYCLLEQSYWURRAVFSFSPLLAPQKVALLPLMVKPELLATISEIRQELVMRGISVRVDDSSVTI
 VETVLPNVIEPSFGIGRILYCLLEHNYWTRRGVLSFTPVVAPTKVLIVPLSRHDDFVPFVQKISQKLRSVGVSSRVDDSSATI VADALPHVIEPSYGIDRIFYGIMEHAFDEERLVMHFSSAVAPVQVAVLPLLTRKELADPAKEIIAKLREKTLLVNYDDSG-TI VLDY LPNVIEPSFGLGRIMYTVFEHTFRVRRTFFSFPAIVAPYKCSVLPLSQNQEFAPFVRELSEALTRNGVSHKVDDSSGSI VLEYLPSVIEPSFGLGRIMYTILEHTFHVRRTFFSFPAVVAPFKCSVLPLSQNQEFMPFVKELSEALTRNGVSHKVDDSSGSI VEAAIPNVIEPSFGIGRI LYSLLEHNYVRRGVISFPPAVAPVKVLIVPISSKAEFAPHVRRLSQKLRSVGI SSRVDDSSASI VVEALPSVIEPSFGIGRIIYCLFEHSFYTRLNVFRFPPIVAPIKCTVFPLVKNQEFDDAAKVIDKALTTAGISHIIDTTAISI VTDALPSVIEPSFGIGRIMYCMFEHAFYIRKTVLRLTPVVAPIKTTIFPLVNDDKLNAIAAEMNKMLTTNGISAKLDATAI SV VMEY LPNVIEPSFGLGRIMYTVFEHTFHVRRTFFSFPAVVAPFKC SVLPLSQNQEFMPFVKELSEALTRHGVSHKVDDSSGSI LIKYVPHVIEPAFGIGRILQAII EHSFNQRKTFFKFSPRVAPVKCSI LSVVQSEEFDNVIFELTSSLKKLGI SCKTDNAGVAL I QKWLPNVIEPSFGIGRILYSIFEHQYWARRGVLSLPPLVAPTKVLLVPLSNSADLQPIVTKVSAYLRKQQI PFKVDDSSASI IYACLPNVIEPSFGIGRLIFCILEHSFRIRRQYLSLPYKLAPIKCSILSI SNNKAFYPYIKQI QMLLNQYNISSKIDNSSVSI IYQWLPNVIEPSFGIGRLIFCIIEHSFRTRRHYLSLPYTLAPIKCSVLTI SNHKTFIPFVKQVQMI LNEFSISSKIDNSSVSI IYKILPNVIEPSFGIGRLIFCILEHSFRVRRHYLSLPYALSPIKCSVLSI SNNKEFYPYIKQIQTI LSENNI SCKLDNSSVSI VVEALPSVIEPSFGIGRIIYCLYEHSFYMRQNVFRFPPLVAPIKCTVFPLVQNQQYEDVAKIISKSLTAAGISHKIDITGTSI VLEYLPSVIEPSFGLGRIMYTI LEHTFHVRRTFFSFPAVVAPFKCSVLPLSQNQEFMPFVKELSEALTRNGVSHKVDDSSGSI IESVLPNVIEPSFGLGRIIYCIFDHCFQVRRGFFSFPLQIAPIKVFVTTI SNNDGFPAILKRI SQALRKREIYFKIDDSNTSI VEEAMPNVIEPSFGLGRILYVLMEHAYWTRRGVLSFPASIAPIKALIVPLSRNAEFAPFVKKLSAKLRNLGI SNKIDDSNANI VVEHLPSVIEPSFGVGRILYSI LEHNFKVRRTYFTLPPIIAPYKCCVLPLSSNKDFEPLVKTLAQALSNASI SHKVDSSSGSI AMDFLPNVIEPSFGIGRIMYTIFEHTFHVRRTFFSFPATVAPYKCSVLPLSQNQEFVPFVRELSEELTRNGVSHKVDSSGSI LTKLIPYVIEPSFGVGRIFSAI LEHSFRMRRTFFHLPPKISPIKCSILPVISHEKYNDAIHKLKVGLTKVGVSSKVDDTGHAI LLNHLPCVIEPSFGLGRLIFSI LEHSYRVRRKYVALNKSIAPTKCSVLPLSSKEVFEPLITRVQAHLRRLGISHKVDKTGASI I LNHLPCVIEPSFGLGRLIFSI LEHAYRVRRKYVSLHKSIAPTKCSI LPLSSKEVEEPLISRVQSQLRSLGISHKVDKTGASI I MDALNI TVCGDKTVTY EMYNI NDTVVTTR ----FFPNALSS---------------------------------------------VMEYLPSVIEPSFGVGRI LYALLEQSYWVRRAVFGLRPLIAPQKVAVFPLLMKPELIRTVEEIKERMLLHGISTRTDDSGASI VMEYLPSVIEPSFGVGRI LYALLEQSYWVRRAVFSMRPVIAPQKVAVLPLLVKPELRVVEEIRGDMVLRGISTRTDDSGASI VEEAIPNVIEPSFGIGRIFYSLLEHSFWTRRGVLSLPPLVAPIKASIVPISSNEKLSPLVKQVSRKLRSAGVASRVDDSNASI VEEWFPNVIEPSFGIGRILYSLIEHCFWTRKGVLSFPPRIAPTKVLVVPLSSQKELAPFTQEVSKKLRQARISAKVDDSSASI

Anopheles_gambiae/1-3220
Arabidopsis_thaliana/1-3286
Ashbya_gossypii/1-3281
A fumigatus/1-3241
Aspergillus_niger/1-3300 Bombyx_mori/1-2389
Bos_taurus/1-3273
C_briggsae/1-3231
Caenorhabditis_elegans/1-3289
Candida_albicans/1-3266 Candida_glabrata/1-3282
Canis_familiaris/1-3286
Ciona intestinalis/1-2712
Cryptococcus_neoformans/1-3301 Cryptosporidium_hominis/1-3020 C_parvum/1-3281
Danio rerio/1-3286
Debaryomyces hansenii/1-3166
Dictyostelium_discoideum /1-3275 D_melanogaster/1-3287 Drosophila_pseudoobscura/1-3286 Encephalitozoon_cuniculi/1-3199 Entamoeba_histolytica/1-3271 Gallus_gallus/1-3286
Homo_sapiens/1-3286
Kluyveromyces_lactis/1-3281
Leishmania_major/1-3286
Macaca mulatta/1-3182
Magnaporthe_grisea/1-3286
Methanosarcina_acetivorans/1-3043
Monodelphis_domestica/1-3196
Mus musculus/1-3286
Neurospora_crassa/1-3264
Oryza_sativa/1-3218
Ostreococcus_lucimarinus/1-3287
Pan troglodyte/1-3286
Paramecium tetraurelia/1-3282
Pichia_stipitis/1-3290
_falciparum/1-3285
P_knowlesi/1-3284
Plasmodium_yoelii/1-3285
Populus_trichocarpa/1-3286
Rattus_norvegicus/1-3286
Saccharomyces_cerevisiae/1-3284
Schizosaccharomyces_pombe/1-3284
Strongylocentrotus_purpuratus/1-3252
Takifugu_rubripes/1-2930
Tetrahymena_thermophila/1-3270
T_annulata/1-3248
Theileria parva/1-3273
Trichomonas_vaginalis/1-3198
T_brucei/1-3264
rypanosoma_cruzi/1-3266
stilago_maydis/1-3291
Yarrowia_lipolytica/1-3278

420430
430440 GKRYARTDELGVPFAITVDS------DTSVTIRERDSKDQVRVT LKEAASVVSSVSEG--KMTWKFEP--AA -PPARVGRKQG GKRYARNDELGTPFGITIDFESIK--DGSVTLRERDSTRQVRGSVTDIIRAIRDITYN--GVTWKYEP--P--VQSKIGRKKG GKRYARNDELGTPFGITVDFQSVK - -DNTFTLRDRDTTKQVRASEDEI LQAIKSLVDG--EKTWKYEP--PPPPTTRIGRKKG GKRYARNDELGI PLGITIDFDSVK - -DGTITLRERDSTKQVRASKEEI LGAIESLI SG--KMNWRYEP--PPPPTTRLGRKKG GRRYARTDELGVPYAVTVDFDTIK - EPHTVTLRERDSMRQVRLPMADVPTVVRDLSNS - -KI LWGRRYARTDEI GVAFGITI DFDTVNKTPHTATLRDRDSMRQIRAEVSELPSVVRDLANG--SIMWKYEP--P--VPTRVGKKKG GRRYARTDEIGI PFGITVDFESGKTTPYTVTIRHAETMSQIRLEVSELGRLISDLVSG--RQQWKYEA--P--IPSRIGKKKG GRRYARTDEIGI PFGITVDFDSLKTTPFTVTIRHAETMSQIRLEVSELGRLISDLVAG--RQQWKYEA--P--IPSRIGKKKG GKRYARNDELGTPFGITIDFDSVK - -DDSVTLRERDSTKQVRGSI QEIVEAIKDITYN--DGTWKYEP--P--VESKFGKKKG GKRYARNDELGTPFGITIDFDSVK - -DGSVTLRERDSTKQVRGSVEAVIKAVREITYN--GASWKYEP--P--VQSKFGRKKC GRRYARTDEIGVAFGITIDFDTVNKTPHTATLRDRDSMRQIRAEVSELPSVVCDLANG--SITWKYEP--P--VPTRVGKKKC GKRYARTDEI AVPFGITVDFDTVKI EPHTATLRERDSLVQIRATVEEI PQIVYDLVQE - -NTTWKYER - - P--LPTRVGKRRG GKKYARNDELGTPFGCTVDFATIQ--NGTMTLRERDSTSQLIGPIEDVI SVVDQLVKG--VLDWKWEP--P--VPTRIGKKKG GRRYARTDEIGI PFGITIDFQSVK - -DDTVTLRERDSMKQVRISSSEVPSVI SKIINQ--QITWKLES--A--PPPMEMKRKC GRRYARTDEIGI PFGITIDFQSVK - -DDTVTLRERDSMKQVRISSSEVPSVISKIINQ--QITWKLES--A--PPPMEMKRKG GRRYARTDEIGVAFGITIDFDTVNKNPHTATLRDRDSMRQIRAEVGELPEIIRDLANG--AITWKYEP--P--IPTRVGKRKG GKRYARNDELGTPFGITIDFDSVK - -DESVTLRDRDSTKQVRGSLEDIVEAIKDIAYN--NVSWKYEP--P--VESKFGKKRG GRKYARTDEI GVPFGVTI DFQTVE--DNTVTLRERDTTKQVRI PI SELASTLRKLCDL--TVSWKYQP--PP-PPTQFGKKKG GRRYARTDEIAI PYGITVDFDTLK - EPHTVTLRDRNTMKQVRVGLEEVVGVVKDLSTA - -RTTWKYEP--P--I PTRVGKKKC GRRYARTDEIAI PYGITVDFDTLK - EPHTVTLRDRNTMKQVRVGLEQVVGVVKDLATA - -RTSWKYEP--P--I PTRVGKKKG GRKY S SCDELGI PFFIT FDPDFLK - -DRMVTIRERDSMQQIRVDVEKCPSIVLEYIRG--QSRWNLQD--T--TTINLRRRRE
 GRRYARTDEVGVAFGITIDFDTVNRTPHTATLRDRDSMRQIRAEVSELPAIIRDLANG--YLTWKYEP--P--VPTRVGKKKG GRRYARTDEI GVAFGVTI DFDTVNKT PHTATLRDRDSMRQIRAEISELPSIVQDLANG--NITWKYEP--P--VPTRVGKKKG GKRYARNDELGTPFGVTIDFDSVT--DGSIT LRERDSTKQVRGSVADVIKAIREITYQ--GVSWKYEP--P--VESKFGRKKG GKKYARVDELGI PFAITCDFEG----DGSVTLRERDTA SQVRVPKLEVASVVVDLCNPLQPLTWKWEP--P--VAPEIGKRKG GRRY SRNDELGTPLGITVDFQTVK - -DGTITLRDRDTTVQVRADQDKIVEAIQELVSG--NKVWKYEP--PPRPTTRVGRKK GRRYRRNDEIGTPYSVTVDYDTLQ - -DGTVTIRDRDSMRQVRAPINGI ENVLYELIYR - -GRDF - _ GKRYARTDEI GVAFGITI DFDTVNKT PHTAT LRDRDSMRQIRAEISELPKVVCSLANG--TMTWKYEP--P--VPTRVGKKKG GRRYARTDEI GVAFGITI DFDTVNKT PHTAT LRDRDSMRQIRAEVSELPNVVRDLANG--NITWKYEP--P--VPTRVGKKKG GRRYARNDELGTPFGLTIDFQTLQ--DGTFTLRERDSTRQVRAEEEKIVDAIKALVEG--SKTWKYEP--PPKPTTRIGRKKC GRRYARTDEI GVPFAVTVDS------ATSVTIRERDSKEQIRVGIDEVASVVKQLTDG--QSTWKFEP--PA -APSRVGRKQC GKRYARTDELGVPFAVTVDHR SVT - -ENTVTVRERDSCGQVRVPI PEVPGLLGR LCKM--TVDWKYVP--PA-PPMRVGKKKG GRRYARTDEI GVAFGVTI DFDTVNKT PHTATLRDRDSMRQIRAEI SELPSIVRDLANG--NITWKYEP--P--VPTRVGKKKG GKKYARTDEIGI PFAITVDKETLT--AQSVTLREI ETTKQVRVPIAEVPRLI LELSAG--LI LWKKPP--P--PPQRVGRKKG GKRYARNDELGTPFGITIDFDSIK--DESVTLRERDSTKQVRGSFEDVVAAIKEITYT--GTTWKYEP--P--VESKFGKKRG GKKYARTDEIGI PFAVTIDFQTLK - -DKTITLRERDSMLQIRISMSHLVDI INSMLHA - -KKNWKLES--V--PISHMGKKKG GKKYARTDEIGI PFAVTI DFQT LK - -DKTVTLRERDSMLQVRIDLSDLVEIVTSLLRQ--KKTWKLES--A--PPSHIGKRKG GKKYARIDEIGI PFAVTIDFQTLK - -DKTITLRDRDSMLQIRVNI SEVSDIINSLLSQ--KSSWKLES--S--PPSHIGKRKG GKRYARTDELGVPFAITVDS-----TSSVTIRERDSKDQIRVNVEEAASVVKSVTDG--HTTWKFEP--AA-PPARVGRKQG GRRYARTDEIGVAFGITIDFDTVNKTPHTATLRDRDSMRQIRAEVSELPSVVRDLANG--NITWKYEP--P--VPTRVGKKKC GKKYARNDELGTPFGITIDFETIK--DQTVTLRERNSMRQVRGTITDVISTIDKMLHNPDESDWKYEP--P--VQSKFGRKKG GRRYARNDELGTPFGLTVDFETLQ--NETITLRERDSTKQVRGSQDEVIAALVSMVEG--KSSFKYEP--P--VPTRTGRRKG GRRYARTDEI SVPFCITVDFDSLK - EPHTVTLRDRDTFEQVRTLVSDVADIIRDLSSD--KIRWKYEP--P--VPTRVGKKKG GRRYARTDEIGVAFGITIDFDTVNKTPHTATLRDRDSMRQIRAEVRELPGIIRDLANG--TLSWKYEP--P--IPTRVGKKKG GRRYARTDELGI PFGITIDNDTLV--DDSVTLREI LTTKQIRIPINDVFRVVSDLADG--LITWKDQM--P--------REKGKRYARTDEIGI PFCVTLDFQSVN--DDTVTLRERDTMQQVRIKLDDLGELINNLLKD--DITWQSRS--D--QPVTFGKRKR GKRYARTDEIGVPFCVTLDFQSVN - -DDTVT LR ERDSMQQVRVK LDQVGQLLSNLLK - - - DI TW

EESVNPKKK - -MIKRIK
 GKKYARVDELGI PFCVTCDFET----DGCVTLRERDSARQVRI PREAVADVVAELSR PLR PR EWKWQP--P - -VASDIGKKKG GRRYARNDELGT PFACT LDFASLS - -KGTMT LRERDTTAQRIGPIDQVIDVIRQLCDG--S LDWKWEP--P--LPTRVGKKKG GKRYARND EMGT PFGI TVDFDTVK - -DNSVT LRERDSTRQVRGSIDAVIAAI NVMTAD - -DVAWKYEP--P--VNTRSHRKKG

Anopheles_gambiae/1-3220
Arabidopsis_thaliana/1-3286
Ashbya_gossypii/1-3281
A fumigatus/1-3241
Aspergillus_niger/1-3300 Bombyx_mori/1-2389
Bos_taurus/1-3273
_briggsae/1-3231
Caenorhabditis_elegans/1-3289
Candida_albicans/1-3266 Candida_glabrata/1-3282
Canis_familiaris/1-3286
Ciona intestinalis/1-2712
Cryptococcus_neoformans/1-3301 Cryptosporidium_hominis/1-3020 C_parvum/1-3281
Danio rerio/1-3286
Debaryomyces hansenii/1-3166
Dictyostelium_discoideum/1-3275 D_melanogaster/1-3287
Drosophila_pseudoobscura/1-3286 Encephalitozoon cuniculi/1-3199 Entamoeba_histolytica/1-3271 Gallus_gallus/1-3286
Homo_sapiens/1-3286
Kluyveromyces_lactis/1-3281 Leishmania_major/1-3286 Macaca_mulatta/1-3182 Magnaporthe_grisea/1-3286 Methanosarcina_acetivorans/1-3043 Monodelphis_domestica/1-3196
Mus musculus/1-3286
Neurospora_crassa/1-3264
Oryza_sativa/1-3218
Ostreococcus_lucimarinus/1-3287
Pan troglodyte/1-3286
Paramecium tetraurelia/1-3282
Pichia_stipitis/1-3290
_falciparum/1-3285
P_knowlesi/1-3284
Plasmodium_yoelii/1-3285
Populus_trichocarpa/1-3286
Rattus_norvegicus/1-3286
Saccharomyces_cerevisiae/1-3284
Schizosaccharomyces_pombe/1-3284
Strongylocentrotus_purpuratus/1-3252
Takifugu_rubripes/1-2930
Tetrahymena_thermophila/1-3270
T_annulata/1-3248
Theileria parva/1-3273
Trichomonas vaginalis/1-3198
T_brucei/1-3264
rypanosoma_cruzi/1-3266
stilago maydis/1-3291
Yarrowia_lipolytica/1-3278
 PEAAARLPTV------IPSTKCKLRLLKLERIKDYLLMEEEFVANQEDDLRGTPMSVGNLEELIDENHAIVSSSVGPEYYVGI PATAEKLPNV-----YPSTRCKLKLLRMERIKDHLLLEEEFVTNSEEEIRGTPLSIGTLEEIIDDDHAIVTSPTTPDFYVSI PSAASKLPDI------FPTSRCKLRY LRMQRVHDHLLLEEEYVENMEDDMRGSPMGVGNLEELIDDDHAIVSSATGPEYYVSI PSTASKLPDI-----FPTSRCKLRYLRMQRVHDHLLLEEEYVENMEDDMRGSPMGVGNLEELIDDDHAIVSSATGPEYYVSI -MEEEFIRNQE
 PDVASKLPLV------TPHTQCRLKLLKLERIKDYLLMEEEFIRNQEDDLRGTPMSVGTLEEIIDDNHAIVSTSVGSEHYVS
PDAASKLPAV------TPHARCRLKLLKSERIKDYLLMEQEFIQNQEDELRGTPMAVGSLEEIIDDQHAIVSTNVGSEHYVNI PDAASKLPAV------IPHARCRLKLLKSERIKDYLLMEQEFIQNQEDELRGTPMAVGSLEEIIDDQHAIVSTNVGSEHYVNI PDTAVKLPSV-----YPNTRCKLKLLKLERIKDHLLLEEEFVTNQEDELRGYPMAIGTLEEII DDDHAIVSSTASSEYYVSI PATAEKLPNV-----VPSTRCKLKLLRMERIKDHLLLEEEFVTNSEEEIRGNPLSIGTLEEIIDDDHAIVTSPTMPDYYVSI PDAASKLPLV-----TPHTQCRLKLLKLERIKDYLLMEEEFIRNQEDDLRGTPMSVGTLEEIIDDNHAIVSTSVGSEHYVSI PDAANKLPSV------TPHTRCRLRLLKQERIKDYLLMEEEYIRNQEDDLRGSPMSVGTLEEIIDENHAIVSTSVGSEHYVSI PDASSRLPAV-----YPTTRCKLKLLKMERI QDYLLMEEEFVSNQADELRGSPMGVGTLEEIIDDDHAIVSSGGGEYYVGI PPQYARLPAV-----VPNAKCRLRLLKYERIKDYLMMEQEFITSMEDDLRGSPMNIGTLEEIIDENHAIVSSSVGSEYYVNI PPQYARLPAV-----VPNAKCRLRLLKYERIKDYLMMEQEFIT SMEDDLRGSPMNIGTLEEIIDENHAIVSSSVGSEYYVNI PDAASKLPLV------IPHTQCRLKLLKQDRIKDYLLMEEEFIRNQEDDLRGTPMSVGTLEEIIDDNHAIVSTSVGSEHYVSI PDTAVKLPSV------YPSTRCKLKLLKLERIKDHLLLEEEFVTNQEDELRGYPMSIGTLEEIIDDDHAIVSSTAGSEYYVSI AETSTKLPVI-----TPHSKCKLKQLKLERIKDYLLMEQEFLQNYDEELRGDPLTVGNLEEIIDDNHAIVSSTVGPEHYVRI PDAAMKLPQV-----TPHTRCRLKLLKLERIKDYLMMEDEFIRNQEDDLRGTPMSVGNLEEIIDDNHAIVSTSVGSEHYVSI PDAAMKLPLV-----IPHTRCRLKLLKLERIKDYLMMEDEFIRNQEDDLRGTPMSVGNLEEIIDDNHAIVSTSVGSEHYVSI GKAASKPPQV-----YPLMKCKLRYLKLKKLAHLLSLEDNILSLCEEQLRGSPLSVGTLEEFVDDHHGIITTGVGLEYYVNI YGNNNKLPQI-----NPRTQCNLKKLRLERLKDILLI QRDFIENQEEELRGSPLEVSKLHEMIDDHHAIISSGNTMQYCVPV PDAASKPPLV-----TPHTQCRLKLLKLERIKDYLLMEEEFIRNQEDDLRGTPMSVGTLEEIIDDNHAIVSTSVGSEHYVSI PDAASKLPLV-----IPHTQCRLKLLKLERIKDYLLMEEEFIRNQEDDLRGTPMSVGTLEEIIDDNHAIVSTSVGSEHYVSI PSTVEKLPSV------YPSTRCKLKLLRMERIKDHLLLEEEYVTNSEDDIRGTPLSIGTLEEIVDDDHAIVTSPTTPDYYVSI PDAATRIPKV------YPNRACLLRKYRLERCKDYLLLEEEFLRTINEDIRGTPLEVATLEEAVDDSHAIVSIS-GTEYYVPL PDAASKLPLV------IPHTQCRLKLLKLERIKDYLLMEEEFIRNQEDDLRGTPMSVGTLEEIIDDNHAIVSTSVGSEHYISI TSAAAKLPAI-----YPTSRCKLRLLRMQRTHDHLLLEEEFVENQEDDMRGSPMGVGTLEEMIDDDHAIVSSTTGPEYYVSI EQLT EPPLFIATILEVNGEIALIRQHGNQQE
PDAASKLPLV-----TPHTQCRLKLLKLERIKDYLLMEEEFIRNQEDDLRGTPMSVGTLEEIIDDNHAIVSTSVGSEHYVSI PDAASKLPLV-----IPHTQCRLKLLKLERIKDYLLMEEEFIRNQEDDLRGTPMSVGILEEIIDDNHAIVSTSVGSEHYVSI ASAAAKLPSV-----YPTSRCKLRLLRMQRIHDHLLLEEEYVENQEDDMRGSPMGVGVLEELIDDDHAIVSSTSGPEYYVSI PEAAARLPNV-----APLSKCRLRLLKLERVKDYLLMEEEFVAAQEDDLRGTPMSVGSLEEIIDESHAIVSSSVGPEYYVGI I EGSTRLPNV-----APQSKCKLRMLKLERVKDYLLMEEEFVGNQEDEMRGAPMSVGSLEEIIDDTHGIVSSSIGPEYYVNI PDAASKLPLV-----TPHTQCRLKLLKLERIKDYLLMEEEFIRNQEDDLRGTPMSVGTLEEIIDDNHAIVSTSVGSEHYVSI VEQASKLPAA-----IPITKCRLKLLKNERIKDYLLLEQEFIENQQEQLRGIPMIVGTLEEFVNENHAIVSSSVGPESYSGI PDTAVKLPSV-----YPNTRCKLKLLKLERIKDHLLLEEEFVTNQEDELRGYPMSIGTLEEIIDDDHAIVSSTAGSEYYVSI TSGHSKLPNV-----TPNTKCRLKLLKLERIKDYLLLEEEYITNQEDDLRGSPVSVGTLEELIDENHGIIATSVGPEYYVNI VPGHSKLPTV------TPNTKCRLKLLKLERIKDYLLLEEEFITNQEDDLRGSPMSVGTLEELIDENHGIIATSVGPEYYVNI ATGHSKLPTV------TPNTKCRLKLLKLERIKDYLLLEEEFITNQEDDLRGSPMSVGILEELIDENHGIIATSVGPEYYVNI PEAAARLPTV-----TPHTKCKLRLLKMERIKDYLLMEEEFVANQEDDLRGSPMSVGNLEELIDENHAIVSSSVGPEYYVGI PDAASKLPLV-----TPHTQCRLKLLKLERIKDYLLMEEEFIRNQEDDLRGTPMSVGTLEEII DDNHAIVSTSVGSEHYVSI PATAEKLPNI-----YPSTRCKLKLLRMERIKDHLLLEEEFVSNSEEEIRGNPLSIGTLEEIIDDDHAIVTSPIMPDYYVSI PDASAKLPTV------IPTTRCRLRLLKMQRIHDHLLMEEYVQNQEDEIRGTPMSVGTLEEIIDDDHAIVSTA-GPEYYVSI PDTASKLPQV-----TPHTKCRLRLLKMERIKDYLLMEEEFIRNQEDDLRGTPMSVGSLEEIIDDNHAIVSASVGSEYYVSI PDAASKLPLV-----TPHTQCRLKLLKQERIKDYLLMEEEFIRNQEDDLRGTPMSVGTLEEIIDDNHAIVSTSVGSEHYVSI --KPAPVRIV-----TPISKCRLRQLKLDRIKDYLLMEQEFIRNQEEQLRGSPMLIGTLEEFIDEDHAIVSS-IGPEYYANI QLAPVRIPTV------TPNSKCRLRLLKLERIKDYLLLEEEYITNKSDDIRGSPMSVGTLEEIIDENHAIVTSSIGPEYYVNI PVLINRLPLVNVKGKLTPNSKCRLRLLKLERIKDYLLLEEEYITNKSDDIRGSPMSVGTLEEIIDENHAIVTSSIGPEYYVNI GAKNTHI PTV-----IPNAKCQLRLLKLERVKDWLKMEEEFINNCEEEVRGSPMMVGTLEEIVDDDHAIVSRSV-QDFYVII PDAAAKLPKI------YPSRACLLKQLRLERCKDYLLLEDELLTMITDALRGMPLEVGTLEEVIDDTHAIVSTA-GSEYYVAM PDTAAKLPKI------YPVKACLLKQLRLERCKDYLLLEEELLKTIGDALRGMPLEVGTLEEVIDDTHAIVSTA-GSEYYVPM PDSSSKLPTV------YPNTRCRLKLLKLERIKDHLLLEEEFVQNQEDDLRGSPMAVGTLEEIIDDEHAIVSSATGPEYYVSI PENANKLPGV------YPTTRCKLKLLKMERIKDHLLLEEEFVQNQEDT LRGSPMGVGTLEEIIDDDHAIVSST SGPEYYVSI

# 590 

Anopheles_gambiae/1-3220 Arabidopsis_thaliana/1-3286 Ashbya_gossypii/1-3281
A fumigatus/1-3241 Aspergillus_niger/1-3300 Bombyx_mori/1-2389
Bos_taurus/1-3273
C_briggsae/1-3231
Caenorhabditis elegans/1-3289
Candida_albicans/1-3266 Candida_glabrata/1-3282
Canis_familiaris/1-3286
Ciona intestinalis/1-2712
Cryptococcus_neoformans/1-3301 Cryptosporidium_hominis/1-3020 C_parvum/1-3281
Danio rerio/1-3286
Debaryomyces hansenii/1-3166
Dictyostelium_discoideum/1-3275 D_melanogaster/1-3287
Drosophila_pseudoobscura/1-3286 Encephalitozoon_cuniculi/1-3199 Entamoeba_histolytica/1-3271 Gallus_gallus/1-3286
Homo_sapiens/1-3286
Kluyveromyces_lactis/1-3281 Leishmania_major/1-3286 Macaca_mulatta/1-3182 Magnaporthe_grisea/1-3286 Methanosarcina_acetivorans/1-3043 Monodelphis_domestica/1-3196
Mus musculus/1-3286
Neurospora_crassa/1-3264
Oryza_sativa/1-3218
Ostreococcus_lucimarinus/1-3287
Pan_troglodyte/1-3286
Paramecium tetraurelia/1-3282
Pichia_stipitis/1-3290
_falciparum/1-3285
P_knowlesi/1-3284
Plasmodium_yoelii/1-3285
Populus_trichocarpa/1-3286
Rattus_norvegicus/1-3286
Saccharomyces_cerevisiae/1-3284
Schizosaccharomyces_pombe/1-3284
Strongylocentrotus_purpuratus/1-3252
Takifugu_rubripes/1-2930
etrahymena_thermophila/1-3270
T_annulata/1-3248
Theileria parva/1-3273
Trichomonas vaginalis/1-3198
T_brucei/1-3264
rypanosoma_cruzi/1-3266
stilago maydis/1-3291
Yarrowia_lipolytica/1-3278

LSFVDKD-- QLEPGCSV LLNHKV'HAVVGVLGDDTDPMVTVMK'LEKAPQETYADIGGLDTQIQEIKESVELPLTHPEYYEEMG LS FVDKD---QLEPGCSI LMHNKVLSVVGI LQDEVDPMVSVMKVEKAPLESYADIGGLEAQIQEIKEAVELPLTHPELYEDIG LS FVDK E-- - LLEPGCSVLLHHKTMSIVGV LQDDADPMV SVMKI DK SPTESYNDIGGLEAQIQEIKEAVELPLTHPELYEEMG MSFVDKD-- - LLEPGASI LLHHK SVSVVGVLTEESDPLVSVMK LDKAPTESYADIGGLESQIQEVRESVELPLLHPELYEEMG MS FVDKD MS FVDKD --LLEPGASI LLHHKS
 - - LLEPGCSV L LNHKVHAVIGVLMDDTDPLVTVMKVEKA PQETYADIGGLDNQIQEIKESVELPLTHPEYYEEMC MS FVDK E---QLEPGCSVLLNHKNHAVIGVLSDDTDPMVSVMKLEKAPQETYADVGGLDQQIQEIKEAVELPLTHPEYYEEMG MS FVDKE---QLEPGCSVLLNHKNHAVIGVLSDDTDPMVSVMKLEKAPQETYADVGGLDQQIQEIKEAVELPLTHPEYYEEMC MS FVDKG-- - L E PGC SV L LHHKTVAVVGV LQDDAD PMV SVMK LDK SPTESYADIGGLESQIQEIKESVELPLTHPELYEEMC LS FVDK E-- - L L EPGCSV L LHHKTMSIVGV LQDDADPMV SVMKI DK SPTESYSDIGGLESQIQEIKESVELPLTHPELYEEMC LS FVDKD-- - LLELGCSVLLNHKVHAVIGVLMDDTDPLVIVMKVEKTPQETYADIRALDNQIQEIKESVELPLTHPEYYEEMG LS FVDKD-- - LLEPGCTV LMNHKVHAVVGFLGDDVDPLVTVMKLEKAPKESYADIGGLDTQITEIKESVELPLTHPEYYEEMC MS FVDKD-- - LLEPGCSV L LHHKTHAVVGV LADDTDPMV SVMK LDKAPTESYADIGGLESQI QEIKESVELPLTHPELYEEMC LS FVDKN -- -QLEPGS SV LLHNKVY SVVGI MNDEVDPLVSVMKVDKAPLESYADIGGLEQQIQEIKEAVEIPLTHPELYDDIG LSFVDKN-- QLEPGSSVLLHNKVYSVVGI MNDEVDPLVSVMKVDKAPLESYADIGGLEQQIQEIKEAVEIPLTHPELYDDIG LS FVDKD---LLEPGCSVLLNHKVHAVIGVLMDDTDPLVTVMKVEKAPQETYADIGGLDNQIQEIKESVELPLTHPEYYEEMC MSFVDKG-- - LLEPSCSVLLHHKTVSIVGV LQDDADPMVSVMK LDKSPTESYADIGGLESQIQEIKEAVELPLTHPELYEEMG MS FVDK $---K$ LY LGATV LLNNKT LSVVGVI DGEVDPMVNVMKVEKAPTESYSDIGGLEAQVQEMKEAI ELPLTHPELYEEIG LS FVDKD-- QLEPGCSVLLNHKVHAVVGVLSDDTDPMVTVMKLEKAPQETYADIGGLDTQIQEIKESVELPLTHPEYYEEMC LS FVDKD-- -QLEPGCSVLLNHKVHAVVGVLSDDTDPMVTVMKLEKAPQETYADIGGLDTQIQEIKESVELPLTHPEYYEEMC MS FVDKD---LLEPGCTVLLNYKDNSVVGVLEGEMDPMVNVMKLEKAPSETYADIGGLEEQIQEIKESVELPLTNPELYQEMG LSIVDRE-- - LLEPGVQVLTHNHNKAIVGVLQNDEDPHVSVMKVDKAPLESYADVGGLEKQIQEIKEAVELPLSHPELYEEIG LS FVDKD-- LLEPGCSVLLNHKVHAVIGVLMDDTDPLVTVMKLEKAPQETYADIGGLDNQIQEIKESVELPLTHPEYYEEMG LS FVDKD-- - LLEPGCSV LLNHKVHAVIGV LMDDTDPLVTVMKVEKAPQETYADIGGLDNQIQEIKESVELPLTHPEYYEEMG LS FVDK E-- LLEPGCSV L LHHKTMSVVGV LQDDAD PMV SVMKMDKSPTENYSDIGGLEAQIQEIKEAVELPLTHPELYEEMG MS FVDK E-- -QLELGCSVLLHDRQHSIVGVLKDDVDPLVSVMKVDKAPEDTYADIGGLEQQIQEIKEAVEFPLSHPELYDEIG LS FVDKD-- LLEPGCSVLLNHKVHAVIGVLMDDTDPLVTVMKVEKAPQETYADIGGLDNQIQEIKESVELPLTHPEYYEEMG MSFVDKD---LLEPGASVLLHHKSVSIVGVLTDDADPLVSVMK LDKAPTESYADIGGLEQQI QEVRESVELPLLHPELYEEMC LTQI PEECLGKI EPGMRVAVN-GAYSIISIVSRAADVRAQVMELINSPGVDYSMIGGLDDVLQEVRESVELPLTEPELFEDLG LSFVDKD---LLEPGCSVLLNHKVHAVIGVLMDDTDPLVTVMKVEKAPQETYADIGGLDNQIQEIKESVELPLTHPEYYEEMC LS FVDKD-- LLEPGCSVLLNHKVHAVIGVLMDDTDPLVTVMKVEKAPQETYADIGGLDNQIQEIKESVELPLTHPEYYEEMG MS FVDKD-- LLEPGASV LLHHKSVSIVGVLTDDTDPAVSVMK LDKAPTESYADIGGLEQQI QEVRESVELPLLHPELYEEMC LS FVDKD---QLEPGCSI LMHNKVLSVVGI LQDEVDPMVSVMKVEKAPLESYADIGGLDAQIQEIKEAVELPLTHPELYEDIG ASFVDKS---QLEPGCAVLLHHKNSAVVGTLADDVDPMVSVMKVDKAPLESYADVGGLEDQIQEIKEAVELPLTHPELYEDIG LSFVDKD---LLEPGCSVLLNHKVHAVIGVLMDDTDPLVTVMKVEKAPQETYADIGGLDNQIQEIKESVELPLTHPEYYEEMG MS FVDKD---QLEPGCSVLLNQR SYAVVGI MQDEI DPLLNVMKVDKAPLESYADIGGLEQQIQEIKEAVELPLTHPEIYEDMG MS FVDKG-- - LLEPGC SV L LHHKTV SVVGV LQDDAD PMV SVMK LDK SPTESYADIGGLESQI QEIKEAVELPLTHPELYEEMG LSFVDKD-- LLEPGCSVLLNNKTNSVVGI LLDEVDPLVSVMKVEKAPLESYADIGGLESQIQEIKEAVELPLTHPELYEDIG LSFVDKD- - - LLEPGCSVLLNNKTNSVVGI LLDEVDPLVSVMKVEKAPLESYADIGGLESQIQEIKEAVELPLTHPELYEDIG LSFVDKD-- - LLEPGCSVLLNNKTNSVVGI LLDEVDPLVSVMKVEKAPLESYADIGGLESQIQEIKEAVELPLTHPELYEDIG LS FVDKD-- QLEPGCAI LMHNKVLSVVGLLQDEVDPMVSVMKVEKAPLESYADIGGLDAQIQEIKEAVELPLTHPELYEDIG LS FVDKD-- LLEPGCSVLLNHKVHAVIGVLMDDTDPLVTVMKVEKAPQETYADIGGLDNQIQEIKESVELPLTHPEYYEEMG LS FVDK E-- - LLEPGC SV L LHHKTMSIVGV LQDDAD PMV SVMKMDK SPTESYSDIGGLESQIQEIKESVELPLTHPELYEEMG MS FVDKD---MLEPGCSVL LHHKAMSIVGLLLDDTDPMINVMK LDKAPTESYADIGGLESQIQEIKEAVELPLTHPELYEEMG LS FVDKD-- -QLEPGCTVLLNHKVLAIVGVLGDDTDPMV SVMKLEKAPQESYADIGGLDTQIQEIKESVELPLTHPEYYEEMG LS FVDKD-- LLEPGCSVLLNHKVHAVIGVLMDDTDPLVTVMKVEKAPQETYADIGGLDNQIQEIKESVELPLTHPEYYEEMG LS FVDKD---QLEPGSTVLLNNRTMAVVGI MQDEVDPMLNVMKVEKAPLECYADIGGLEQQIQEVKEAVELPLTHPEIYEDMG LSFVDKE---LLEPGCSVLLHNKTNSIVGI LLDDVDPLVSVMKVEKAPLESYDDIGGLEEQIQEIKEAVELPLTRPELYDDIG LSFVDKE---LLEPGCSVLLHNKTNSIVGI LLDDVDPLVSVMKVEKAPLESYDDIGGLEEQIQEIKEAVELPLTRPELYDDIG SS FVDRK - - -A LQI GC SV LLHEKALTIVGLLDDDANPLVDVMKVENAPLESFADIGGLEDQIVDIKEAVELPLTHPEQFDEIG LS FVDK E-- -K LELGC SV L LHDRYHNVVGLLESNTDPLVSVMKVDKAPQETYADIGGLEDQIQEIKEAVEFPLSHPELFDEVG LS FVDK E-- K L ELGCSV L LHDRQHSVVGVLQNSI DPHVSIMKVEKAPQETYADIGGLEEQIQEIKEAVEFPLSHPELYDEVG MSFVDKD---L LEPGCSV L LHHKAMAIVGVLSDDADPMV SVMKLDKAPSESYADIGGLETQIQEIKEAVELPLTHPELYEEMG MSFVDKD- - LLEPGCSVLLHHKTVSVVGVLQDDADPMV SVMKLDKAPTESYADIGGLESQIQEIKESVELPLTHPELYEEMC

Anopheles_gambiae/1-3220 Arabidopsis_thaliana/1-3286
Ashbya_gossypii/1-3281
A fumigatus/1-3241
Aspergillus_niger/1-3300 Bombyx_mori/1-2389
Bos_taurus/1-3273
C briggsae/1-3231
Caenorhabditis elegans/1-3289
Candida_albicans/1-3266
Candida_glabrata/1-3282
Canis_familiaris/1-3286
Ciona intestinalis/1-2712
Cryptococcus_neoformans/1-3301 Cryptosporidium_hominis/1-3020 _parvum/1-3281
Danio_rerio/1-3286
Debaryomyces hansenii/1-3166
Dictyostelium_discoideum /1-3275 D_melanogaster/1-3287
Drosophila_pseudoobscura/1-3286 Encephalitozoon cuniculi/1-3199 Entamoeba_histolytica/1-3271 Gallus_gallus/1-3286
Homo_sapiens/1-3286
Kluyveromyces_lactis/1-3281 Leishmania_major/1-3286 Macaca_mulatta/1-3182 Magnaporthe_grisea/1-3286 Methanosarcina_acetivorans/1-3043 Monodelphis_domestica/1-3196
Mus musculus/1-3286
Neurospora_crassa/1-3264
Oryza_sativa/1-3218
Ostreococcus_lucimarinus/1-3287
Pan_troglodyte/1-3286
Paramecium tetraurelia/1-3282
Pichia_stipitis/1-3290
_falciparum/1-3285
P_knowlesi/1-3284
Plasmodium_yoelii/1-3285
Populus_trichocarpa/1-3286
Rattus_norvegicus/1-3286
Saccharomyces_cerevisiae/1-3284
Schizosaccharomyces_pombe/1-3284
Strongylocentrotus_purpuratus/1-3252
Takifugu_rubripes/1-2930
etrahymena_thermophila/1-3270
T_annulata/1-3248
Theileria parva/1-3273
Trichomonas vaginalis/1-3198
T_brucei/1-3264
Trypanosoma_cruzi/1-3266
stilago maydis/1-3291
Yarrowia_lipolytica/1-3278

PPKGVI LYGPPGTGKTLLAKAVANQT SATFLRVVGSELI QKY LGDGPKLVRELFRVAEEHAPSIVFIDEIDAVGTKRYDSN K PPKGVI LYGEPGTGKT LLAKAVANST SATFLRVVGSELI QKY LGDGPKLVRELFRVADDLSPSIVFIDEIDAVGTKRYDAN K PPKGVI LYGAPGTGKT LLAKAVANQT SATFLRIVGSELI QKY LGDGPR LCRQI FKVAAENAPSIVFIDEIDAIGTKRYDSN K PPKGVI LYGAPGTGKTLLAKAVANQT SATFLRIVGSELI QKY LGDGPRLVRQI FQVAAEHAPSIVFIDEIDAIGTKRYDST K PPKGVI LYGAPGTGKT LLAKAVANQT SATFLRIVGSELI QKYLGDGPRLVRQI FQVAAEHAPSIVFIDEIDAIGTKRYDST K PPKGVI LYGPPGTGKTL LAKAVANQT SATFLRVVGSELI QKY LGDGPKLVRELFRVAEEHAPSIVFIDEIDAVGTKRYDSN K PPKGVI LYGPPGTGKTLLAKAVANQT SATFLRVVGSELI QKYLGDGPKLVRELFRVAEEHAPSIVFIDEIDAIGTKRYDSN R PPKGVI LYGCPGTGKT L LAKAVANQT SATFLRIVGSELI QKY LGDGPKMVRELFRVAEENAPSIVFIDEIDAVGTKRYDSN R PPKGVI LYGCPGTGKT L LAKAVANQT SAT FLRIVGSELI QKY LGDGPKMVRELFRVAEENAPSIVFIDEIDAVGTKRYDSN K PPKGVI LYGAPGTGKT LLAKAVANQT SATFLRIVGSELI QKY LGDGPR LCRQI FQI AADHAPSIVFIDEIDAIGTKRYEST K PPKGVI LYGAPGTGKTLLAKAVANQT SATFLRIVGSELI QKY LGDGPRLCRQI FKVAAENAPSIVFIDEIDAIGTKRYDSN K PPKGVI LYGPPGTGKTLLAKAVANQT SATFLRVVGSQLI QKY LGNGPKLIRELFRVVEEHAPSIVFIDEIDAIGTKRYDSN KPPKGVI LYGPPGTGKTLLAKAVANQT SATFLRVVGSELI QKY LGDGPKLVRELFRVAEEHAPSIVFIDEIDAIGTKRYESN R PPKGVI LYGVPGTGKTLLAKAVANQT SATFLRIVGSELI QKY LGDGPKLVRELFRVAEENAPSIVFIDEIDAIGTKRYDST K PPKGVI LYGPPGTGKTLLAKAVANETSATFLRVVGSELI QKYLGDGPKLVRELFRVAEENAPSIVFIDEIDAVGTKRHDSQ K PPKGVI LYGPPGTGKT LLAKAVANET SATFLRVVGSELI QKY LGDGPKLVRELFRVAEENAPSIVFIDEIDAVGTKRHDSQ K PPKGVI LYGAPGTGKTLLAKAVANQT SATFLRVVGSELI QKY LGDGPKLVRELFRVAEEHAPSIVFIDEIDAIGTKRYDSN K PPKGVI LYGAPGTGKT LLAKAVANQT SATFLRIVGSELI QKY LGDGPR LCRQI FQIAGELAPSIVFIDEIDAIGSKRYESS K PPKGVI LYGEPGTGKTLLAKAVANQT SATFLRVVGSELI QKY LGDGPKLVRELFRVADECAPSIVFIDEIDAVGTKRYDSQ K PPKGVI LYGPPGTGKTLLAKAVANQT SATFLRVVGSELI QKY LGDGPKLVRELFRVAEEHAPSIVFIDEIDAVGTKRYDSN K PPKGVI LYGPPGTGKTLLAKAVANQT SATFLRVVGSELI QKY LGDGPKLVRELFRVAEEHAPSIVFIDEIDAVGTKRYDSN KPPKGVI LYGLPGTGKT LLAKAVANQT SATFLRVVGTELIQEY LGEGPKLVRELFRVADMHAPSIIFIDEIDAIGGKRYNTS K PPKGVI LYGPPGTGKT L LAKAVANET SAT FLRIVGSELI QKY LGDGPKLVRELFQAAKDSAPSIVFIDEIDAVGTKRYDAH K PPKGVI LYGPPGTGKT LLAKAVANQT SAT FLRVVGSELI QKY LGDGPKLVRELFRVAEEHGPSIVFIDEIDAIGTKRYDSN K PPKGVI LYGPPGTGKTLLAKAVANQT SAT FLRVVGSELI QKY LGDGPKLVRELFRVAEEHAPSIVFIDEIDAIGTKRYDSN K PPKGVI LYGAPGTGKT LLAKAVANQT SAT FLRIVGSELI QKY LGDGPRLCRQI FKVAAENAPSIVFIDEIDAIGTKRYESN K PPKGVI LYGVPGTGKT LLAKAVANRT SATFLRVVGSELI QKYSGEGPKLVRELFRVAEEHSPAIVFIDEIDAIGTKRYDTD K PPKGVI LYGPPGTGKT LLAKAVANQT SATFLRVVGSELI QKY LGDGPKLVRELFRVAEEHAPSIVFIDEIDAIGTKRYDSN I K PPKGVI LYGAPGTGKT LLAKAVANQT SATFLRIVGSELIQKYLGDGPRLVRQLFQVAAENAPSIVFIDEIDAIGTKRYDST I EPPSGVLLHGAPGTGKTLIAKAIASQAKATFIRMSGSDLVQKFVGEGSRLVKDI FQLARDKSPSILFIDEIDAVGSMRTYDG I K PPKGVI LYGPPGTGKTLLAKAVANQT SATFLRVVGSELI QKY LGDGPKLVRELFRVAEEHAPSIVFIDEIDAIGTKRYDSN I K PPKGVI LYGPPGTGKTLLAKAVANQT SATFLRVVGSELI QKYLGDGPKLVRELFRVAEEHAPSIVFIDEIDAIGTKRYDSN I K PPKGVI LYGAPGTGKT LLAKAVANQT SATFLRIVGSELI QKY LGDGPRLVRQLFQVAAENAPSIVFIDEIDAIGTKRYDST I R PPKGVI LYGEPGTGKTLLAKAVANST SATFLRVVGSELI QKY LGDGPKLVRELFRVADELSPSIVFIDEIDAVGTKRYDAH K PPKGVI LYGAPGTGKTLLAKAVANST SATFLRIVGSELI QKY LGDGPKLVRELFRVADEMSPSIVFMDEIDAVGTKRYDSQ K PPKGVI LYGPPGTGKT LLAKAVANQT SATFLRVVGSELI QKY LGDGPKLVRELFRVAEEHAPSIVFIDEIDAIGTKRYDSN KPPKGVI LYGEPGTGKT LLAKAVANETSATFLRVVGSELI QKYQGDGPKLVRELFRVAEEHAPSIVFIDEIDAVGTKRYDSH K PPKGVI LYGAPGTGKT LLAKAVANQT SATFLRIVGSELI QKY LGDGPRLCRQI FQI AGEHAPSIVFIDEIDAIGTKRYEST K PPKGVI LYGPPGTGKT LLAKAVANET SAT FLRVVGSELI QKY LGDGPKLVREMFKVAEEHAPSIVFIDEIDAVGTKRYEAT KPPKGVI LYGPPGTGKT LLAKAVANET SATFLRVVGSELI QKY LGDGPKLVREMFKVAEDHAPSIVFIDEIDAVGTKRYEAT K PPKGVI LYGPPGTGKT LLAKAVANET SATFLRVVGSELI QKY LGDGPK LVREMFKVAEDHAPSIVFIDEIDAVGTKRYEAT I KPPKGVI LYGEPGTGKT LLAKAVANST SATFLRVVGSELI QKY LGDGPKLVRELFRVADDLSPSIVFIDEIDAVGTKRYDAH K PPKGVI LYGPPGTGKTLLAKAVANQT SATFLRVVGSELI QKYLGDGPKLVRELFRVAEEHAPSIVFIDEIDAIGTKRYDSN K PPKGVI LYGAPGTGKTLLAKAVANQT SATFLRIVGSELI QKYLGDGPRLCRQI FKVAGENAPSIVFIDEIDAIGTKRYDSN K PPKGVI LYGAPGTGKT L LAKAVANQT SAT FLRVVGSELI QKY LGDGPR LVRQLFNAAEEHSPSIVFIDEIDAIGTKRYDAQ R PPKGVI LYGAPGTGKT LLAKAVANQT SATFLRVVGSELI QKY LGDGPKLVRELFRVAEEHAPSIVFIDEIDAIGTKRYESN K PPKGVI LYGPPGTGKTLLAKAVANQT SAT FLRVVGSELI QKY LGDGPKLVRELFRVAEEHAPSIVFIDEIDAIGTKRYDSN K PPKGVI MYGPPGTGKT L LAKAVANET SAT FLRIVGSELI QKYAGEGPKLVRELFRVAEEHAPSIVFIDEIDAVGSKRYNTS K PPKGVI LYGPPGTGKTLLAKAVANETSATFLRVVGSELI QKYLGEGPKLVREMFKVAEDNAPSIIFIDEIDAIGTKRYDAT I K PPKGVI LYGPPGTGKT LLAKAVANET SATFLRVVGSELI QKY LGEGPKLVREMFKVAEDNAPSIIFIDEIDAIGTKRYDAT VQPPKGVI LFGPPGTGKT LLARAVAKST SATFLRVVGSELI QKYLGEGPKLVRELFKTAHELAPSIVFIDEIDAVGTKRYDST VK PPKGVI LYGV PGTGKTLLAKAVANQT SATFLRVVGSELI QKYSGEGPKLVRELFRVAEENSPSIVFIDEIDAIGTKRYDTD I K PPKGVI LYGVPGTGKT LLAKAVANQT SAT FLRVVGSELIQKYSGDGPKLVRELFRVAEENSPSIVFIDEIDAIGTKRYDTD I R PPKGVI LYGV PGTGKTLLAKAVANQT SATFLRVVGSELI QKY LGDGPKLVRELFRVADEHAPSIVFIDEIDAVGTKRYDSN IKPPKGVI LYGAPGTGKTLLAKAVANQT SATFLRIVGSELIQKYLGDGPRLCRQI FQIAAEHAPSIVFIDEIDAIGTKRYEST

Anopheles_gambiae/1-3220 Arabidopsis_thaliana/1-3286
Ashbya_gossypii/1-3281
A fumigatus/1-3241
Aspergillus_niger/1-3300 Bombyx_mori/1-2389
Bos_taurus/1-3273
C_briggsae/1-3231
Caenorhabditis_elegans/1-3289
Candida_albicans/1-3266
Candida_glabrata/1-3282
Canis_familiaris/1-3286
Ciona intestinalis/1-2712
Cryptococcus_neoformans/1-3301 Cryptosporidium_hominis/1-3020 C_parvum/1-3281
Danio_rerio/1-3286
Debaryomyces hansenii/1-3166
Dictyostelium_discoideum /1-3275
D_melanogaster/1-3287
Drosophila_pseudoobscura/1-3286
Encephalitozoon_cuniculi/1-3199
Entamoeba_histolytica/1-3271
Gallus_gallus/1-3286
Homo_sapiens/1-3286
Kluyveromyces_lactis/1-3281
Leishmania_major/1-3286
Macaca_mulatta/1-3182
Magnaporthe_grisea/1-3286
Methanosarcina_acetivorans/1-3043
Monodelphis_domestica/1-3196
Mus musculus/1-3286
Neurospora_crassa/1-3264
Oryza_sativa/1-3218
Ostreococcus_lucimarinus/1-3287
Pan troglodyte/1-3286
Paramecium tetraurelia/1-3282
Pichia_stipitis/1-3290
_falciparum/1-3285
P knowlesi/1-3284
Plasmodium_yoelii/1-3285
Populus_trichocarpa/1-3286
Rattus_norvegicus/1-3286
Saccharomyces_cerevisiae/1-3284
Schizosaccharomyces_pombe/1-3284
Strongylocentrotus_purpuratus/1-3252
Takifugu_rubripes/1-2930
Tetrahymena_thermophila/1-3270
T_annulata/1-3248
Theileria parva/1-3273
Trichomonas vaginalis/1-3198
__brucei/1-3264
rypanosoma_cruzi/1-3266
stilago maydis/1-3291
Yarrowia_lipolytica/1-3278
 SGGEREI QRTMLELLNQLDGFD-SRGDVKVI LATNRIESLDPALLRPGRIDRKIEFPLPDIKTRRRIFQIHTSKMTLAEDVNL SGGEREI QRTMLELLNQLDGFD-DRGDVKVIMATNKIESLDPALIRPGRIDRKI LFENPDVSTKRKILGIHT SKMNLSADVDL SGGEREI QRTMLELLNQLDGFD-DRGDVKVIMATNKIETLDPALIRPGRIDRKILFENPDQNTKKKIFTLHT SKMSLADDVDL SGGEREI QRTMLELLNQLDGFD-DRGDVKVIMATNKI ET LDPALIRPGRIDRKILFENPDQNTKKKI FT LHT SKMS LGDDVDL SGGEREIQRTMLELLNQLDGFD-SRGDVKVIMATNRIETLDPALIRPGRIDRKIEFPLPDEKTKRRIFTIHTSRMT LADDVNL SGGEREI QRTMLELLNQLDGFD-SRGDVKVIMATNRIETLDPALIRPGRIDRKIEFPLPDEKTKKRIFQIHTSRMT LADDVTL SGGEREI ORTMLELLNOLDGFD SGGEREI ORTMLELLNQLDGFD SGGEREI QRTMLELLNQLDGFD信 GG SGGEREIRTMLLLLNQLDGFD-SRGDVKVIMATNRI ETLDPALIRPGRIDRKI EFPLPDEKTKKRIFQIHTSRMI LADAVT SSGEREI QRTMLELLNQLDGFD-TRGDVKVIMATNKI ENLDPALIRPGRIDRKIEFPLPDTKTKRHI FKLHT SRMSLADDVDI SGGERDI QRTMLELLNQLDGFE-ARGDVKVIMATNKIESLDPALIRPGRIDRKIELPNPDCKTKRRIFQIHTSKMTLSDDVDL SGGERDI QRTMLELLNQLDGFE-ARGDVKVIMATNKIESLDPALIRPGRIDRKIELPNPDCKTKRRIFQI HT SKMT LSDDVDL SGGEREI QRTMLELLNQLDGFD-SRGDVKVIMATNRIETLDPALIRPGRIDRKIEFPLPDEKTKRRIFQIHT SRMTVAEDVSL SGGEREVQRTMLELLNQLDGFD-DRGDIKVIMATNKIESLDPALIRPGRIDRKI LFENPDSNTKKRILHIHTSKMSLADDVKL SGGEREI QRTMLELLNQLDGFD-ARTDVKVIMATNRIETLDPALIRPGRIDRKIEFPLPDIKTKRKIFEI HTAKMNLSEDVNL SGGEREIQRTMLELLNQLDGFD-SRGDVKVIMATNRIETLDPALIRPGRIDRKIEFPLPDEKTKRRIFTIHTSRMTLAEDVNL SGGEREI QRTMLELLNQLDGFD-SRGDVKVIMATNRIETLDPALIRPGRIDRKIEFPLPDEKTKRRIFTIHTSRMT LAEDVNL SGGRREVQRTMLELLNQLDGFD-TRNDIKVIMATNKI EALDPALIRPGRIDRKIEFGMPDAATKKKIFDIHT SRMT LDESVNI SGGEKEI QRTMLELLNQLDGFD-TRGEVKVIIATNRIESLDSALIRPGRIDRKIEFPLPDIKTKRKIFEI HTSKMTLEEGVDM SGGEREIQRTMLELLNQLDGFD-SRGDVKVIMATNRIETLDPALIRPGRIDRKIEFPLPDEKTKKRIFQIHTSRMT LADDVTL SGGEREI QRTMLELLNQLDGFD-SRGDVKVIMATNRIETLDPALIRPGRIDRKIEFPLPDEKTKKRIFQIHTSRMTLADDVTL SGGEKEI QRTMLELLNQLDGFD-DRGDVKVIMATNKIESLDPALIRPGRIDRKILFENPDITTKRKIVGIHTSKMNLAEDVDL SSGTKEVQRTMLELLTQLDGFD-SSNDVKVIMATNRIDTLDPALIRPGRIDRKIEFPFPDEKTKRRIFEI HT SRMSLAEDVDI SGGEREI QRTMLELLNQLDGFD-SRGDVKVIMATNRIETLDPALIRPGRIDRKI EFPLPDEKTKKRIFQIHTSRMT LADDVTL SGGEREVQRTMLELLNQLDGFD-DRGDVKVIMATNKIETLDPALIRPGRIDRKI LFENPDQNTKRKIFTLHTSKMSLNEDVDL TSGSAEVNRTMLQLLAEMDGFD-PKGNVKVVAATNRIDLLDPALLRPGRFDRSI EVPLPDEKGRVEI LKI HTRKMK LADDVDF SGGEREI QRTMLELLNQLDGFD-SRGDVKVIMATNRIETLDPALIRPGRIDRKI EFPLPDEKTKKRIFQIHT SRMT LADDVT SGGEREI QRTMLELLNQLDGFD-SRGDVKVIMATNRIETLDPALIRPGRIDRKIEFPLPDEKTKKRIFQIHT SRMT LADDVTL SGGEREI QRTMLELLNQLDGFD-DRGDVKVIMATNKIESLDPALIRPGRIDRKILFENPDQNTKRKI FTLHT SKMSLNEDVDL SGGEREI QRTMLELLNQLDGFD-SRCDVKVI LATNRIESLDPALLRPGRIDRKIEFPLPDIKTRRRIFQI HTSKMTLADDVNL SGGEREI QRTMLELLNQMDGFD-SRGDVKVIMATNRIESLDPALLRPGRIDRKIEFPLPDVKTKRHIFNI HTGRMNLSADVQL SGGEREI QRTMLELLNQLDGFD-SRGDVKVIMATNRIETLDPALIRPGRIDRKIEFPLPDEKTKKRIFQIHT SRMT LADDVTL SGGEKEI QRTMLELLNQLDGFD-SRADVKVI LATNKIESLDPALIRPGRIDRKIEFPLPDVKNKKKIFQIHTSKMNLGEDANL SGGEREI QRTMLELLNQLDGFD-DRGDIKVIMATNKI ESLDPALIRPGRIDRKI LFENPDANTKKKI LTIHTSKMSLADDVNL SGGEREI QRTMLELLNQLDGFD-SRGDVKVIMATNRIDSLDPALIRPGRIDRKIQLPNPDTKTKRRIFQIHTSKMTMSPDVDL SGGEREI QRTMLELLNQLDGFD-SRGDVKVIMATNRIDSLDPALIRPGRIDRKIQLPNPDTKTKRRIFQIHT SKMTMSPDVDL SGGEREI QRTMLELLNQLDGFD-SRGDVKVIMATNRIDSLDPALIRPGRIDRKIQLPNPDTKTKRRIFQIHTSKMTMSPDVDL SGGEREI QRTMLELLNQLDGFD-SRGDVKVI LATNRIESLDPALLRPGRIDRKIEFPLPDIKTRRRIFQIHTARMT LADDVNL SGGEREI QRTMLELLNQLDGFD-SRGDVKVIMATNRIETLDPALIRPGRIDRKIEFPLPDEKTKKRIFQIHT SRMTLADDVTL SGGEREI QRTMLELLNQLDGFD-DRGDVKVIMATNKIETLDPALIRPGRIDRKI LFENPDLSTKKKILGIHT SKMNLSEDVNL SGAEREI QRTMLELLNQLDGFDTSQRDIKVIMATNRISDLDPALIRPGRIDRKI LFENPDEATKRKI FTIHT SKMNLGEDVNL SGGEREI QRTMLELLNQLDGFD-SRGDVKVIMATNRIETLDPALIRPGRIDRKI EFPLPDEKTKRRIFNIHTSRMTLSNDVNL SGGEREI QRTMLELLNQLDGFD-SRGDVKVIMATNRIETLDPALIRPGRIDRKIEFPLPDEKTKRRIFQIHTSRMTVADDVTL SGGEKEI QRTMLELLNQLDGFD-SRTDVKVI LATNKIESLDPALIRPGRIDRKIEFPVPDMKTKKKIFEIHTSKMALGEEVNF SGGEKEI QRTMLELLNQLDGFD-SQSDVKVIMATNKIESLDPALIRPGRIDRKIQLPNPDSKTKRKIFEIHTSKMTMSKDVDL SGGEKEIORTMLELLNOLDGFD SSGEREVQRTMLELLNQLDGFD SGGAK EVQRTMLELLTQLDGFD SSGAK EVQRTMLELLTQLDGFD SGGEREI QRTLLELLNQLDGFD
SGGEREVQRTMLELLNQLDGFD

## SQSDVKVIMATNKIESLDPALIRPGRIDRKIQLPNPDSKTKRKIFEIHTSKMTMSKDVDL

 DRGDIKVIMATNRIETLDPALIRPGRIDRKIELPFPDNKTKLKIFQI HTANMHLAPDVNL SCNDVKVIMATNRIETLDPALIRPGRIDRKIEFPFPDEKTKKMIFEI HT SRMSLAEDVDL SSNDVKVIMATNRI ETLDPALIRPGRIDRKI EFPFPDEKTKKMI FEI HT SRMSLAEDVDI -TRHDVKVIMATNRI ESLDPALIRPGRIDRKI EFPLPDOKTKMHI FKLHT SRMN LDSDVDL
## 840

Anopheles_gambiae/1-3220 Arabidopsis_thaliana/1-3286
Ashbya_gossypii/1-3281
A fumigatus/1-3241
Aspergillus_niger/1-3300 Bombyx_mori/1-2389
Bos_taurus/1-3273
C_briggsae/1-3231
Caenorhabditis elegans/1-3289
Candida_albicans/1-3266
Candida_glabrata/1-3282
Canis_familiaris/1-3286
Ciona intestinalis/1-2712
Cryptococcus_neoformans/1-3301 Cryptosporidium_hominis/1-3020 C_parvum/1-3281
Danio rerio/1-3286
Debaryomyces hansenii/1-3166
Dictyostelium_discoideum/1-3275 D_melanogaster/1-3287
Drosophila_pseudoobscura/1-3286
Encephalitozoon_cuniculi/1-3199
Entamoeba_histolytica/1-3271
Gallus_gallus/1-3286
Homo_sapiens/1-3286
Kluyveromyces_lactis/1-3281 Leishmania_major/1-3286 Macaca_mulatta/1-3182 Magnaporthe_grisea/1-3286 Methanosarcina_acetivorans/1-3043 Monodelphis_domestica/1-3196
Mus musculus/1-3286
Neurospora_crassa/1-3264
Oryza_sativa/1-3218
Ostreococcus_lucimarinus/1-3287
Pan_troglodyte/1-3286
Paramecium tetraurelia/1-3282
Pichia_stipitis/1-3290
_falciparum/1-3285
P_knowlesi/1-3284
Plasmodium_yoelii/1-3285
Populus_trichocarpa/1-3286
Rattus_norvegicus/1-3286
Saccharomyces_cerevisiae/1-3284
Schizosaccharomyces_pombe/1-3284
Strongylocentrotus_purpuratus/1-3252
Takifugu_rubripes/1-2930
Tetrahymena_thermophila/1-3270
T_annulata/1-3248
Theileria parva/1-3273
Trichomonas vaginalis/1-3198
T_brucei/1-3264
Trypanosoma_cruzi/1-3266
stilago maydis/1-3291
Yarrowia_lipolytica/1-3278

850
860


Anopheles_gambiae/1-3220 Arabidopsis_thaliana/1-3286
Ashbya_gossypii/1-3281
A fumigatus/1-3241
Aspergillus_niger/1-3300 Bombyx_mori/1-2389
Bos_taurus/1-3273
C_briggsae/1-3231
Caenorhabditis elegans/1-3289
Candida_albicans/1-3266
Candida_glabrata/1-3282
Canis_familiaris/1-3286
Ciona intestinalis/1-2712
Cryptococcus_neoformans/1-3301 Cryptosporidium_hominis/1-3020 C_parvum/1-3281
Danio_rerio/1-3286
Debaryomyces hansenii/1-3166
Dictyostelium_discoideum/1-3275
D_melanogaster/1-3287
Drosophila_pseudoobscura/1-3286
Encephalitozoon cuniculi/1-3199
Entamoeba_histolytica/1-3271
Gallus_gallus/1-3286
Homo_sapiens/1-3286
Kluyveromyces_lactis/1-3281
Leishmania_major/1-3286
Macaca_mulatta/1-3182
Magnaporthe_grisea/1-3286
Methanosarcina_acetivorans/1-3043
Monodelphis_domestica/1-3196
Mus musculus/1-3286
Neurospora_crassa/1-3264
Oryza_sativa/1-3218
Streococcus_lucimarinus/1-3287
Pan troglodyte/1-3286
Paramecium tetraurelia/1-3282
Pichia_stipitis/1-3290
_falciparum/1-3285
P_knowlesi/1-3284
Plasmodium_yoelii/1-3285
Populus_trichocarpa/1-3286
attus_norvegicus/1-3286
Saccharomyces_cerevisiae/1-3284
Schizosaccharomyces_pombe/1-3284
Strongylocentrotus_purpuratus/1-3252
Takifugu_rubripes/1-2930
Tetrahymena_thermophila/1-3270
T_annulata/1-3248
Theileria parva/1-3273
Trichomonas vaginalis/1-3198
__brucei/1-3264
rypanosoma_cruzi/1-3266
stilago maydis/1-3291
Yarrowia_lipolytica/1-3278
 HEYGNPHSRTHLYGWEAENAV ENARNQVAKLIEA-SPKEIVFVSGAT EANNMAVKGVMHFKKHVITTQTEHKCVLDSCRHLQQ GLYGNPHSNTHSYGWETSQEV EKARKNVADVIKA-DPKEII FTSGATESNNMALKGVARFKNHIITTRTEHKCVLEAARSMKD GIYGNPHSRTHAYGWESEKAV EQAREYIAKLIGA-DPKEIIFTSGATESNNMSIKGVARFKKHIITSQTEHKCVLDSCRHLQD GIYGNPHSRTHAYGWESEKAVEQAREHVAKLIGA-DPKEIIFTSGATESNNMSIKGVARFKKHIITTQTEHKCVLDSCRHLQD HKCVLDSCRALEG NYYGNPHSRTHAYGWESEAAMECARQQVASLIGA-DPREIIFTSGATESNNIAIKGVARFKKHLITTQTEHKCVLDSCRSLEA NDFGNPHSRTHSYGWKAEEGVEQARKYVADLIKA-DPRDIVFTSGATESNNLAIKGVAKFKNHIITLQTEHKCVLDSCRYLEN NDFGNPHSRTHSYGWKAEEGVEQAREHVANLIKA-DPRDIIFTSGATESNNLAIKGVAKFKNHIITLQTEHKCVLDSCRYLEN GLYGNPHSSTHAYGWETDK EVEKARTYIADVINA-DPKEII FT SGAT ETNNMAI KGVPRFKKHI ITTQTEHKCVLDSARHMQD GLYGNPHSNTHAYGWETNKEVETARDHVAKVIRA-DPKEIIFTSGATESNNLAIKGVGRFKKHIITTRTEHKCALEAARGMIN NYYGNPHSRTHAYGWESEAAMEHARQQVASLIGA-DPREIIFTSGATESNNIAIKGVARFKKHLITTQTEHKCVLDSCRSLEA SYYGNPHSRTHAYGWQA EDAVEVARQQVADVINA-DPREII FTSGATESNNLAVKGVGRFKKHIITTQIEHKCVLDSCRALEN EQYGNPHSRTHAYGWEAEKAVDEARQQVAQLVGA-QPKDIVFTSGATESNNMLIKGIAKFKKHIITTQTEHKCVLDSCRWLST EK FGNSHSRTHGYGWEAEEAVENARTNIANLIKC-LPKEII FTSGATESNNTII RGVCDIKNHIITTQIEHKCVLSTLRELEL EK FGNSHSRTHGYGWEAEEAVENARTNIANLIKC-LPKEII FTSGATESNNTIIRGVCDIKNHIITTQIEHKCVLSTLRELEL NYYGNPHSRTHAYGWESESAMEKARKQVAGLIGA-DPREIVFTSGATESNNMSIKGVARFKMHIITTQIEHKCVLDSCRVLET GLYGNPHSSTHSYGWETDK EV EKARKYVADVINA-DPKEIIFTSGATESNNMAVKGVPRFKKHIITTQTEHKCVLDSARHMQD ENYGNPHSKTHAYGWT SNDLVEDAR EKVSKIIGA-DSKEII FT SGATESGNIAIKGVARFKNHIITTVTEHKCI LDSCRHLEM NFYGNPHSRTHAYGWET ESAVEKAREQVATLIGA-DPKEII FTSGATESNNIAVKGVARFKRHVITTQTEHKCVLDSCRALEN NYYGNPHSRTHAYGWESETAVEKAREQVANLIGA-ETKEII FT SGATESNNIAVKGVARFKKHVVTTQTEHKCVLDSCRALEN TVFGNPHSRTHRYGWQAEAAVEKARSQVASLIGC-DPKEIIFT SGATESNNLALKGVSGFAAHIITLQTEHKCI LDTCRNLEE EI YGNPNS-LHAFGQKARKALSDSLDI IYECIGASDDDTVLITANSTEGNNTVLKTMLARRNKIIVSQIEHPSISESEKYLKE GCYGNPHSRTHAYGWESEAATERARRQVADLIGA-DPREVIFTSGATESNNMAIKGVARFKKHIITTQTEHKCVLDSCRSLEA NYYGNPHSRTHAYGWESEAAMERARQQVASLIGA-DPREIIFTSGATESNNIAIKGVARFKKHLITTQTEHKCVLDSCRSLEA GLYGNPHSNTHSYGWETNK EI EQARKYIADVIKA-DPKEIIFTSGATESNNMALKGVSRFRNHIITTRTEHKCVLEAARAMKN EEYGNPNSRTHQYGWSAEEAVEKARRQVADLIGA-SPKEIFFTSGATECNNIAIKGVGNFKNHIITLQTEHKCVLDSCRYLEM NYYGNPHSRTHAYGWESEAAMERARQQVASLIGA-DPREIIFTSGATESNNIAIKGVARFKKHLITTQTEHKCVLDSCRSLEA GVYGNPHSRTHAYGWESEKAV EDARAHVASLIGA-DPKEIIFT SGATESNNMSI KGVARFKKHIITTQTEHKCVLDSCRHLQD ENFGNP-SSIYELGKI SKHAVENARKRVADAIGA-EENEIYFTSGGTESDNWTVKGVAFAGKHIITSSIEHHAVLHACAWLEG NYYGNPHSRTHAYGWESEAAVEHARQQVASLIGA-DPREIIFTSGATESNNLAIKGVARFKKHVITTQTEHKCVLDSCRSLEA NYYGNPHSRTHAYGWESEAAMERARQQVASLIGA-DPREII FTSGATESNNIAI KGVARFKKHLVTTQTEHKCVLDSCRSLEA NVYGNPHSRTHAYGWETDKAVEEARKHIADLIGA-DPKEIIFTSGATESNNMSIKGVARFKKHIIT SQTEHKCVLDSCRHLQD SRYGNPHSRTHLYGWE SDAAVEEARARVASLVGA-DPREI FFT SGAT ECNNIAVKGVMR FRRHVVTTQTEHKCVLDSCRYLQQ EQYGNPHSRTHMYGWET EDAI EKARGELASLIGA-NAKEIVFTSGATESNNMSLKGVARFKKHIITTTTEHKCVLDSCRQLER NYYGNPHSRTHAYGWESEAAMERARQQVASLIGA-DPREII FTSGATESNNIAIKGVARFKKHLITTQTEHKCVLDSCRSLEA NQYGNPHSKTHSFGWETEKAVENARSQIANLINT-QPQSIIFTSGATESNNAALKGLYGFKNHIITTQTEHKCVLDSCRYLEE GMYGNPHSSTHAYGWETDK EV EKAR EYVAAVIKA-DPKEII FT SGAT ETNNMAIKGVPRFKKHI ITTQTEHKCVLDSARHMQD YIYGNAHSRNHFFGWESEKAVEDARTNLLNLINGKNNKEII FTSGATESNNLALIGICTYKNHIITSQIEHKCI LQTCRFLQT YIYGNAHSRNHFFGWESEEAVEDARKNI LHLINGKNNKEII FTSGATESNNLALIGICTYKNHIITSQIEHKCI LQTCRYLQT YIYGNAHSRNHFFGWESEQAVEDARANLIKLLNGNNNKEIIFTSGATESNNLALIGTCTYKNHIITSQIEHKCI LQTCRYLQT ARYGNPHSRTHLYGWESDQAVETARSQIADLIGA-SPKEIVFTSGATESNNI SVKGVIKFKRHVVTTQTEHKCVLDSCRHLQQ NYYGNPHSRTHAYGWESEAAMERARQQVASLIGA-DPREII FT SGATESNNIAIKGVARFKKHLVTTQTEHKCVLDSCRSLEA G LYGNPHSNTHSYGWETNTAVENARAHVAKMINA-DPKEII FT SGAT ESNNMV LKGVPRFKKHIITTRTEHKCVLEAARAMMK GIYGNPHSRTHAYGWEAEKAVENARQEIASVINA-DPREII FT SGATESNNAI LKGVARFKKHLVSVQT EHKCVLDSLRALQE AYYGNPHSRTHSYGWESDDAV EHARKQVANLIGA-DAREIIFT SGATESNNI SVKGTARFKKHVITTQTEHKCVLDSCRVLEG NYYGNPHSRTHAYGWESETAMETARKQVADLIGA-DPREIIFTSGATESNNMAIKGVARFKRHVITTQTEHKCVLDSCRVLES NMYGNPHSR SHEYGWATEKATEDARAQVADLIGA-DPKEITFTSGATESNNQALKGLAAFKKHIITTQIEHKCI LDTCRNLEE HA FGNPHSRTHSYGWEAEKAVETARADVANLINC-ESKNVIFTSGATESNNLAIKGSKSFKNHVITTQIEHKCVLQCCRQLEN HAFGNPHSRTHSYGWEAEKAVETARADIANLINC-ESKNVIFT SGATESNNLAIKGSK SFKNHVITTQIEHKCVLQCCRQLEN YVHGNAHSKQHGFGQEAMAAVEKARK SVADLINA-KPNEIIFTSGATECNNIAIKGAMGYKKHVIVSSIEHKCVIESARALQK EMYGNPHSRTHSYGWTAEEAVEKARTQVADLIRA-SPKGVFFT SGATESNNIAIKGVANYKNHLITLQTEHKCVLDSCRYLEM ERYGNPHSRTHRYGWTAEDAVEKARAEVADLIGT-SPKGVFFT SGATESNNIAIKGVAYYKNHIITLQTEHKCVLDSCRYLEM NQYGNPHSRTHAYGWESEKGVEEGREHIASLIGA-DPKEII FT SGATESNNMAIKGVAHFKNHIITTQTEHKCVLDSCRRLQE DMYGNPHSRTHSYGWETDTAVEKAREEIAALIGA-DPKEIIFTSGATESNNMVIKGIARFKRHIITTQTEHKCI LDSCRYLQD

Anopheles_gambiae/1-3220
Arabidopsis_thaliana/1-3286
Ashbya_gossypii/1-3281
A fumigatus/1-3241
Aspergillus_niger/1-3300 Bombyx_mori/1-2389
Bos_taurus/1-3273
C_briggsae/1-3231
Caenorhabditis elegans/1-3289
Candida_albicans/1-3266
Candida_glabrata/1-3282
Canis_familiaris/1-3286
Ciona intestinalis/1-2712
Cryptococcus_neoformans/1-3301 Cryptosporidium_hominis/1-3020 C_parvum/1-3281
Danio rerio/1-3286
Debaryomyces hansenii/1-3166
Dictyostelium_discoideum /1-3275 D_melanogaster/1-3287
Drosophila_pseudoobscura/1-3286 Encephalitozoon_cuniculi/1-3199 Entamoeba_histolytica/1-3271 Gallus_gallus/1-3286
Homo_sapiens/1-3286
Kluyveromyces_lactis/1-3281 Leishmania_major/1-3286 Macaca_mulatta/1-3182 Magnaporthe_grisea/1-3286 Methanosarcina_acetivorans/1-3043 Monodelphis_domestica/1-3196
Mus musculus/1-3286
Neurospora_crassa/1-3264
Oryza_sativa/1-3218
Ostreococcus_lucimarinus/1-3287
Pan_troglodyte/1-3286
Paramecium tetraurelia/1-3282
Pichia_stipitis/1-3290
_falciparum/1-3285
P_knowlesi/1-3284
Plasmodium_yoelii/1-3285
Populus_trichocarpa/1-3286
Rattus_norvegicus/1-3286
Saccharomyces_cerevisiae/1-3284
Schizosaccharomyces_pombe/1-3284
Strongylocentrotus_purpuratus/1-3252
Takifugu_rubripes/1-2930
Tetrahymena_thermophila/1-3270
T_annulata/1-3248
Theileria parva/1-3273
Trichomonas vaginalis/1-3198
T_brucei/1-3264
rypanosoma_cruzi/1-3266
stilago maydis/1-3291
Yarrowia_lipolytica/1-3278

EGFHVTYLPVQSNGLISMEELEKAITP-ETSLVSIMTVNNEIGVKQPIAEIGRLCVFFHTDAAQAVGKIPLDVNKMNIDLMSI EGFEVTYLPVKTDGLVDLEMLREAIRP-DTGLVSIMAVNNEIGVVQPMEEIGMICVPFHTDAAQAIGKIPVDVKKWNVALMSM EGFDVTFLNVNEDGLVSLEELEQAIRP-ETSLVSVMSVNNEIGVVQPIKEIGAICVFFHSDAAQAYGKIPIDVDEMNIDLLSI EGFEVTYLPVQNNGLIRMEDLEAAIRP-DTALVSIMAVNNEIGVIQPLEEIGKLCVFFHTDAAQAVGKIPLDVNKLNIDLMSI EGFDVTYLPVQSNGLIRMEELEAAIRP-DTALVSIMAVNNEIGVIQPMEEIGKLCIFFHTDGAQAVGKIPLDVNKLNIDLMSI EGFRITYLPVQQNGIINLKDLEDAITP-ETSLVSIMTVNNEIGVRQPIEAIGAICVFFHTDAAQAVGKVPLDVNTMNIDLMSI EGFKVTYLPVKKSGIIDLKELEAAIQP-DTSLVSVMTVNNEIGVKQPIKEIGQICVYFHTDAAQAVGKIPLDVNDMKIDLMSI EGFKVTYLPVDKGGMVDMEQLEQSITP-ETCLVSIMFVNNEIGVVQPIKQIGELCVYFHTDAAQATGKVPIDVNDLKIDLMSI EGFKVTYLPVDKGGMVDMEQLTQSITA-ETCLVSIMFVNNEIGVMQPIKQIGELCVYFHTDAAQATGKVPIDVNEMKIDLMSI EGFEVTYLPVSSEGLINLDDLKKAIRK-DTVLVSIMAVNNEIGVIQPLKEIGKICVFFHTDAAQAYGKIPIDVNEMNIDLLSI EGFDVTFLSVDNQGLIDMKELEEAIRP-DTCLVSVMAVNNEIGVMQPLKEIGALCIYFHTDAAQAYGKVPIDVNEMNIDLLSV EGFQVTY LPVKKSGIIDLKELESAIQP-DTSLV SVMTVNNEI GVKQPIAEIGQICVY FHTDAAQAVGKI PLDVNDMKIDLMSI EGFKVTYLPVK PNGIVDLKVLEESFQP-DTSLVSII FVNNEIG-
QGFEVTY LPVLPNGLVSINELKAALRP-DTSLVSIMAVNNEIGVIQPLAEISQAIPLFHTDAAQAVGKIPIDVEALGIDAMSI KGFRVTY LKVNNKGLISLEELEKSIIPGETI LASIMHVNNEIGVIQPMNLIGEICVLFHSDVAQGLGKINIDVDKWNADFLSL KG FRVTY LKVNNKGLISLEELEKSIIPGETILASIMHVNNEIGVIQPMNLIGEICVLFHSDVAQGLGKINIDVDKWNADFLSL EGFDITYLPVKSNGLIDLKQLEDTIRP-DTSLVSIMAINNEIGVKQPVKEIGHLCVFFHTDAAQAVGKIPVDVTDWKVDLMSI EGFEVTYLPVNEEGLISLDDLRKSIRK-DTSLVSIMAVNNEIGVVQPLKEIGKICIFFHTDAAQAYGKIDIDVNEMNIDLMSI EGFKVTYLPVGENGLVDLELLKNTITP-QTSLVTIMAVNNEIGVVQPIKEIGKICVFFHTDAAQAVGKIPIDVNDMNIDLLSI EGFKVTYLPVLANGLIDLQQLEETITS-ETSLVSIMTVNNEIGVRQPVDEIGKLCVFFHTDAAQAVGKVPLDVNAMNIDLMSI EG FTVTYLPVQTNGIIDLKQLEEALTP-ETSLVSIMAVNNEIGVKQPIDEIGRLCVFFHTDAAQAVGKIPMDVNAMNIDLMSI NGVEVTYLPVGNDGVVDIDDVKKSIKE-NTVLVSIGAVNSEIGTVQPLKEIGMLCVLFHTDAAQGVGKI QIDVNEMNIDLLSM RGI EVIKMPVNEDGVVDPKDLERLIDD-KTALV SCMWVNNETGLIMPVEELCKIAALFHSDATQAMGKIKVSVKDVPVDYLTF EGFQITYLPVQKNGLIDLKELEAAFQP-DTSLVSVMAVNNEIGVKQPIRDIGEICVFFHTDAAQAVGKIPLDVNDSKIDLMSI EGFQVTYLPVQKSGIIDLKELEAAIQP-DTSLVSVMTVNNEIGVKQPIAEIGRICVYFHTDAAQAVGKIPLDVNDMKIDLMSI EGYEITFLNVDEQGLINLEELEAAIRP-ETCLVSVMAVNNEIGVMQPLKEIGELCVFFHTDAAQAYGKIPIDVNEMKIDLMSI EGFEVTYLPVQKNGI LDLKVLEAAIKP-TTCLVSCMAAHNEIGVLQPIREIGALCVLFHTDAAQALGKVKVDVNADNIDLMSM EGFQVTYLPVQKSGIIDLKELEAAIQP-DTSLVSVMTVNNEIGVKQPIAEIGQICVYFHTDAAQAVGKIPLDVNDMKIDLMSI EGFEVTYLPVQNSGLVDLKELEAAMRP-ETALVSIMTVNNEIGVIQPVEEIGKMCIFFHTDAAQAVGKIPMDVNAMNIDLMSI QGFEVTYLPVDRYGMVSPEELKNAIRD-DTILISIMLANNEIGTIQPVEEIGKISIYFHTDAVQAIGHVPIDVKKMNVDLLSL EG FQVTYLPVQK SGIIDLKELEAAIQP-DTSLVSIMTVNNEIGVKQPIADIGRICVYFHTDAAQAIGKIPLNVNDMKIDLMSI EGFRVTYLPVQKSGIIDLKELEAAIQP-DTSLVSVMTVNNEIGVKQPIAEIRQICVYFHTDAAQAVGKIPLDVNDMKIDLMSI EGFEVTYLPVKSSGLIDMAELEAAIRP-DTAIVSIMAVNNEIGVIQPLEEIGKLCIFFHTDAAQAVGKI PVDVNAMNIDLMSI EGFEVTYLPVRPDGLVDVAQLADAIRP-DTGLVSVMAVNNEIGVVQPLEEIGRICVPFHTDAAQALGKIPIDVNQMGIGLMSL EGFDVTYLPVK ENGLVDLKELEAAMRD-DTAIVSVMAVNNEIGVIQPLKAIGELCIFFHTDGAQAVGKVPMDVNDMNIDLMSI EG FQVTYLPVQK SGIIDLKELEAAIQP-DTSLVSVMTVNNEIGVKQPIAEIGRICVYFHTDAAQAVGKIPLDVNDMKIDLMSI KGVEVTYLPVDSNGLISLQQLQESIKS-NTLCVSVMLVNNEIGVIQNLKEISRICVYVHSDMAQAIAKIPVDVQDLDIDLGSI EG FDVTY LPVDEHGLISLDDLKAAIRK -DTI LV SVMAVNNEIGVVQPLKEIGKICIFFHTDAAQAYGKIDIDVNDMNIDLLSI KGFEVTYLKPDTNGLVKLDDIKNSIKD-NTIMASFIFVNNEIGVIQDIENIGNLCI LFHTDASQAAGKVPIDVQKMNIDLMSM KGFEVTY LKPEPNGIVKLEDIEKNIKE-NTIMASFIHVNNEIGVIQDIENIGLLCVIFHTDASQAIGKIPIDVQKMNIDLLSM KGFEVTYLKPDANGLIKLEDLKNSIKE-NTILASFIYVNNEIGVIQDIENIGKICIIFHTDASQAVGKIKIDVQKLNIDLLSL EGFEVTYLPVGNDGIVDLEKLKGSIRP-DTGLV SVMAVNNEIGVIQPMEEIGEICVPFHTDAAQALGKIPIDVDKWNVSLMSL EGFRVTYLPVQKSGIIDLKELEAAI QP-DTSLVSVMTVNNEIGVKQPIAEIGQICLYFHTDAAQAVGKIPLDVNDMKIDLMSI EGFEVTFLNVDDQGLIDLKELEDAIRP-DTCLV SVMAVNNEIGVIQPIKEIGAICIYFHTDAAQAYGKIHIDVNEMNIDLLSI EGFEVTFLPVQTNGLINLDELRDAIRP-DTVCVSVMAVNNEIGVCQPLEEIGKICVFFHSDAAQGYGKIDIDVNRMNIDLMSI EGFDITYLPVKPNGIIDLKELEAAFRP-DTVLCSIMAINNEIGVKQPMKQIGEMCVFFHTDAAXAVGKIPVDVNDMKIDLMSI EGFSVTYLPVQKNGLVDLELLEASIRP-DTSLLSVMTVNNEIGVQQPIDEIGRICVFLHTDAAQAVGKIPINVSDWKVDLMSI QGYEITYLPVQKNGLVDLEVFKNAIRP-DTLVASIILVHNEIGVIQDIKTIGKICVFFHTDAAQALGKIPINVDEMNIDLMSM EGY SVTY LK PDKYGMI LPDLVRKNIRP-ETFLCSVIHVNNEIGVIQNISEIGRICVIFHTDAAQSFGKLPIDLKNLDVDLLSI EGYSVTYLKPDKYGMI LPEEVRKNIRP-ETFLCSVIHVNNEIGVIQDIAEIGKVCVIFHTDAAQSFGKLPIDLKNLEVDLLSI EGFDATFLQVGKDGRVDPK EVAKNIRP-DTGLVSCMLVNNEIGSINPVQEISKICVWFHTDAAQGFGKIPIDVKKIGANFMSI EGFEVTYLPVEKNGIVNLQKLEEAIRP-TTALVSCMYVNNEIGVIQPIGEIGKICVLFHTDAAQAVGKLDIDVDRDNIDLMSV DGFEVTYLPVEKNGLVNLQKIEEAIRP-TTALVSCMYVHNEIGVIQPISEIGNLCVLFHTDAAQALGKVSIDVERDNIDLMSL EGFEVTYLPVQSNGLIDLKQLEEALRP-TTALVSIMTVNNEIGVIQPIKEIGQLLPFFHTDAAQAAGKIRLDVNELGIDLMSL EGFEVTYLPVLSSGLIDMKQLEAAIRP-DTALVSIMAVNNEIGVIQPIAEIGALCVFFHTDAAQAVGKIPIDVNADKIDVMSI

1090
Anopheles_gambiae/1-3220
Arabidopsis_thaliana/1-3286
Ashbya_gossypii/1-3281
A fumigatus/1-3241
Aspergillus_niger/1-3300 Bombyx_mori/1-2389
Bos_taurus/1-3273
C_briggsae/1-3231
Caenorhabditis_elegans/1-3289
Candida_albicans/1-3266
Candida_glabrata/1-3282
Canis_familiaris/1-3286
Ciona intestinalis/1-2712
Cryptococcus_neoformans/1-3301 Cryptosporidium_hominis/1-3020 C_parvum/1-3281
Danio_rerio/1-3286
Debaryomyces_hansenii/1-3166
Dictyostelium_discoideum /1-3275 D_melanogaster/1-3287
Drosophila_pseudoobscura/1-3286 Encephalitozoon_cuniculi/1-3199 Entamoeba_histolytica/1-3271 Gallus_gallus/1-3286
Homo_sapiens/1-3286
Kluyveromyces_lactis/1-3281 Leishmania_major/1-3286 Macaca_mulatta/1-3182 Magnaporthe_grisea/1-3286 Methanosarcina_acetivorans/1-3043 Monodelphis_domestica/1-3196
Mus musculus/1-3286
Neurospora_crassa/1-3264
Oryza_sativa/1-3218
Ostreococcus_lucimarinus/1-3287
Pan troglodyte/1-3286
Paramecium_tetraurelia/1-3282
Pichia_stipitis/1-3290
_falciparum/1-3285
P_knowlesi/1-3284
Plasmodium_yoelii/1-3285
Populus_trichocarpa/1-3286
Rattus_norvegicus/1-3286
Saccharomyces_cerevisiae/1-3284
Schizosaccharomyces_pombe/1-3284 Strongylocentrotus_purpuratus/1-3252 Takifugu_rubripes/1-2930 etrahymena_thermophila/1-3270
T_annulata/1-3248
Theileria parva/1-3273
Trichomonas vaginalis/1-3198
T_brucei/1-3264
rypanosoma_cruzi/1-3266
stilago_maydis/1-3291
Yarrowia_lipolytica/1-3278

SGHK I YGPKGIGALYV SAHK I YGPKGVGA LYV SHK IYGPKGI GACYV SSHKIYGPKGMGACYV SGHKVYGPKGVGA LYI SGHK I YGPKGVGAI YI SGHK I YG PKGAGA LYV SAHK I YGPKGAGA LYV S SHK I YGPKGI GACYV SSHK I YGPKGI GALYV -

SGHK LYG PKGVGAAYV SAHKVYGPKGI GAFYI SAHKVYGPKGI GAFYI SAHK I YGPKGVGALFV S SHK I YGPKGI GACYV SGHK I YGPKGVGALFV SGHK I YG PKGVGA LYV SGHK I YGPKGVGA LYV CAHK I YGPKGI GA LYV TAHK FHG PKGVGALFI SGHK I YGPKGVGA I YV SGHK I YGPKGVGA I Y I S SHK I YG PKGI GA I YV S SHKVYGPKGCGA LYV SGHK I YGPKGVGA I Y I SGHK I YGPKGI GACYV SGHK FGGPKGCGA LYI SGHK LYG PKGVGA I Y I SGHK LYGPKGVGA I Y I S SHK I YGPKGI GACYV SAHK I YGPKGVGA LY LSGHK FYG PKGI GA LYV SGHK I YGPKGVGA I Y I SAHK LYG PKGI GA LYV S SHK I YG PMGI GACYV SGHK LYGPKGI GALYIK SGHK LYGPKGIGALYIK S SHK LYGPKGVGALYIK SGHK I YGPKGVGA LYMSGHK LYG PKGVGAIYI SSHK I YGPKGIGAIYVSAHK I YGPKGI GAAYV SGHK I YGPKGI GA LYV SGHK I YG PKGVGA LYV S SHKVYGPKGI GGLYV SGHK I YG PKGVGALFV SGHK I YGPKGVGALFV SGHK I HG PKGI GALYV S SHK I YG PKGCGA LYM S SHK I YG PKGCGA LYM SGHK LYG PMGI GACYV

1110
1130

1100
1120
 RPRIRLEPLMNGGGQERGLR SGTGATQQIVGFGAACELAMK EMEYDEKWIKGLQER LNGS --M RPRVRMEPLLSGGGQERGFR SGT LPPPLVVGLGHAAK LMV EEY EYDSAHVRRLSDRLNGS--A RPRVRLEPII SGGGQERGLR SGT LAPHLVVGFGEACRIASQDMEYDRKHVERLSKRLNGD--A RPRVRLEPII SGGGQERGLR SGTIAPHLVVGFGEACRIAY EDMEYDSKHIARLSKRLNGD--P PRVRVEPI QSGGGQERGMR SGTVPTPLVVGL- $\qquad$
PRVVEALQSGGGQERGMR SGTV PT PLVVGLGACEVAQQEMEYDHKRISKLAERLNGD--P PRVRV EAQMSGGGQERGLR SGTVAAPLCIGLGEAARIAGR EMEMDKAHV ER LSRMLNVD--E RPRVRI EAQMSGGGQERGLRSGTVAAPLCIGLGEAAKIADK EMAMDKAHVER LSQMLNGD--A R PRVR LDPIITGGGQERGLR SGTLAPPLVAGFGEAAR LMK QESSFDKRHIEKLSSKLNGCNDA R PRVR LEPLLSGGGQERGLRSGTLAPPLVAGFGEAAR LMHEEYNADIAHIDKLSSKLNGS--A R PRVRVEA LQSGGGQERGMR SGTV PT PLVVGLGAAC EVAQQEMEYDHKRISKLADR LNGD--P

RPRVRLEPLIHGGGERGLR SGTVAAPLVVGLGEACRIAENEMAADHARIKALSDRLNGD---
KPRRRIKPLIFGGGQERGMRSGTMPVPLAVGFGEACKIASSEMNSDSIHVKSLYDKLNGC PRRRRIKPLIFGGGQERGMR SGTMPVPLAVG FG EACKIA S S EMN SDSI HVK S LYDK LNGC -PRVR LEPLQSGGGQERGLR SGTV PT PLAVGLGACEIAQQELEYDHKRVSLLANR LNGD--P PRVRLDPIVtGGGQerg Lr SGt lap plvag fa eas lmk eemd -
PRVRIEPIITGGGQERGIR SGTVPSTLAVGLGACDIALK EMNHDAAWVKY LYDRLNGD--L PRVRLEPI QSGGGQERGLR SGTVPAPLAVGLGAAAELSLREMDYDKKWVDFLSNRLNGD--A PRVRLEPI QSGGGQERGLRSGTVPASLAVGLGAAELSQQEMEYDKKWIDFLSNR LNGD--A PRVRMVPLINGGGQERGLR SGTVASPLVVGFGKAAEICSKEMKRDFEHIKELSKKLNGS---KP---ITPLLHGGEQMGGLRSGTIDTPSVVGMAVALKKATHDINI ENTYVRKLRDKLVGK--P PRVR LEPLQSGGGQERGLR SGTVPT PLAVGLGAACEVAQEEMEYDHKRI SQLAER LNGD--R PPRVRVEALQSGGGQERGMR SGTVPTPLVVGLGAACEVAQQEMEYDHKRISKLSERLNGD--P KPRVR LDPLI SGGGQERGLR SGT LAPPLVAGFGEAAR LMMK EY END SNHIKR LSDK LNGS--A PRVRLRSPVSGGGQERGVR SGTVAAALVVGMGAAC EVAMK EWKRDAAHTER LQER LNGD--L R PRVRVEALQSGGGQERGMR SGTVPTPLVVGLGAACEVAQQEMEYDHKRISKLAERLNGD--P RPRVR LDPIISGGGQERGLRSGTLAPPLIVGFGEACRIAKQEMEYDSKRVKY LSDR LNGH--P -GTKI EAFLHGGAQERKRRAGTENVPSIVGLGKAIGLATGEMEETNKPLLEMRERLNGH--P R PRVRVEALQSGGGQERGMR SGTV PTPLVVGLGAACEVAQEEMENDHKRI SMLAER LNGD--P RPRVRVEALQSGGGQERGMR SGTVPTPLVVGLGAACELAQQEMEYDHKRISKLAERLNGD--P RPRVR LDPII SGGGQERGLR SGT LAPPLVVGFGEACRIAKEEMPYDSKRIKHLSDR LNGD--P RPRIRVEPQMSGGGQERGIR SGTVPTPLVVGFGAACEIAAK EMDYDHRRASV LQQR LNGS--M RPRVRMEPI INGGGQERGLR SGTLPTPLIVGIGEAARVAQK ELQRDEEHVNR LAKR LNGD--R R PRVRVEALQSGGGQERGMR SGTV PT PLVVGLGAACEVAQQEMEYDHKRISKLSERLNGD--P KPRVR LQQI I HGGGQERGLR SGT LAPHLCVGFGKAAEIALTELPYDI QHVDKLYNRLNGS--L R PRVR LDPI ITGGGQERGLR SGT LSPP LVAGFGEAAR LMK EEMDYDKAHITR LSNKLNG SNNP KPNIR LNALI HGGGQERGLR SGTLPTHLIVGFGEAAKVCSLEMNRDEKKVRYFFNYVNGC--Q KPNLR LNALI HGGGQERGLR SGT LPTHLIVGLGEAANLGSI EMNRDHKKMK FFFDYVNGC--Q KPNIRLNAII HGGGQERGLR SGTLPTHLIVGLGEAANICLSEMDRDNKKMNFFFNYVNGC--Q -RPRIRVEPQMNGGGQERGIRSGTVPTPLVVGMGAACELAKK EMEYDDKRIRALHERMNGS--V R PRVRV EA LQSGGGQERGMR SGTV PT PLVVGLGAACELAQQEMEYDHKRISK LAER LNGD--P PRRVRLEPLLSGGGQERGLRSGT LAPPLVAGFGEAAR LMKKEFDNDQAHIKRLSDK LNGS --P R PRVRLEPLI SGGGQERGLR SGT LAPSQVVGFGTAARICK EEMKYDYAHISKLSQRLNGD--P PRVRV EALQSGGGQERGMR SGTLPAPLVVGLGAACEVSQQEMEYDHKRISALSERLNGD--P RPRVR LEPLQSGGGQERGLR SGTVPT PLAVGLGAAC SVAQQEI EYDHQRVSMLANR LNGD--P PRKVRI LPI INGGGQERGLR SGT LAPHLCVGFGEACEIAKR EMDNDKKHI QR LSEKFNGD--K KPRIRLQPIIDGGGQERGLRSGTLPTALVVGLGTAAKIAKMEMKRDQLHMENLFFKLNGSIKP KPRIRLQPIIDGGGQERGLR SGTLPTALVVGLGTAAKIAKMEMERDHRHMENLFFKLNGSIKP RPR SRVEPI INGGGQERNIR SGTLAVPLIVGLGKAAEIAKREMKYDSPYIESLGKHLNGS--L R PRVRVR SPVSGGGQERGVR SGTIATPLAVGLGAACELAKVEMKRDSERIAQLSKR LNGD--V RPRVRVRSPV SGGGQERGVR SGTVATAQVVGMGAACAI AKV EMERD SAHI SR LSKR LNGD--L R PRVR LEP I INGGGQERG LR SGT LAPPLIAG FG EAAR LAKQELAYDHAHI SK LSQR LNGD---

Anopheles_gambiae/1-3220 Arabidopsis_thaliana/1-3286
Ashbya_gossypii/1-3281
A fumigatus/1-3241
Aspergillus_niger/1-3300 Bombyx_mori/1-2389
Bos_taurus/1-3273
C_briggsae/1-3231
Caenorhabditis elegans/1-3289
Candida_albicans/1-3266
Candida_glabrata/1-3282
Canis_familiaris/1-3286
iona intestinalis/1-2712
Cryptococcus_neoformans/1-3301 Cryptosporidium_hominis/1-3020 C_parvum/1-3281
Danio rerio/1-3286
Debaryomyces hansenii/1-3166
Dictyostelium_discoideum /1-3275
D_melanogaster/1-3287
Drosophila_pseudoobscura/1-3286
Encephalitozoon cuniculi/1-3199
Entamoeba_histolytica/1-3271
Gallus_gallus/1-3286
Homo_sapiens/1-3286
Kluyveromyces_lactis/1-3281
Leishmania_major/1-3286
Macaca mulatta/1-3182
Magnaporthe_grisea/1-3286
Methanosarcina_acetivorans/1-3043
Monodelphis_domestica/1-3196
Mus musculus/1-3286
Neurospora_crassa/1-3264
Oryza_sativa/1-3218
Streococcus_lucimarinus/1-3287
Pan troglodyte/1-3286
Paramecium tetraurelia/1-3282
Pichia_stipitis/1-3290
_falciparum/1-3285
P_knowlesi/1-3284
Plasmodium_yoelii/1-3285
Populus_trichocarpa/1-3286
Rattus_norvegicus/1-3286
Saccharomyces_cerevisiae/1-3284
Schizosaccharomyces_pombe/1-3284
Strongylocentrotus purpuratus/1-3252
Takifugu_rubripes/1-2930
etrahymena_thermophila/1-3270
T_annulata/1-3248
heileria parva/1-3273
Trichomonas vaginalis/1-3198
T_brucei/1-3264
rypanosoma_cruzi/1-3266
Ustilago maydis/1-3291
Yarrowia_lipolytica/1-3278

VQ'SYPGCiNLSFA'YVEGESLLMA'L-
1210 DHRY PGCVNV SFAFVEGESLLMAL--RDIALSSGSACTSASLEPSYVLHAIGRDDALAHSSIRFGIGRFTTEAEVDYVIKAIT ERHYPGCVNV SFAYI EGESLLMAL--KDIALSSGSACT SASLEPSYVLRALGSSDESAHSSIRFGIGRFTTDSEIDYVLKAVQ DRHYPGCVNI SFAYI EGESLLMAL--KDIALSSGSACT SASLEPSYVLRALGSSDESAHSSIRFGIGRFTTDSEIDYVLKAVQ EHHY PGC KHAY PGCVNLSFAYVEGESLMALI-K RHAY PGCVNLSFAYVEGESLLMAL--KSIALSSGSACT SASLEPSYVLRAIGSEEDLAHSSIRFGLGRFTTDEEVKHTIDLCI K SQY PGCVNV SFAYI EGESLLMAL--KDIALSSGSACT SASLEPSYVLHALGADDALAHSSIRFGIGRFTTEAEVDYVIQAIN EKRYPGCVNVSFAYVEGESLLMAL--RDIALSSGSACTSASLEPSYVLHALGKDDALAHSSIRFGIGRFTTEEEVDYVLKAIT EHHYPGCINLSFAYVEGESLLMAL--KDVALSSGSACT SASLEPSYVLRAIGTDEDLAHSSIRFGIGRFTTEEEVDYTVEKCI

VNGYPGCVNLSFSYVEGESLLMAL--KDIALSSGSACT SASLEPSYVLRALGAAEDMAHSSLRFGIGRFTTEEEIDLVVQRIV VNRMFGNLNLSFTGVEGESLMMK L--YSLALSSGSACT SASLEPSYVLRAIGVGEDVAHTSIRFGLGRFTKHEDVDKAVKEIV VNRMFGNLNLSFTGVEGESLMMK L--Y SLALSSGSACT SASLEPSYVLRAIGVGEDVAHTSIRFGLGRFTKHEDVDKAVKEIV DQRYPGCINLSFAYVEGESLLMAL--KDVALSSGSACT SASLEPSYVLRAIGADEDLAHSSIRFGIGRFTTEEEVDYTAEKCI NARYYGNLNI SFSYVEGESLLMAI--KDVACSSGSACTSSSLEPSYVLRSLGVEEDMAHSSIRFGIGRFTTEQEIDYTIEILK KATYNGCLNLSFAYVEGESLLMAL--KDVALSSGSACTSASLEPSYVLRAIGTDEDLAHSSIRFGIGRFTTVEEVDYTADKCI VATYNGCLNLSFAYVEGESLLMAL--KDVALSSGSACT SASLEPSYVLRAIGADEDLAHSSIRFGLGRFTTVEEVDYTADKCI EKGFPGCVNVSFPFVEGESLLMHL--KDIALSSGSACT SASLEPSYVLRALGRDDELAHSSIRFGIGRFTMAKEIDIVANKTV ELRVPNTI LVAFKGVEGEAMLWDLNKHGIAASTGSACASESLQANPTFKAMKFGEDLSHTGIRLSLSRFNTEEEIDYTIDIIK EHRYPGCINLSFAYVEGESLLMAL--KDVALSSGSACTSASLEPSYVLRAIGADEDLAHSSIRFGIGRFTTEEEIDYTVQKCI KHHY PGCINLSFAYVEGESLLMAL--KDVALSSGSACTSASLEPSYVLRAIGTDEDLAHSSIRFGIGRFTTEEEVDYTVEKCI DHRYPGCVNI SFAYVEGESLLMAL--RDIALSSGSACT SASLEPSYVLHALGKDDALAHSSIRFGIGRFTTDEEIDYVIKAIT KHR LPGNLNI SFSCVEGESLLMGM--RDVAVSSGSACT SASLEPSYVLRALGVDAENAHTSIRFGIGRFTTAKEVDLVI EECV EHHYPGCINLSFAYVEGESLLMAL--KDVALSSGSACT SASLEPSYVLRAIGTDEDLAHSSIRFGIGRFTTEEEVDYTVEKCI DH FY PGCVNV SFAYVEGESLLMAL--KDIALSSGSACT SASLEPSYVLRALGNSDESAHSSIRFGIGRFTTEREIDYVLKAVQ T ER LANNVNVTFEYI EGESLLLLLNAKGI FASTGSACNSTSLEPSHVLTACGVPHEIVHGSLRLSLGRMNTLEDVDRVLEVLP QQHYPGCINLSFAYVEGESLLMAL--KDVALSSGSACTSASLEPSYVLRAIGTDEDLAHSSIRFGIGRFTTEEEVDYTAEKCI KQHYPGCINLSFAYVEGESLLMAL--KDVALSSGSACT SASLEPSYVFRAIGTDEDLAHSSIRFGIGRFTTEEEVDYTAEKCI NH FY PGCVNV SFAYVEGESLLMAL--KDIALSSGSACT SASLEPSYVLRALGNSDESAHSSIRFGIGRFTTEQEIDYVLKAVT EHRYPGNLNLSFAYVEGESLLMGL--K EVAVSSGSACTSASLEPSYVLRALGVEEDMAHTSIRFGIGRFTTEEEVDRAI ELTV EARYHGNVNMS FAYVEGESMLMGL--KEIAVSSGSACT SASLEPSYVLRALGVNEEMAHT SVRYGLGRFTTEAEVDRAI EATV KHHY PGCINLSFAYVEGESLLMAL--KDVALSSGSACT SASLEPSYVLRAIGTDEDLAHSSIRFGIGRFTTEEEVDYTVEKCI EHRYKGNLNVSFAFVEGESLIMAI--KQVAVSSGSACTSASLEPSYVLRALGVQEDMAHTSLRIGIGRFTTEKEVDFLLDQLS ESQYPGCVNI SFAYI EGESLLMAL--KDIALSSGSACT SASLEPSYVLHALGADDALAHSSIRFGIGRFTTEEEVDYVIKAIN I NRYYGNMNI SFLFVEGESLLMSL--NEIALSSGSACTSSTLEPSYVLRSIGI SEDIAHTSIRIGFNRFTTFFEVQQLCINLV TNRYFGNMNVSFLFVEGESLLMSL--NEIALSSGSACTSSTLEPSYVLRSIGI SEDIAHTSIRIGFNRFTTFFEVQQLCENLV I NRYFGNMNI SFLFVEGESLLMSL--NDIALSSGSACTSSTLEPSYVLRSIGITEEIAHTSIRIGFNRFTTFFEVQQLCKNLV ERRYAGNLNLSFAYVEGESLLMGL--KDVAVSSGSACT SASLEPSYVLRALGVDEDMAHTSIRFGIGRFTTEEEIDRAIELTV KQHYPGCINLSFAYVEGESLLMAL--KDVALSSGSACT SASLEPSYVLRAIGTDEDLAHSSIRFGIGRFTTEEEVDYTVQKCI DHRY PGCVNV SA FVEGESLLMAL--RDIALSSGSACT SASLEPSYVLHALGKDDALAHSSIRFGIGRFSTEEEVDYVVKAVS K SRY PGCVNI SFNYVEGESLLMGL--KNIALSSGSACTSASLEPSYVLRAIGQSDENAHSSIRFGIGRFTTEAEIDYAI ENVS DETYPGCVNLSFAYVEGESLLMAL--KDVALSSGSACT SASLEPSYVLRAIGAQEDLAHSSIRFGI SRFTTEEEVDYTAEKCV NQRY PGCVNLSFAYVEGESLLMAL--KDVALSSGSACT SASLEPSYVLRAIGADEDLAHSSIRFGIGRHHRRRSGLHGRKMYL DQRYVGNINI SFEFVEGESLMMGI--KQCAVSSGSACTSASLEPSYVLRALGVNEELAHTSLRIGFGRFTTDEEVDYLINLLS GERYFGNLNMSFEFI EGESLLMSL--SNFALSSGSACT SASLEPSYVLRSLDVSEELAHTSIRFGLGRFTMESEVDMALESIT GQRYFGNLNMSFEFI EGESLLMSL--SNFALSSGSACT SASLEPSYVLRSLDVSEELAHTSIRFGMGRFTIESEVDMALDSIT EHRWFGCVNI SFEAVEGESLMATI--PNFGVSSGSACTSASLEPSYVLKGIGVGDELAHTSLRIGISKFTTREEVDQFVELLE ERR FHGNLNI SFACVEGESLLMGM--KKVAVSSGSACT SASLEPSYVLRALGIDAENAHTSIRFGIGRFTTEREVDVTVEECA EKRYPGNLNI SFSCVEGESLLMGM--KNVAVSSGSACT SASLEPSYVLRALGIDAENAHTSIRFGIGRFTTEREIDVTIEECV QNGY PGCLNLTFQYVEGESLLMALEHHYPGCVNI SFAYVEGESLLMAL

Anopheles_gambiae/1-3220
Arabidopsis_thaliana/1-3286
Ashbya_gossypii/1-3281
A fumigatus/1-3241
Aspergillus_niger/1-3300
Bombyx_mori/1-2389
Bos_taurus/1-3273
C_briggsae/1-3231
Caenorhabditis_elegans/1-3289
Candida_albicans/1-3266
Candida_glabrata/1-3282
Canis_familiaris/1-3286
Ciona intestinalis/1-2712
Cryptococcus_neoformans/1-3301 Cryptosporidium_hominis/1-3020 C_parvum/1-3281
Danio_rerio/1-3286
Debaryomyces hansenii/1-3166
Dictyostelium_discoideum /1-3275 D_melanogaster/1-3287
Drosophila_pseudoobscura/1-3286 Encephalitozoon cuniculi/1-3199 Entamoeba_histolytica/1-3271 Gallus_gallus/1-3286
Homo_sapiens/1-3286
Kluyveromyces_lactis/1-3281
Leishmania_major/1-3286
Macaca mulatta/1-3182
Magnaporthe_grisea/1-3286
Methanosarcina_acetivorans/1-3043
Monodelphis_domestica/1-3196
Mus musculus/1-3286
Neurospora_crassa/1-3264
Oryza_sativa/1-3218
Ostreococcus_lucimarinus/1-3287
Pan troglodyte/1-3286
Paramecium_tetraurelia/1-3282
Pichia_stipitis/1-3290
_falciparum/1-3285
P_knowlesi/1-3284
Plasmodium_yoelii/1-3285
Populus_trichocarpa/1-3286
Rattus_norvegicus/1-3286
Saccharomyces_cerevisiae/1-3284
Schizosaccharomyces_pombe/1-3284
Strongylocentrotus purpuratus/1-3252
Takifugu_rubripes/1-2930
Tetrahymena_thermophila/1-3270
T_annulata/1-3248
heileria parva/1-3273
Trichomonas_vaginalis/1-3198
T_brucei/1-3264
rypanosoma_cruzi/1-3266
stilago maydis/1-3291
Yarrowia_lipolytica/1-3278


## 1330

1340

Anopheles_gambiae/1-3220
Arabidopsis_thaliana/1-3286
Ashbya_gossypii/1-3281
A fumigatus/1-3241
Aspergillus_niger/1-3300 Bombyx_mori/1-2389
Bos_taurus/1-3273
C briggsae/1-3231
Caenorhabditis_elegans/1-3289
Candida_albicans/1-3266
Candida_glabrata/1-3282
Canis_familiaris/1-3286
iona intestinalis/1-2712
Cryptococcus_neoformans/1-3301 Cryptosporidium_hominis/1-3020 C_parvum/1-3281
Danio_rerio/1-3286
Debaryomyces hansenii/1-3166
Dictyostelium_discoideum /1-3275 D_melanogaster/1-3287
Drosophila_pseudoobscura/1-3286 Encephalitozoon_cuniculi/1-3199 Entamoeba_histolytica/1-3271 Gallus_gallus/1-3286
Homo_sapiens/1-3286
Kluyveromyces_lactis/1-3281 Leishmania_major/1-3286 Macaca_mulatta/1-3182 Magnaporthe_grisea/1-3286 Methanosarcina_acetivorans/1-3043 Monodelphis_domestica/1-3196
Mus musculus/1-3286
Neurospora_crassa/1-3264
Oryza_sativa/1-3218
Ostreococcus_lucimarinus/1-3287
Pan troglodyte/1-3286
Paramecium tetraurelia/1-3282
Pichia_stipitis/1-3290
_falciparum/1-3285
P_knowlesi/1-3284
Plasmodium_yoelii/1-3285
Populus_trichocarpa/1-3286
Rattus_norvegicus/1-3286
Saccharomyces_cerevisiae/1-3284
Schizosaccharomyces_pombe/1-3284
Strongylocentrotus_purpuratus/1-3252
Takifugu_rubripes/1-2930
etrahymena_thermophila/1-3270
T_annulata/1-3248
Theileria parva/1-3273
Trichomonas_vaginalis/1-3198
T_brucei/1-3264
Trypanosoma_cruzi/1-3266
Ustilago_maydis/1-3291
Yarrowia_lipolytica/1-3278

GVGDMFVATVKKGKPELRKKVMPAVVIRQRKPFRRRDGVFLYFEDNAGVIVNNKGEMKGSAITGPVAKECADLWPRIA---SN CVGDMVMATVKKGK PDLRKKVLPAVIVRQRKPWRRKDGVFMYFEDNAGVIVNPKGEMKGSAITGPIGKECADLWPRIA---SA S LGDMVMATVKKGKPELRKKVMPAIVVRQSKPWRRKDGVYLYFEDNAGVIANPKGEMKGSAITGPVGKECADLWPRIA---SN GVGDMVMATVKKGKPELRKKVMPAVVVRQSK PWRR PDGIY LYFEDNAGVIVNAKGEMKGSAITGPVGKEAAELWPVSSLLFSN GVGDMVMATVKKGK PELRKKVMPAVVVRQSK PWRR PDGIYLYFEDNAGVIVNAKGEMKGSAITGPVGKEAAELWPRIA---SN GSGDMI VATVKKGKPELRKKVMPAVVIRQRKPFRRRDGVFIYFEDNAGVIVNNKGEMKGSAITGPVAKECADLWPRIA---SN GVGDMVMATVKKGK PELRKKVHPAVVIRQRK SYRRKDGVFLYFEDNAGVIVNNKGEMKGSAITGPVAKECADLWPRIA---SN GVGDMFVCSVKKGKPELRKKVLQGVVIRQRKQFRRKDGTFIYFEDNAGVIVNNKGEMKGSAITGPVAKECADLWPRIA---AN GVGDMFVCSVKKGKPELRKKVLQGVVIRQRKQFRRKDGTFIYFEDNAGVIVNNKGEMKGSAITGPVAKECADLWPRIA---AN AAGDMVMATVKKGKPELRKKVMPAIVIRQSKPWRRRDGVYLYFEDNAGVIVNPKGEMKGSAITGPVAKECADLWPRIA---SN SLGDMVMATVKKGKPELRKKVMPAIVVRQSKAWRRKDGVYLYFEDNAGVIANPKGEMKGSAITGPVGKECADLWPRVA---SN GVGDMV MATVKKGK PX LRKKVHPAVVIRQRK SYRRKDGVFLYFEDNAGVIVNNKGEMKGSAITGPVXKECADLWPXIA---SN GVGDIVLATVKKGKPELRKKVHPAVIIRQSK SYRRKHGQMI Y FEDNAGVIVNQKGEMKG-
AAGDMVMA SVKKGKPELRKKVMPAVICRQRKPWRRRDGIFLYFEDNAGVIVNAKGEMKGSAINGPVAKECADLWPRIA---SN SIGDMV LATVKKGKPELRKKVWPAVIVRQRKAFRRPEGTFLYFEDNAGVIVNPKGEMKGSAITGPVGKECAELWPKVS---AA SIGDMV LATVKKGKPELRKKVWPAVIVRQRKAFRRPEGTFLYFEDNAGVIVNPKGEMKGSAITGPVGKECAELWPKVS---AA GVGDMVMATVKKGKPELRKKVHPAVVIRQRKSYRRKDGVFLYFEDNAGVIVNNKGEMKGSAITGPVAKECADLWPRIA---SN SAGDMVMATVKKGKPELRKKIMPAIVVRQAR PWRRKDGVYLYFEDNAGVIVNPKGEMKGSAITGPVAKECADLWPRIA---SN GVGDMVMATVKKGKPELRKKVCTGLVVRQRKHWKRKDGVYIYFEDNAGVMCNPKGEVKGN-I LGPVAKECSDLWPKVA---TN GVGDMFVATVKKGKPELRKKVMPAVVIRQRKPFRRRDGVFIYFEDNAGVIVNNKGEMKGSAITGPVAKECADLWPRIA---SN GVGDMFVATVKKGKPELRKKVMPAVVIRQRKPFRRRDGVFIYFEDNAGVIVNNKGEMKGSAITGPVAKECADLWPRIA---SN APGDICVVSVKKGKPELRKKVHYAI LIRQKKI WRRTDGSHIMFEDNAAVLINNKGELRGAQIAGPVPREVADMWPKIS---SQ GCGDMVVATCKKGKPEYRKKMHTAVI I RQRRTWRRKDGVTLYFEDNAAVIVNMKGEMKGSAITGPVSKESADLWPKIS---SN GVGDMVMATVKKGKPELRKKVHPAVVIRQRKSYRRKDGVFLYFEDNAGVIVNNKGEMKGSAITGPVAKECADLWPRIA---SN GVGDMV MATVKKGK PELRKKVHPAVVIRQRK SYRRKDGVFLYFEDNAGVIVNNKGEMKGSAITGPVAKECADLWPRIA---SN SLGDMVMATVKKGKPELRKKVMPAIVVRQSKAWRRKDGVFLYFEDNAGVIANPKGEMKGSAVTGPVGKECADLWPRIA---SN ALGDMVMC SVKKGKPELRKKVLNAVIIRQRK SWRRKDGTVIYFEDNAGVIVNPKGEMKGSGIAGPVAKESADLWPKIS---TH GVGDMVMATVKKGKPELRKKVHPAVVIRQRK SYRRKDGVFLYFEDNAGVIVNNKGEMKGSAITGPVAKECADLWPRIA---SN GVGDMV MATVKKGKPELRKKVHPAVIVRQSK PWKRTDGVFLYFEDNAGVIVNPKGEMKGSAITGPVGKEAAELWPRIA---SN GIGDMCVV SVKKGTPEMRKQVLLAVVVRQKQEFRR PDGLHVSFEDNAMVITDEEGI PKGTDIKGPVAREVAERFPKIG---TT GVGDMVMATVKKGKPELRKKVHPAVVIRQRK SYRRKDGVFLYFEDNAGVIVNNKGEMKGSAITGPVAKECADLWPRIA---SN GVGDMVMATVKKGKPELRKKVHPAVVIRQRK SYRRKDGVFLYFEDNAGVIVNNKGEMKGSAITGPVAKECADLWPRIA---SN GVGDMVMATVKKGKPELRKKVHPAVIVRQSK PWKR FDGVFLYFEDNAGVIVNPKGEMKGSAITGPVGKEAAELWPRIA---SN CVGDMVMATVKKGKPDLRKKVMPAVIVRQRKPWRRKDGVYMY FEDNAGVIVNPKGEMKGSAITGPIGKECADLWPRIA---SA NPG SMV MATVKKGKPDLRKKVFPAI IVRQRKPIRRKEGLI IYFEDNAGVICNPKGEMKGSAIAGPVAKECADLWPRVA---SA GVGDMVMATVKKGKPELRKKVHPAVVIRQRKSYRRKDGVFLYFEDNAGVIVNNKGEMKGSAITGPVAKECADLWPRIA---SN SI GDMV LC SVK QGK PA LRKKVMQAVVVRQRK PYRRREGYYIYFEDNAGVIINPKGEMKGSAITGPVGKEAADLWPKIA---SA SAGDMVMATVKK SLGDMV LATVKKGKPDLRKKVLNAI I CRQSKAWRRHEGYYI YFEDNAGVIVNPKGEMKGSAITGPVARECAELWPKL SLGDMV LATVKKGKPDLRKKVLNAI ITRQSKAWRRHEGYFIYFEDNAGVIVTPR-RMKGSAITGPVARECAELWPK CVGDMVMATVKKGKPDLRKKVMPAVIVRQRK PWRRKDGVFMYFEDNAGVIVNPKGEMKGSAITCPIGKECADLWPRIA ---SA GVGDMVMATVKKGKPELRKKVHPAVVI RQRK SYRRKDGVFLYFEDNAGVIVNNKGEMKGSAITGPVAKECADLWPRIA---SN SLGDMVMATVKKGKPELRKKVMPAIVVRQAK SWRRRDGVFLYFEDNAGVIANPKGEMKGSAITGPVGKECADLWPRVA---SN SCGDMV LATVKKGK PDLRKKI MPAIVVRQRKAWRRKDGVYLYFEDNAGVIVNPKGEMKGSAITGPVAKECADLWPRIA---SN GLGDMIVATVKKGKPELRKKVMPAVVIRQRKPIRRREGIVLYFEDNAGVIVNNKGEMKGSAITGPVAKECADLWPRIA---SN GVGDMVMATVKKGKPELRKKVHPAVVIRQRKSYRRKDGVFLYFEDNAGVIVNNKGEMKGSAITGPVAKECADLWPRIA---SN SI GDMI LC SVKKGSPK LRKKVLQAIVI RQRR PWRRRDGVFIYFEDNAGVIANPKGEMKGSQITGPVAKECADIWPKVA---SN SVGDMV LATVKKGR PDLRKKVLPAVIVRQRKAWRRREGYFIYFEDNAGVIVNPKGEMKGSAINGPVAKECAELWPKIS---AA SVGDMV LATVKKGR PD LRKKVLPAVIVRQRKAWRRREGYFIYFEDNAGVIVNPKGEMKGSAINGPVAKECAELWPKIS---AA SV SDLIVVTCKKGK PA LRKKVSMGVVVRQRAI WRRKDGVVIGFQDNAGVI I NDKGEMKGSAITGPVAKEAAELWPKVA---SV A LGDMV MA SVKKGK PE LRRKV LNAVI I RQRK SWRRKDGTVI YFEDNAGVIVNPKGEMKGSGIAGPVAKEAAELWPKI S---TH ALGDIVMA SVKKGKPELRRKVLNAVIIRQRK SWRRKDGTVIYFEDNAGVIVNPKGEMKGSGIAGPVAKEAADLWPKIS---SH AAGDMVVASVKKGKPELRKKVMPAVVVRQRKPWRRRDGVFLYFEDNAGVIVNPKGEMKGSAITGPVAKECADIWPRIA---SN

Anopheles_gambiae/1-3220 Arabidopsis_thaliana/1-3286
Ashbya_gossypii/1-3281
A fumigatus/1-3241
Aspergillus_niger/1-3300 Bombyx_mori/1-2389
Bos_taurus/1-3273
C_briggsae/1-3231
Caenorhabditis elegans/1-3289
Candida_albicans/1-3266
Candida_glabrata/1-3282
Canis_familiaris/1-3286
Ciona intestinalis/1-2712
Cryptococcus_neoformans/1-3301 Cryptosporidium_hominis/1-3020 C_parvum/1-3281
Danio rerio/1-3286
Debaryomyces hansenii/1-3166
Dictyostelium_discoideum/1-3275
D_melanogaster/1-3287
Drosophila_pseudoobscura/1-3286
Encephalitozoon_cuniculi/1-3199
Entamoeba_histolytica/1-3271
Gallus_gallus/1-3286
Homo_sapiens/1-3286
Kluyveromyces_lactis/1-3281
Leishmania_major/1-3286
Macaca mulatta/1-3182
Magnaporthe_grisea/1-3286
Methanosarcina_acetivorans/1-3043
Monodelphis_domestica/1-3196
Mus musculus/1-3286
Neurospora_crassa/1-3264
Oryza_sativa/1-3218
Ostreococcus_lucimarinus/1-3287
Pan troglodyte/1-3286
Paramecium tetraurelia/1-3282
Pichia_stipitis/1-3290
_falciparum/1-3285
P_knowlesi/1-3284
Plasmodium_yoelii/1-3285
Populus_trichocarpa/1-3286
Rattus_norvegicus/1-3286
Saccharomyces_cerevisiae/1-3284
Schizosaccharomyces_pombe/1-3284
Strongylocentrotus_purpuratus/1-3252
Takifugu_rubripes/1-2930
etrahymena_thermophila/1-3270
T_annulata/1-3248
Theileria parva/1-3273
Trichomonas vaginalis/1-3198
T_brucei/1-3264
Trypanosoma_cruzi/1-3266
stilago maydis/1-3291
Yarrowia_lipolytica/1-3278

1420
1430
1440
1450
1460
ANAIG--I SRDSI HKKKRKY ELGRQPANTKL--SIRVRGGNVKWRALRLDTGNFSWGSEAVTRKTRI LDVAYNASNNELVRTQ SGVVG--I SRDSRHKKKRKFELGRQPANTKI - -GVRTRGGNKKFRALRI ETGNFSWASEGVSRKTRIVGVVYHPSNNELVRTN TSS-G--I SRDSRHKKKRAFEKGRQPANTRI--GVRTRGGNRKFRALRLESGNFSWGSEGI SRKTRVIVVAYHPSNNELVRTN SGVVG--I SRDSRHKKKRAFEKGRQPSNTRI--GVRTRGGNQKFRALRLESGNFSWGSEGI SRKTRVIVVAYHPSNNELVRTN ASSIG--I SRDHWHKKKRKY ELGRPAANTRL--GVRSRGGNTKYRALRLDTGNFSWGSECSTRKTRIIDVVYNASNNELVRTK AGSIG--I SRDNWHKKKRKY ELGRPAANTKI --GVRVRGGNKKYRALRLDVGNFSWGSECCTRKTRIIDVVYNASNNELVRTK AGSIG--I SRDSWHKKKRKFELGRPAANTKI --GVRTRGGNLKYRALRLDNGNFSWASEQTTRKTRIVDTMYNATNNELVRTK AGSIG--I SRD SWHKKKRKFELGRPAANTKI--GVRTRGGNEKYRALRLDSGNFSWASEQTTRKTRIVDTMYNATNNELVRTK SGVVG--I SRDSRHKKKRKFELGRQPANTKI--GVRTRGGNQKFRALRVETGNFSWGSEGVSRKTRIAGVVYHPSNNELVRTN SGVVG--I SRDSRHKKKRKFELGRQPANTKI - -GVRTRGGNQKFRALRI ETGNFSWASEGVAKKTRIVGVVYHPSNNELVRTN AGSIG--I SRDNWHKKKRKYELGRPAANTKI--GVRVRGGNKKYRALRLDVGNFSWGSECCTRKTRIIDVVYNASNNELVRTK ---MG--I SQD SWHKKKRKFELGRPAANTKI--GVRTRGGNTKFRGLR LDTGNFSWGSEACARKTRI IDVMYNASNNELLRTK AGTVG--ITRDSRHKKKRKFELGRQPAMTKLD-SVRTRGGNVKYRALRLDSGNFAWGSESVTRKTRLI QVRYNATNNELLRTQ APSIG--I SRDSRHKKKRKY EMGR PA SNTKL--GVRCRGGNKK FRALR LDSGNY SWG SQGVTRKARIMEVVYNA SNNELVRTK APSIG--I SRDSRHKKKRKY EMGRPASNTKL--GVRCRGGNKKFRALR LDSGNY SWG SQGI SRKARIMEVVYNA SNNELVRTK AGSIG--I SRDNWHKKKRKY ELGRPAANTKI --GIRVRGGNKKYRALRLDVGNFSWGSECCTRKTRIIDVVYNASNNELVRTK SGVVG--I SRDSRHKKKRKFELGRQPANTKI--GVRTRGGNEKFRALRI ETGNFSWGSEGVARKTRLAGVVYHPSNNELVRTN AGTIG--I SRDA LHKKKRKYELGRQAAKTKI - -CI RVRGGHQKFRALRLDTGNFSWATEKITRKCRI LNVVYNATSNDLVRTN ASSIG--I SRDSAHKKKRKFELGRPAANTKL--GVRTRGGNTKLRALRLETGNFAWASEGVARKTRIADVVYNA SNNELVRTK ASSIG--I SRDSAHKKKRKFELGRPAANTKL--GVRTRGGNSK LRALRLENGNFAWASEGVARKTRIADVVYNASNNELVRTK ASSIG--I NHRGDHKKKRNNRAGSQPSSTKI--GVRVRGGNRKYKALRLDMGHFKFITTGKFRMAKLLQVVYHPSSNELVRTN APTIG--ITRDSRHKKKKKNTMGRQPANTRL--GVRCRYGIIKRRALRLENGNFSWASQSITKGTKI LNVVYNASDNDFVRTN AGSIG--I SRDNWHKKKRKY ELGRPPANTKI --GVRVRGGNKKYRALRLDVGNFSWGSECCTRKTRIIDVVYNASNNELVRTK AGSIG--I SRDNWHKKKRKYELGRPAANTKI --GVRVRGGNKKYRALRLDVGNF SWGSECCTRKTRIIDVVYNASNNELVRTK SGVVG--I SRDSRHKKKRKFELGRQAANTKI--GVRTRGGNQKFRALRIETGNFSWASEGVARKTRITGVVYHPSNNELVRTN APAIG--IVRSR LHKKRMKAELGRLPANTRL--GVRARGGNFKIRALRLDTGNFAWASEAIAHRVRLLDVVYNATSNELVRTK AGSIG--I SRDNWHKKKRKY ELGRPAANTKI --GVRVRGGNKKYRALRLDVGNFSWGSECCTRKTRI IDVVYNASNNELVRTK SGVVG--I SRDSR HKQKRAWEAGRQPASTKI --GVRVRGGNTKYRALRLDSGNFSWGSEGVTRKTRVIAVAYHPSNNELVRTN ASII---MRWQGS SRGKRKFEMGRESAETRI--SVPTMGGNRKVRLLQSNVANVTNPKDGKTVTAPI ETVIDNTANKHYVRRN AGSIG--I SRDNWHKKKRKY ELGRPPANTKI --GVRVRGGNKKYRALRLDVGNFSWGSECCTRKTRIIDVVYNASNNELVRTK AGSIG--I SRDNWHKKKRKY ELGRPAANTKI--GVRVRGGNKKYRALRLDVGNFSWGSECCTRKTRIIDVVYNASNNELVRTK SGVVG--I SRDSRHKKKRAFEAGRQPANTRI--GVRTRGGNHKYRALRLDSGNFAWASEGCTRKTRVIVVAYHPSNNELVRTN ANAIG--I SRDSMHKKKRKY ELGRQPANTKL--SVRVRGGNLKWRALR LDTGNY SWG SEAVTRKTRI LDVVYNASNNELVRTQ ASSIG--I SRDSLHKKKRKY ELGRQPANTKL--SVRCRGGNIKHRALRLDTGNFAWGSENCTRKTRI LDVVYNASNNELVRTK AGSIG--I SRDNWHKKKRKY ELGRPAANTKI --GVRVRGGNKKYRALRLDVGNFSWGSECCTRKTRIIDVVYNASNNELVRTK AGSVG--I SRDSRHKKKRAFEKGRQAAMTKLVSGIRVRGGNFKFRALRLSEGNFSWGSQGIAKKAKIVEVVYHPSNNELVRTK SGVVG--I SRDSRHKKKRKFELGRQSANTKI --GVRTRGGNQK FRALRVETGNFSWGSEGVSRKTRIATVVYHPSNNELVRTN A SAI G--I SRDGRHKKKRKY ELGRPPSNTKL--GVRGRGRNYKYRAIKLDSGSFSWPTFGI SKNTRIIDVVYNASNNELVRTK ASAIG--I SRDGRHKKKRKY ELGRPPSNTKL--GVRGRGKNLKYRAIKLDSGSFSWPAFGVSKITRIIDVVYNASNNELVRTK ASAIG--I SRDGRHKKKRKY ELGRPPSNTKL--GVRGRGRNYKYRAIKLDSGSFSWPAFGI SKMTRIIDVVYNASNNELVRTK ANAIG--I SRDSMHKKKRKY ELGRQPASTKL--SIRVRGGNVKWRALR LDTGNY SWG SEAVTRKTRI LDVVYNA SNNELVRTQ AGSI G--I SRDNWHKKKRKY ELGRPAANTKI --GVRVRGGNKKYRALRLDVGNFSWGSECCTRKTRIIDVVYNASNNELVRTK SGVVG--I SRDSRHKKKRKFELGRQPANTKI --GVRTRGGNKKYRALRI ETGNFSWASEGI SKKTRIAGVVYHPSNNELVRTN AGTVG--I TRDSRHKKKRKFELGRQPSNTRI--GVRVRGGNKKFRALRLDSGNFSWGSEGVSKKTRII QVAYHPSNNELVRTN ASTI GGRI PDDTTRKAHYALPLARKKGAKLL--GVRCMGGNIKRRALRLDNGNFSWGSEHTTRKTRIIDVVYNASNNELVRTK AGSIG--I SRDNWHKKKRKY ELGRPPANTKL--GVRVRGGNKKYRALR LDVGNFSWGSECCTRKTRIIDVVYNASNNELVRTK AGSVG--I SRDSKHKKKRAFEKGRPI SMTKL--TVRVRGGHLKFRALRLCEGNFSWGSENITRKTKI LDVKYNATNNELVRTK APSIG--I SRDSRHKKKRKY ELGRPSSNTKL--GVRCRGGNLKFRALRLDSGNFSWGSQNVTRKTRVMDVVYNASSNELVRTK APSIG--I SRDSRHKKKRKY ELGRPSSNTKL--GVRCRGGNLKFRALRLDSGNFSWGSQNVTRKTRVMDVVYNASSNELVRTK APAVG--ITRMGDLKKKRNFLAGRPSAQTRI--GVRVRGGNLKMRALRLETGTFAWASENCTRKTRI LNVTYHPADNDLVRTN APAIG--IVRSR LHKKRMKAELGRLPANTKL--GVRARGGNFKLRGLR LDTGNFAWGTEASAQRARI LDVVYNATSNELVRTK APAIG--IVRSR LHKKRMKA ELGRLPAHTKL--GVRARGGNFKLRGLRLDTGNFAWGTEAIAQRARI LDVVYNATSNELVRTK AGTVG--I TRDSRHKKKRAFELGRQAANTRI--GVRVRGGNLKHRALRLESGNFAWGSEHITAKTRVLGVVYNASNNELVRTN SGVVG--I SRDSRHKKKRKFECGRQGAVTRI--GVRTRGGNKKFRAIRIETGNFSWGSEGTTRKTRVLGVSFHPSNNELIRTN

Anopheles_gambiae/1-3220
Arabidopsis_thaliana/1-3286
Ashbya_gossypii/1-3281
A fumigatus/1-3241
Aspergillus_niger/1-3300 Bombyx_mori/1-2389
Bos_taurus/1-3273
_briggsae/1-3231
Caenorhabditis_elegans/1-3289
Candida_albicans/1-3266 Candida_glabrata/1-3282
Canis_familiaris/1-3286
Ciona intestinalis/1-2712
Cryptococcus_neoformans/1-3301 Cryptosporidium_hominis/1-3020 C_parvum/1-3281
Danio_rerio/1-3286
Debaryomyces_hansenii/1-3166
Dictyostelium_discoideum /1-3275 D_melanogaster/1-3287
Drosophila_pseudoobscura/1-3286 Encephalitozoon cuniculi/1-3199 Entamoeba_histolytica/1-3271 Gallus_gallus/1-3286
Homo_sapiens/1-3286
Kluyveromyces_lactis/1-3281 Leishmania_major/1-3286 Macaca_mulatta/1-3182 Magnaporthe_grisea/1-3286 Methanosarcina_acetivorans/1-3043 Monodelphis_domestica/1-3196
Mus musculus/1-3286
Neurospora_crassa/1-3264
Oryza_sativa/1-3218
Ostreococcus_lucimarinus/1-3287
Pan_troglodyte/1-3286
Paramecium_tetraurelia/1-3282
Pichia_stipitis/1-3290
_falciparum/1-3285
P_knowlesi/1-3284
Plasmodium_yoelii/1-3285
Populus_trichocarpa/1-3286
Rattus_norvegicus/1-3286
Saccharomyces_cerevisiae/1-3284
Schizosaccharomyces_pombe/1-3284
Strongylocentrotus_purpuratus/1-3252
Takifugu_rubripes/1-2930
Tetrahymena_thermophila/1-3270
T_annulata/1-3248
Theileria parva/1-3273
Trichomonas vaginalis/1-3198
T_brucei/1-3264
Trypanosoma_cruzi/1-3266
stilago_maydis/1-3291
Yarrowia_lipolytica/1-3278
$1500 \quad 1510$
1510 - 1520
1520
LVKNAI IVI DASPFRQWY E'SHY SK SNLRKYVKR T LVK SAI VQVDAA PFKQGY LQHY SNHVQRK LEMR T LTK SA I VQI DAT PFRQWY E SHY SR NA ERKWAAR TLTK SAVVQI DAA PFR QWY EAHY SN SVVKKQAAR T LTK SAVVQI DAAPFRQWY EAHY SN SVVKKQAAR T LVKNA I VVVDAT PFRQWY ESHY SQKTARKY LAR T LVKNCIVLIDSTPYRQWY ESHY SKKI QKKYDER T LVKGAI V SVDAA PFRQWY EAHY SNHT LKKYT ER TLVKGAI I SVDAAPFRQWY EAHY SHHTMKKYTER T LTK SAVVQI DAT PFRQWY ENHY SRKVERK LAAR T LTKAA I VQI DAT PFRQWY EAHY SK SA ERKWAAR TLVKNCIVLIDST PYRQWY ESHY SKKI QKKYDER TLVKNAII QIDSTPFRQWY EAHY SKKTQKKYEER T LVKGAVVDI DAT PFRQWY ESHY SNHVKRI LEER TLVKNAIVVI DAT PFRQFY LQRY SGHLLATRKAR T LVKNA I VVI DAT PFRQFY LQRY SGHLLATRKAR T LVKNCVV LVDST PYR QWY ESHY SKKVQKKFT LR TLTKAAIVQI DAT PFKQWFETHY SRKVERK LAQR T LVKGSIVQI DAT PYK QWY ETHY SA S LLAK LA SR TLVKNSIVVI DAT PFRQWY EAHY SEKVMKKY LER T LVKNSIVVI DAT PFRQWY ESHY S EKVMKKY LER LTK SSVVK I SAEPFKNDIK $\qquad$ TLVKGAII EIDPAPFR LWFLK FY SKTMQKKYAKKL T LVKNCIVLVDST PYRQWY EAHY SKKI QKKYDER TLVKNCIVLIDST PYRQWY ESHY SKKI QKKYDER T LTKAAI VQI DAT PFRQWY ESHY SKNT ERKWAAR T LVKNCIVAVDAA PFKRWYAKHY SPK LQR EWTRR T LVKNCIVLIDSTPYRQWY ESHY SKKI QKKYDER TLTKSAVI QIDAAPFRQWY EAHY SK SVEKKQAERF I LTKGSVIRTSMGT

1540
1550
1560
 ---QEGRALDSHLEEQFSSGR LLACI ASRPGQCGRADGY LEGKELEFY --AGDAKI EGAVDSQFSAGRLYACI SSRPGQSGRCDGYI LEGEELAFY ---ADHGKVEPAI EKQFESGRLYAVIASRPGQSGRVDGILEGEELAFY F--AEQGKVESAV ERQFESGRLYAVVSSRPGQSGRVDGI LEGEELAFY ---QR LAKVEGALEEQFHTGR LLACVASRPGQCGRADGYI LEGKELEFY ---KKNAKISSLLEEQFQQGKLLACIASRPGQCGRADGYVLEGKELEFY ---QKTAAVDALLT EQFNTGRLLARI SSSPGQVGQANGYI LEGK ELDFY ---QKTAAVDALLIEQFNTGRLLARI SSSPGQVGQANGYI LEGKELDFY ---SGAAAI ESAVDSQFGSGRLYAVI S SR PGQSGRCDGYI LEGEELAFY ---AASAKVESAVDSQFSAGRLYACI SSR PGQSGRCDGYI LEGEELAFY --KKNAKISSLLEEQFQQGKLLACIASR PGQCGRADGYVLEGKELEFY ---KK EPKVAQALEEQFNQGRILACI SSR PGQSGRCDGYI LEGKELEFY ---KKVAKI DPLLEQQFRAGRLLAVI S SR PGQSGRADGYI LEGKELEFY ---LMNNVIDPLVEEQFGIGRLLACV SSRPGQCGRCDGYI LEGKELEFY ---LMNNVIDPLVEEQFGI GRLLACV SSRPGQCGRCDGYI LEGKELEFY ---RKTAKI SPLLEEQFLQGKLLACI SSRPGQCGRADGYVLEGKELEFY ---SGASNI ESAV EHQFNAGRLYAAI S SR PGQSGRCDGYI LEGEELAFY ---AKGRVLDSAI ESQIGEGRFARITSRPGQVGKCDGYILEAKELEFY ---QKYGKV EQAL EDQFT SGRI LACI S SR PGQCGR SDGYI LEGKELEFY ---QK FGKV EQALEDQFT SGRI LACI S SR PGQCGR SDGYI LEGK ELEFY --DVARDVDPSLHESFEKGHLYAIITSRPGQVGMAQGHV LQGDELKFY EV LKNMK FDEALLEGFQSGRVLACI SSR PGQTGSVEGYI LEGK ELDFY --KKNAKIASILEEQFQQGKLACIASRPGQCGRADGYVLEGKELEFY ---KKNAKI SSLLEEQFQQGKLLACIASR PGQCGRADGYVLEGKELEFY ---AAEAKI EHAVDSQFGAGR LYAAI S SR PGQSGRCDGYILEGEELAFY --RRNHRVEKAIADQLREGRVLARITSRPGQSGADGI LLEGAELQFY --KKNAKISSLLEEQFQQGKLLACIASRPGQCGRADGYVLEGKELEFY --AARGKVDSALEKQFEAGRVFAVV SSRPGOSGRCDGYI LEGEELAFY TLVKNCIVLVDST PYRQWY ESHY SKKIQKKYEER---KKNAKISPLLEEQFQQGKLLACIASRPGQCGRADGYVLEGKELEFY TLVKNCIVLIDSTPYRQWY ESHY SKKI QKKYDER---KKNAKISSLLEEQFQQGKLLACIASRPGQCGRADGYVLEGKELEFY TLTKSAVVQIDAAPFRQWY EAHY SK SVEKKQAERF--AAAGKVDPALEKQFEAGRLYAVI SSRPGQSGRCDGYILEGEELAFY TLVK SAIVQVDAA PFKQWY LTHY SNHVVRKLEKR---QQTRTLDSHI EEQFGSGRLLACI SSR PGQCGRADGYI LEGKELEFY TLVK SAVIAVDAAPFRAWYAQHY SK SVTMK LRSR---NQKHEVAKAIDEQFATGRLLAIIISRPGQCGRADGYVLEGKELEFY TLVKNCIVLIDST PYRQWY ESHY SKKI QKKYDER---KKNAKI SSLLEEQFQOGKLLACIASRPGQCGRADGYVLEGKELEFY TLTRGVIVQVDATPFRQWYAKKY SRSLIKKLEQR---AKDNAIDALVQEQFTNQRLLVRITSRPGQSGRADGYI LEGKELEFY TLTKAAIVQIDAIPFRQWY ENHY SRKVERKLASR---AGQAAI ESAVDAQFGSGKLYAAI SSRPGQSGRCDGYILEGEELAFY TLVKNCIVVIDSHPFTTWY ENTFTYGVIKKI-----GK SKNIDPLLLEQFK QGRVLACISSRPGQCGKADGY।IEGDELLFY TLVKNCIVLIDSHPFTTWY ENTFSYSVIKKI------GK SKQIDPALLEQFK QGRVLACISSRPGQCGKADGYIIEGDELLFY TLVKNCIVLIDSHPFTAWY ENTFSYSVIKKI------GKAKQIDPALLEQFKQGRVLACI SSRPGQCGKADGYII EGDELLFY TLVKSAIVQVDAAPFKQWY LQHY SNHVIRKLEKR---QQVRKLDPHIEEQFGSGRLLASISSRPGQCGRADGYILEGKELEFY TLVKNCIVLIDSTPYRQWYESHY SKKI QKKYDER---KKNAKISSLLEEQFQQGKLLACIASRPGQCGRADGYVLEGKELEFY TLTKAAIVQIDAIPFRQWFEAHY SKNAERKWAAR---AASAKI ESSVESQFSAGRLYACI SSRPGQSGRCDGYILEGEELAFY TLTKSAIVQIDAAPFRVWY ETHY SKHVQRKHSAR---LGDSKVDSALETQFAAGRLYAVVSSRPGQSGRCDGYILEGEELHFY TLVKNAIVQIDSTPFRQWY EAHY SKKVVKKFFEER---KKTAKVAQALEEQFGTGRLLACIASRPGQCGRADGYILEGKELDFY TLVKNCIILVDSLPFRQWY EAHY SKKTQKKYDER---KKTAKI STLLEEQFQQGKLLACIASRPGQCGRADGYILEGKELEFY TLVKNSIVEIDSTPFREWYK LHY SRHVQKRV-KR---TKAQALEKNIEEQFVSQRI LACITSRPGQSGRADGYILEGKELEFY TLVKNAIVTVDPTPFKLWFKTHYSEKV------------AGLVPKTLLEQFSSGRLLACISSRPGQCGRCDGYVLEGEELNFY TLVKNAIVTVDPTPFKLWFKTHYSEKV------------AALVPRTLLDQFSSGRLLACI SSRPGQCGRCDGYVLEGEELNFY TLARGSVVSIDAAPFKQWY ERQFTDKMT QRWAAN---KDGGVVAPELVAEFDQGRLLAVITSRPGQCGRADGYILEGEELAFY TLVKNCIVVVDAAPFRLWYAKHY SSKLKRKWEYR--RKHHKI EKALADQLREGRLLARITSRPGQTGRADGALLEGAELQFY TLVKNCIVVVDAAPFKLWYAKHY SDELKRKWMLR---R ENHKI EKAVADQLKEGRLLARITSRPGQTARADGALLEGAELQFY TLVKGCIVQVDATPFRQAYEKHY SNNVTRKLENR---RKEGKLDSLVEQQFGAGRLYAAVSSRPGQSGRCDGYI LEGKELEFY TLTK SAIVQIDATPFRQWY ESYYAEADQAAVAAR---QADAKLDPAVEAQFGAGRLYACVSSRPGQSGRVDGYVLEGEELAFY

Anopheles_gambiae/1-3220
Arabidopsis_thaliana/1-3286
Ashbya_gossypii/1-3281
A fumigatus/1-3241
Aspergillus_niger/1-3300
Bombyx_mori/1-2389
Bos_taurus/1-3273
C_briggsae/1-3231
Caenorhabditis elegans/1-3289
Candida_albicans/1-3266
Candida_glabrata/1-3282
Canis_familiaris/1-3286
Ciona intestinalis/1-2712
Cryptococcus_neoformans/1-3301 Cryptosporidium_hominis/1-3020 C_parvum/1-3281
Danio rerio/1-3286
Debaryomyces hansenii/1-3166
Dictyostelium_discoideum /1-3275 D_melanogaster/1-3287
Drosophila_pseudoobscura/1-3286 Encephalitozoon cuniculi/1-3199 Entamoeba_histolytica/1-3271 Gallus_gallus/1-3286
Homo_sapiens/1-3286
Kluyveromyces_lactis/1-3281 Leishmania_major/1-3286 Macaca_mulatta/1-3182 Magnaporthe_grisea/1-3286 Methanosarcina_acetivorans/1-3043 Monodelphis_domestica/1-3196
Mus musculus/1-3286
Neurospora_crassa/1-3264
Oryza_sativa/1-3218
Ostreococcus_lucimarinus/1-3287
Pan_troglodyte/1-3286
Paramecium_tetraurelia/1-3282
Pichia_stipitis/1-3290
_falciparum/1-3285
P_knowlesi/1-3284
Plasmodium_yoelii/1-3285
Populus_trichocarpa/1-3286
Rattus_norvegicus/1-3286
Saccharomyces_cerevisiae/1-3284
schizosaccharomyces_pombe/1-3284
Strongylocentrotus_purpuratus/1-3252
Takifugu_rubripes/1-2930
Tetrahymena_thermophila/1-3270
T_annulata/1-3248
Theileria parva/1-3273
Trichomonas vaginalis/1-3198
T_brucei/1-3264
rypanosoma_cruzi/1-3266
stilago_maydis/1-3291
Yarrowia_lipolytica/1-3278

158
1590
1600
1610
1620
1630
1640
1650
 MKK LQKKKMV LAELGGRITRAI QQMSNVIIIDEK -ALNECLNEI TRALLQSDVSFPLVK EMQSNIKEQAIFSELCKMVVMFVG VRRLTAKKMV LADLGKRINAAVAQALNNDTDDYVAGVETMLKAIVTALLENDVNIKLVSSVRSNIKQKTVFEELCALVVMFVG QRAIRK--MVLQDLGRRINAAVNDLTRSSNLDEK-AFDDMLKEICAALLSADVNVRLVQTLRKSIKQKAVFDELVALVIMFVG QRAIRK--MVLQDLGRRINAAVNDLIR SNNLDEK QAFDDMIKEICAALLSADVNVRLVQSLRKSIKQKAVFDELVSLVIMFVG
 LRKIRAKKMV LADLGRKIRNAIGKLGQNTVINEE-ELDLMLKEVCTALI ESDVHIRLVKQLKDNVKQKTVFNELLKLVFMFVG LRKIRAKKMV LADLGRKIRNAIGKLGQSTVINEG-ELDLMLKEVCTALI ESDVHIRLVKQLKDNVKQKTVFNELLKLVFMFVG LRR LTAKKMVLADLGSRLRGALSSVESGS---DD-EI QQMIKDICSALLESDVNVKLVAKLRGNIKQKIIFDELCALVIMFVG LRR LTAKKMVLADLGKRINNAVNSALSNTEDDYVNSIDGMLKGISTALLEADVNIMLVSKVRNNIRQKTVFDELCGLIIMFVG LRKIKARKMVLADLGRKITSALRSLSNATIINEE-VLNAMLKEVCTALLEADVNIKLVKQLRENVKQHAVFKELVKLVIMFVG QRKIKARKMVLADLGRKINNALRSLSNATIINEE-VLQSMLSEICRALLESDVNIRLVKKLRENVRQSAVFRELVKLVIMFVG HHK LQIRKMV LADLGTRLHGAWNQLSKASVIDDK -VIDGVLKELCAALLESDVNVKLVASLRTKVKQKAVFDELVALVLMAVG
 LRKIKAKKMV LADLGRKITSALRSLSNATIINEE-VLNAMLK EVCAALLEADVNIKLVKQLRENVKQHAVFKELVKLVIMFVG LRR LTAKKMV LADLGSRLRGALSNVESGS---ET-EIQSMIKDICNALLESDVNIKLVAKLRDNIKQKIVFDELCSLVIMFVG QRRLQKKKMVLADLGNQLSSALRSLNETTIVNED-TINQLLKEVGNALSK SDVSMSLII QMRKNIKKQVVFDELIRLIVMFVG LKKIKSKKMVLADLGRKITTALHSLSKATVINEE-ALNSMLKEICAALLEADVNIRLVKQLRENVRQSAVFKELVKLVVMFVG LKKIK SKKMV LADLGRKITTALHSLSKATVINEE-ALNSMLKEICAALLEADVNIRLVKQLRENVRQSAVFKELIKLIIMFVG ADK FNKKS-MITELGRSITNTLSNLLSSPATDQH--IETAIREICNSLILSNVNPRYVSDLRDELRQNAVYERLVDLVVVFVG SKKI SDKKMVLSQLGSSLVTALRKMTSSTVVDEE-VINTLLKEIETSLLGEDVNPIFIRQMVNNIKKDSVFEELINLVLMMVG LRKIKARKMVLADLGRKITSALRSLSNATIINEE-VLNAMLKEVCTALLEADVNIKLVKQLRENVKQHAVFKELVKLIIMFVG LRKIKARKMVLADLGRKITSALRSLSNATIINEE-VLNAMLKEVCTALLEADVNIKLVKQLRENVKQHAVFKELVKLVIMFVG LRR LTAKKMV LADLGKRINNAVTNAI SNEQTDY ETTVQSMLKEIATALLENDVNIRLVSRLRENIKQKTVFDELCNLVIMFVG LKR LEKKKMV LAELGQKIGQAIHRMSAKSMLGED-DVK ELMNEIARALLQADVNVTIVKKLQVSIRQNAVFNGLKRIVVMFVG LRKIKARKMVLADLGRKITSALRSLSNATIINEE-VLNAMLKEVCTALLEADVNIKLVKQLRENVKQHAVFKELVKLVIMFVG QRKLHK--MVLQDLGRRINAAVSDLIRAPNLDEK-A-----KICAALLEADVNVR LVGQLRKSIKGKAVFDELVSLVIMFVG LRKIKAR--MVMEKLGDSLQGALKKLIGAGRIDER-TVNEVVKDI QRALLQADVNVKLVMGMSQRIKIRIVYQELMEITIMMVG LRKKARKMV LADLGRKITSALRSLSNATIINEE-VLNAMLK EVCTALLEADVNIKLVKQLRENVKQHAVFKELVKLVIMFVG LRKIKARKMVLADLGRKITSALRSLSNATIINEE-VLNAMLKEVCTALLEADVNIKLVKQLRENVKQHAVFKELVKLVIMFVG QRK LHK--MV LQDLGRRINAAV SDLTRAPNLDEK-AFDGMLKEICSALLEADVNVR LVGQLRK SI KQKAVFDEEVRLVIMFVG MKK LQRKKMVLAQLGGSI SRALAQMSNATVIDEK -VLSDCLNEI SRALLQSDVQFKMVRDMQSNIKQQAVFTELCNMVVMFVG QKKMmKKKMVLNDLGNKIASALRSLNAHVVVDEE-LLDACLKDITNALLASDVAVPLVVrmkKNIVERAVFKELTALVVmFVG LRKIKARKMVLADLGRKITSALRSLSNATIINEE-VLNAMLKEVCTALLEADVNIKLVKQLRENVKQHAVFKELVKLVIMFVG IKKVEQKKMVLAELGKSINAALQK LSKAPVVDEA-LVDQI LGEIAMALLKADVNAKFIKKLREDVKQKAVVDGLTRMVIMFVG LRR LTAKKMVLADLGSRLRGALSSVESAS---DE-EINQMI KDVCTALLESDVNIKLVVKLRDNIKQKIIYDELVGLIIMFVG KRKMDKKKMV LTELGTQITNAFRKLQT STLADDV-VIEECLKEIIRALILSDINVSYLKDIKSNIKQKYVVEELIKLVILFVG KRKMDKKKMVLTELGTQLT SALQK LQA SAVADDS-AIEECLKEVIRALI LADINISYLKDIKSNIKQQYVVEELINLVILFVG KRKMDKKKMV LTELGAQLTSALQKI QAAPVADDN-VIEECLKEIVRALILADINVIYLKDIKSNIKQKYVVEELIKLVILFVG MKKI QRKKMVLAQLGGSI SRAI QQMSNATIIDEK-ALNDCLNEITRALLQSDVQFK LVRDMQTNIKQQAI FNELCKIVVMFVG LRKIKARKMVLADLGRKITSALRSLSNATIINEE-VLNAMLKEVCTALLEADVNIKLVKQLRENVKQHAVFKELVKLVIMFVG LRRLTAKKMV LADLGKRINSAVNNAI SNTQDDFTTSVDVMLKGIVTALLESDVNIALVSKLRNNIRQKTVFDELCKLIIMFVG LRRMAPKKMV FADLGRRLNSALGDFSKAT SVNEE-LVDTLLKNICTALLETDVNVR LVQELRSNI KqKAVFDELCSLVIMMVG mRKMRAKKMV LADLGRKITSALKSLSNATIIDED-VLNSMLNEICRALLEADVNIRLVKALKENVKQTAVFKELVKLIIMFVG LRKIKAKKMVLADLGRKITSALRSLSNATIINEE-VLNAMLKEVCAALLEADVNIKLVKQLRENVKOHAVFKELVKLVIMFVG IRKLQSKKMVLADLGKRINNALQQLNKAPVIDEE-LLNQVLKEI QLALLQSDVNVKYVAKLKSNI IQQAVVQELTQMVVMFVG RRRMDKKKMVLAELSNQITQAFRKLHSTTVISEA-VIEEVIGDIVRALLMADVNVKLVHKLKENVKQKIVVDELVNMVIMFVG RRRMDKKKMVLAELSNQITKAFRKLHSTIVISEA-VIEEVIGDIVRALLMADVNVKLVHK LKENVKQKIVVDELVNMVIMFVG SDKIAKKK-MLQDLGEKLMGSIKKLSESKTIDEK -VYVT FMAEVAK SLIAADCSKEIVFDF SRR LKEKAV FNELVKLI FMMVG LKK LDKKKMV LAELGQKIGAAISKMSSKSFVGED-DVKEFLNEVARALLQADVNVKTVKELQQNVRQTAVFNGIKKMIVMFVG LKKLEKKKMVLAELGQKIGGAISKMSSKPLLGED-DVK EFLNEVARALLQADVHVTTVKELQQTIRQTAVFSGLRKIIVMFVG VRRLKASKMVLSDLGRIINSAFQDLSKVPTVDAA-SIDQLLKSVCNALI EADVNVKLVANLRSQVKGAVFDHLVALVIMFVG LKKIVSKKMVLEDLGKRINGAFANLSKGGDIDE--ALDAMLKEVCSALLESDVNIKLVSQLRQKVKQKALFDELVNLVVMFVG

## 1670

1680

Anopheles_gambiae/1-3220 Arabidopsis_thaliana/1-3286 Ashbya_gossypii/1-3281
A fumigatus/1-3241 Aspergillus_niger/1-3300 Bombyx_mori/1-2389
Bos_taurus/1-3273
C_briggsae/1-3231
Caenorhabditis elegans/1-3289
Candida_albicans/1-3266 Candida_glabrata/1-3282
Canis_familiaris/1-3286
Ciona intestinalis/1-2712
Cryptococcus_neoformans/1-3301 Cryptosporidium_hominis/1-3020 C_parvum/1-3281
Danio rerio/1-3286
Debaryomyces hansenii/1-3166
Dictyostelium_discoideum /1-3275 D_melanogaster/1-3287
Drosophila_pseudoobscura/1-3286 Encephalitozoon cuniculi/1-3199 Entamoeba_histolytica/1-3271 Gallus_gallus/1-3286
Homo_sapiens/1-3286
Kluyveromyces_lactis/1-3281 Leishmania_major/1-3286 Macaca_mulatta/1-3182 Magnaporthe_grisea/1-3286 Methanosarcina_acetivorans/1-3043 Monodelphis_domestica/1-3196
Mus musculus/1-3286
Neurospora_crassa/1-3264
Oryza_sativa/1-3218
Ostreococcus_lucimarinus/1-3287
Pan_troglodyte/1-3286
Paramecium tetraurelia/1-3282
Pichia_stipitis/1-3290
_falciparum/1-3285
P_knowlesi/1-3284
Plasmodium_yoelii/1-3285
Populus_trichocarpa/1-3286
Rattus_norvegicus/1-3286
Saccharomyces_cerevisiae/1-3284
Schizosaccharomyces_pombe/1-3284
Strongylocentrotus_purpuratus/1-3252
Takifugu_rubripes/1-2930
Tetrahymena_thermophila/1-3270
T_annulata/1-3248
Theileria parva/1-3273
Trichomonas_vaginalis/1-3198
T_brucei/1-3264
rypanosoma_cruzi/1-3266
stilago maydis/1-3291
Yarrowia_lipolytica/1-3278

LQGSGKTTTCTK LAYHYQKKNWK SC LVCADTFRAGAYDQI KQNATKARI PFYGSYTEVDPVTI AQDGVEMFKKEGFEFI IVDT LQGAGKTTTCTKYAYYHQKKGYKPA LVCADT FRAGAFDQLKQNATKAKI PFYGSYTESDPVKIAVEGVDTFKKENCDLI IVDT LQGAGK ST SCSK LAVYY SKRGFKVG LVCADTFRAGAFDQLKQNAIKAKI PFYGSYTETNPVRVAADGVAKFKKERFEIIIVDT LQGAGKTTTCTK LAR HY QMRGFKTA LVCADT FRAGAFDQLKQNATKAKI PYYGSLTQTDPAVVAAEGVAKFKKERFEVIIVDT LQGAGKTTTCTK LAR HY QMRGFKTALVCADTFRAGA FDQLKQNATKAKI PYYGSLTQTDPAIVAAEGVAKFKKERFEI I IVDT

LQG SGKTTTCSK LAYYYQRKGWKTCLICADTFRAGAFDQLKQNATKARIPFYGSYTEMDPVIIASEGVEKFKNENFEIIIVDT LQG SGKTTTCSKMAYYYQRKGWKTCLICADT FRAGAFDQLKQNATKARIPFYGSYSEIDPVKIAAEGVEKFTKEGFEIIIVDT LQGSGKTTTCTKMAYYYQRKGWKTCLICADTFRAGAFDQLKQNATKARIPFYGSYSEIDPVKIAAEGVEKFTQEGFEII IVDT LQGAGKTT SCTK LAVYYKKRGFKVGLVCADT FRAGAFDQLKQNAIKANI PYYGSYLEPDPVKIAFEGVQKFKQEKFDI I IVDT LQGSGKTT SCTK LAVYY SKRGFKVGLVCADT FRAGAFDQLKQNAVKARI PFYGSYTETDPVKVAGDGIAKFKKEKFDVI IVDT LQG SGKTTTCSK LAYYYQRKGWKTCLICADT FRAGAFDQLKQNATKARIPFYGSYTEMDPVI I ASEGVEKFKNENFEI I IVDT LQG SGKTTTCTK LAYYYQRKNWKTCLICADT FRAGAFDQLKQNATKARIPFYGSYTEADPVVIASEGVETFKEENFEII IVDT I QGAGKTTTCTK LAVHYQRRGFRTCLVCADT FRAGAFDQLKQNATKAKI PFYGSYTETDPVAIASLGVEKFRKERFDVI IVDT

LQGSGKTTTCTKFANYYQRRGWKTALVCADTFRAGAFDQLKQNATKVKI PFYGSYTETDPVKIARDGVREFRKEGYDLIIVDT LQGSGKTTTCSK LAYY FQRKGWKTCLICADTFRAGAFDQLKQNATKARIPFYGSYTEMDPVIIAAEGVEKFKSENFEIIIVDT LQGAGKTT SCTK LAVYYKKRGYKVGLVCADTFRAGAFDQLKQNAIKSSI PYYGSYI ETDPVKVAYEGVVKFKQEKFDIIIVDT LQGAGKTT SVTK LAY FYKKKGFSTAIVCADT FRAGAFDQVRHNAAKAKI HYYGSETEKDPVVVARTGVDIFKKDGTEI I IVDT LQGSGKTTTCTK LAYHY QKRNWK SC LVCADT FRAGAYDQVKQNATKARI PFYGSYTEIDPVVIAQDGVDMFKREGFEMI IVDT LQG SGKTTTCTK LAYHY QKRNWK SC LVCADT FRAGAYDQI KQNATKARI PFYGSYTEI DPVVI AQEGVDMFKREGFEMI IVDT LQGSGKTT SI CKYANFYKKKGYKVGIVCADT FRAGAFDQVRQNALKIKVPFFGS-SEADPVKVASAGVERFRKERFELI LVDT LQGAGKTTTITK LA LYYKNRGYKPAVVGADTFRAGAY EQLQMNAKRAGVPFFGIKEESDPVKVASEGVRTFRKEKNDII LVDT LQGSGKTTTCSK LAY FY QRKGWKTCLICADTYRAGAFDQLKQNATKARIPFYGSYTEMDPVIIASEGVEKFKNENFEIIIVDT LQG SGKTTTCSK LAYYYQRKGWKTCLICADTFRAGAFDQLKQNATKARIPFYGSYTEMDPVIIASEGVEKFKNENFEIIIVDT LQG SGKTT SCTK LAVYY SKRGYKVGLVCADT FRAGAFDQLKQNAIKAKI PFYGSYTEPNPVKVAKDGVDKFKKEKFEIIIVDT LQG SGKTT SCTKYAAY FQRKGFKTA LVCADT FRAGAYDQLRQNATKAKVRFYGSLTEADPVAIAKEGVAELKKEKYDLI IVDT LQGSGKTTTC SK LAYYYQRKGWKTCLICADTFRAGAFDQLKQNATKARIPFYGSYTEMDPVIIASEGVEKFKNENFEII IVDT LQGAGKTTTCTK LARHYQSRGFKAC LVCADT FRAGAFDQLKQNATKAKI PYYGSLTETDPAVVAREGVDKFKKERFEVI IVDT LQG SGKTT SAAK LARY FQRKG LKAGVVAADT FR PGAYHQLKT LAEK LNVGFYGEEGNPDAVEITKNGLKAL--EKYDIRIVDT LQG SGKTTTCSK LAYYYQRKGWKTCLICADTFRAGAFDQLKQNATKARIPFYGSYTEMDPVIIASEGVEKFKNENFEIIIVDT LQG SGKTTTCSK LAYYYQRKGWKTCLICADTFRAGAFDQLKQNATKARIPFYGSYTEMDPVIIASEGVEKFKNENFEIIIVDT LQGAGKTTTCTK LARHYQSRGFRVGLVCADT FRAGAFDQLKQNATKAKI PYYGSLTETDPVVVARDGVDKFKKEKFEI I IVDT LQGSGKTTTCTKYAYYHQRKGFKPA LVCADT FRAGAFDQLKQNATKAKI PFYGSYMESDPVKIAVEGVERFKKENCDLI IVDT LQGAGKTTTCTK FAHYYAKKGFKPSLVCADT FRAGAFDQLKQNATKAKI PFYGSYTESDPATIAAAGVKRFEEEKSDLI IVDT LQG SGKTTTCSK LAYYY QRKGWKTCLICADT FRAGAFDQLKQNATKARIPFYGSYTEMDPVIIASEGVEKFKNENFEII IVDT LQGSGKTTTCTKYAYYYQKKGWKVALVCADTFRAGAFDQLKQNATKVRVPFYGSYTEADPVQIAQEGVNVFKKEAFEIIIVDT LQGAGKTT SCTK LAVYYKKRGFKVGLVCADT FRAGAFDQLKQNAIKASI PYYGSYLEQDPVKIAYEGVTKFRSEKFDII IVDT LQG SGKTTTCTKYAHYY QKKGFKTALVCADT FRAGAFDQLKQNAAKVKI PFYGSYSEVDPVKIATDGVNAFLKDKYDLI IVDS LQG SGKTTTCTK FAHYYQKKGFKTA LVCADT FRAGAFDQLKQNAAKVKI PFYGSYSEVDPVKIASDGVNAFLKEKYDLI IVDS LQG SGKTTTCTKYAHYYQKKGFKTALICADTFRAGAFDQLKQNAAKVKI PFYGSYSEVDPVKIATDGVNTFLKDKYDLIIVDS LQGSGKTTTCTKYAYYHQKKGWK PA LVCADT FRAGAFDQLKQNATKAKI PFYGSYTESDPVKIAEEGVETFKKENCDLI IVDT LQG SGKTTTCSK LAY FY QRKGWKTCLICADT FRAGAFDQLKQNATKARIPFYGSYTEMDPVIIASEGVEKFKNENFEIIIVDT LQG SGKTT SCTK LAVYY SKRGFKVGLVCADTFRAGAFDQLKQNAIRARIPFYGSYTETDPAKVAEEGINKFKKEKFDIIIVDT LQG SGKTTTCSK LA LHY QRRGLK SC LVAADT FRAGA FDQLKQNAIKARVPYFGSYTETDPVVIAKEGVDKFKNDR FDVI IVDT LQG SGKTTTCTK LAYYYQKKGWKVALICADTFRAGAFDQLKQNATKARIPFYGSYTEVDPVIIAADGVEKFKKENFEII IVDT LQG SGKTTTC SK LAYYYQRKGWKTCLICADT FRAGAFDQLKQNATKARI PFYGSYTEMDPVVIATEGVDKFKMENFEI I IVDT LQGAGKTTTCTKYAY HWQKKGWRTALICADTFRAGAFDQLKQNATKVRVPFYGSYSETDPVAIAEEGVKHFKKENYEMI IVDT LQGAGKTT SCTK FAYHY QRKGWRTALICADT FRAGAFDQLKQNAAKVKI SFYGSY SEANPAKVAADGVARFKEEKYDMI IVDT LQGAGKTT SCTK FAY HY QR KGWRTA LI CADT FRAGA FDQLKQNAAKVKI SFYGSYSEANPAKVAADGVAR FKEEKYDMI IVDT LQGAGKTTTVTK LANFYKRRNWRTGVIAADTFRAGAREQLMQNAQTARIPYFVDFTEQDPVQAALKGIEKFRKDKYEIVIIDT LQG SGKTT SCTKYAAY FQRKGLKTG LVCADT FRAGAYDQLRQNATKAKVRFYGSLTEADPVI I AK EGVLELKKEKYDLI IVDT LQG SGKTT SCTKYAAY FQRKGLKTALVCADT FRAGAYDQLRQNATKAKIRFYGSLTEADPVIIAKEGVAELEKEKYDLIIIDT LQG SGKTT SCTK LA LYY QKRGFKTG LVCADT FRAGA FDQLKQNA SKI NV PFYGSYTETDPVAI SAAGVASFKQNRFEVI IVDT LQGSGKTT SCTK LAVYY QRRGFKVGLVCADT FRAGAFDQLKQNATKAKI PFFGSYTETDPVAVAAEGVAKFKKEKFEII IVDT

1750
1760

Anopheles_gambiae/1-3220 Arabidopsis_thaliana/1-3286
Ashbya_gossypii/1-3281
A fumigatus/1-3241
Aspergillus_niger/1-3300 Bombyx_mori/1-2389
Bos_taurus/1-3273
_briggsae/1-3231
Caenorhabditis_elegans/1-3289
Candida_albicans/1-3266
Candida_glabrata/1-3282
Canis_familiaris/1-3286
Ciona intestinalis/1-2712
Cryptococcus_neoformans/1-3301 Cryptosporidium_hominis/1-3020 C_parvum/1-3281
Danio_rerio/1-3286
Debaryomyces hansenii/1-3166
Dictyostelium_discoideum /1-3275 D_melanogaster/1-3287
Drosophila_pseudoobscura/1-3286 Encephalitozoon cuniculi/1-3199 Entamoeba_histolytica/1-3271 Gallus_gallus/1-3286
Homo_sapiens/1-3286
Kluyveromyces_lactis/1-3281 Leishmania_major/1-3286 Macaca_mulatta/1-3182 Magnaporthe_grisea/1-3286 Methanosarcina_acetivorans/1-3043 Monodelphis_domestica/1-3196
Mus musculus/1-3286
Neurospora_crassa/1-3264
Oryza_sativa/1-3218
Ostreococcus_lucimarinus/1-3287
Pan troglodyte/1-3286
Paramecium_tetraurelia/1-3282
Pichia_stipitis/1-3290
_falciparum/1-3285
P_knowlesi/1-3284
Plasmodium_yoelii/1-3285
Populus_trichocarpa/1-3286
Rattus_norvegicus/1-3286
Saccharomyces_cerevisiae/1-3284
Schizosaccharomyces_pombe/1-3284
Strongylocentrotus_purpuratus/1-3252
Takifugu_rubripes/1-2930
etrahymena_thermophila/1-3270
T_annulata/1-3248
Theileria parva/1-3273
Trichomonas vaginalis/1-3198
T_brucei/1-3264
rypanosoma_cruzi/1-3266
Ustilago maydis/1-3291
Yarrowia_lipolytica/1-3278

SGRHKQEESLFE EMLAV'ANAVNPDNII FVMDATI GQACEAQAKAFKEKVDI GSVI ITKLDGHAKGGGALSAVAATNSPIIFIG SGRHKQEA SLFE EMR QVA EATK PDLVI FVMDSSI GQAAFDQAQAFKQSVAVGAVIITKMDGHAKGGGALSAVAATKSPVIFIG SGRHHQEDALFQEMVEIAQEVK PNQT I MV LDASI GQAAEQQSRAFKEAADFGAII LTKMDGHAKGGGAI SAVAATKTPVIFIG SGRHKQEEELFTEMTQI QNAVTPDQTI LVLDSTIGQAAEAQSAAFKATANFGAIIITKTDGHAAGGGAI SAVAATHTPIIFLG SGRHKQEEELFTEMTQIQTAVTPDQTI LVLDSTI GQAAEAQSSAFKATADFGAIIITKTDGHAAGGGAISAVAATHTPI IYLG - - PDN I I FVMDATIGQAC EAQARAFKDKVDIGSVI I TK LDGHAKGGGALSAVAATQS SPIIFIG SGRHKQEDSLFEEMLQVANAI QPDNIVYVMDASIGQACEAQAKAFKDKVDVASVIVTK LDGHAKGGGALSAVAATKSPIIFIG SGRHKQEASLFEEMLQV SNAVT PDNVVFVMDASIGQACEAQARAFSQTVDVASVIITK LDSHAKGGGALSAVAVTKSPVIFIG SGRHKQEASLFE EMLQV SNAVT PDNVVFVMDASI GQAC EAQARAFSQTVDVASVIITK LDSHAKGGGALSAVAVTKSPVIFIG SGRHRQE EQLFT EMVQI GEAVQPTQT I MVMDGSI GQAAESQARAFKESSNFGSI I LTKMDGHAKGGGAISAVAATKTPIVFIG SGRHHQE E ELFHEMVQI SNVIK PNQT I MV LDASI GQAA EQQSKAFKESSDFGAI I LTKMDGHAKGGGAI SAVAATNTPIAFIG SGRHKQEDSLFEEMLQVANAI QPDNIVYVMDASIGQACEAQAKAFKDKVDVASVIVTK LDGHAKGGGALSAVAATKSPIIFIG SGRHKQEDSLFE EMLQVYNAVA PDNVI FVMDASI GQAC EGQARAFKEKVDVASVIVTKLDGHAKGGGALSAVAATQSPIIFIG SGRHKQESELFEEMVAIGAAVKPDMT LMV LDASI GQAAEGQSRAFKDSADFGAI IVTKLDGHAKGGGAI SAVAATKTPI I FLG MEQVVMETN PDDVV FVMDSHI GQACYDQAMAFCNAVDVGSVIITKLDGHAKGGGALSAVAATGAPIIFIG SGRHKQESSLFVEMEQVVMETNPDDVVFVMD SHI GQACYDQAMAFCNAVDVGSVIITK LDGHAKGGGALSAVAATGAPIIFIG SGRHKQEDSLFEEMLQV SNAVQPDNIVYVMDASI GQAC EAQAKAFKDKVDVASVIVTK LDGHAKGGGALSAVAATKSPIIFIG SGRHKQEQSLFNEMI QI SEMIVPTQTI MVMDGSI GQAAESQAKAFKESSQFGSII LTKMDGHARGGGAISAVATTKTPIVFIG SGRHKQDSELFE EMKQI ETAVK PDNCI FVMDSSI GQAAY EQATAFRSSVKVGSIIITKMDGNSMGGGAISAVAATNTPIIFIG SGRHKQE ESLFE EMLAV SNAVSPDNI I FVMDATI GQAC EAQAKAFKDKVDIGSVIITKLDGHAKGGGALSAVAATQSPIIFIG SGRHKQEESLFEEMLAVANAVNPDNI I FVMDATIGQACEAQAKAFKDKVDIGSVIITKLDGHAKGGGALSAVAATQSPIIFIG SGRHTQET ELFTEMKDIIREI SPSSIVFVMDAGIGQSAEDQAMGFKRAVDVGSII LTKIDGTTKAGGAISSVAATKCPIEFVG SGRHKQDK ELFK EMQSVRDAIK PDSII FVMDGAI GQAAFGQAKAFKDAVEVGSVIITK LDGHSNGGGALSAVAATKSPIIFIG SGRHKQEDSLFE EMLQVANAI QPDNI VYVMDASI GQACEAQAKAFKDKVDVASVIVTK LDGHAKGGGALSAVAATKSPIIFIG SGRHKQEDSLFE EMLQVANAI QPDNIVYVMDASI GQAC EAQAKAFKDKVDVASVIVTK LDGHAKGGGALSAVAATKSPI IFIG SGRHQQEDSLFQEMVEI SQAVKPKQTI MV LDASI GQAAEHQSKAFKESADFGSII LTKMDGHARGGGAISAVASTNTPIIFIG SGRHKQESALFE EMK QV E EAVK PNDIVFVMSATDGQAVEEQARNFK EMVAVGSVIVTK LDCQTKGGGALSAVAATRSPIVFIG SGRHKQEDS LFE EMLQVANAI QPDNIVYVMDASI GQACEAQAKAFKDKVDVASVIVTKLDGHAKGGGALSAVAATKSPIIFIG SGRHRQESALFQEMMDI QKAVKPDETI MV LDASI GQQAEAQAKAFKEAADFGAIIITKTDGHASGGGAISAVAATHTPIVFIG AGRHALEADLI E EMERIHAVAK PDHK FMV LDAGI GQQA SQQAHAFNDSVGITGVIITKLDGTAKGGGALSAVSETKAPIAFIG SGRHKQEDSLFEEMLQVANAI QPDNIVYVMDASI GQAC EAQAKAFKDKVDVASVIVTKLDGHAKGGGALSAVAATKSPIIFIG SGRHKQEDS LFE EMLQV SNAI QPDNIVYVMDASI GQACEAQAKAFKDKVDVASVIVTKLDGHAKGGGALSAVAATKSPIIFIG SGRHRQEEALFQEMMDI QTAVKPDETI MV LDASI GQQAEAQAKAFKEAADFGAIIITKTDGHAAGGGAISAVAATHTPIVFIG SGRHKQEAA LFE EMR QV S EATK PDLVI FVMDSSI GQAA FDQAQA FKQSV SVGAV IVTKMDGHAKGGGALSAVAATKSPVIFIG SGRHKQEEALFEEMREIASVTEPTMTI FVMDSSIGQSA SDQAKAFASTVDVGGVIMTKLDGHAKGGGAI SAVSETKAPI LFIG SGRHKQEDS LFE EMLQVANAI QPDNIVYVMDASI GQACEAQAKAFKDKVDVASVIVTK LDGHAKGGGALSAVAATKSPIIFIG SGRHKQEND LFE EMK QV EAAVK PDDIV FVMDSSI GQACFDQA LAFKKAVNVGSVI ITK LDGHAKGGGALSAVAATESPIVFIG SGRHRQEHQLFQEMVQIGEMI QPTQTI MVMDGSI GQAAESQAKAFKESSNFGSII LTKMDGHAKGGGAISAVAATKTPIVFIG SGRHKQENELFEEMI QVENSI QPEEII FVIDSHIGQSCHDQAMAFKNSVSLGSIIITKIDGHAKGGGALSAVAATGCPITFIG SGRHKQESELFEEMKQVESSINPEEIVFVIDSHIGQSCHDQAMAFKNSVTLGSIIITKIDGHAKGGGALSAVASTGCPITFIG SGRHKQENDLFEEMKQVENSIKPEEIVFVIDSHIGQSCHDQAMAFKNSVKVGSIIITKIDGHAKGGGALSAVSAIGCPITFIG SGRHKQEAA LFE EMR QV S EATK PDLI I FVMDSSI GQAAFDQAQAFKQMVAVGAVI ITKMDGHAKGGGALSAVAATKSPVIFIG SGRHKQEDSLFE EMLQV SNA I QPDNIVYVMDASI GQAC EAQAKAFKDKVDVASVIVTK LDGHAKGGGALSAVAATKSPIIFIG SGRHHQEEELFQEMI EI SNVIK PNQTI MV LDASI GQAA EQQSKAFKESSDFGAII LTKMDGHARGGGAISAVAATNTPIIFIG SGRHQQEQELFAEMVEI SDAIR PDQTIMI LDASIGQAAESQSKAFKETADFGAVIITKLDGHAKGGGALSAVAATKTPIVFIG SGRHKQEDS LFE EMLQVANVTSPDNII FVMDASI GQACESQAKAFKEKVDVASVIITKLDGHAKGGGALSAVAATKSPVIFIG SGRHKQEDS LFE EMLQV SNAVQPDNIVYVMDASI GQACESQAKAFKDKVDVASVIVTK LDGHAKGGGALSAVAATRSPIIFIG SGRHKQESELFDEMKQVQAAVNPDECI FVMDGSI GQACYDQAQAFRNAVNVGSVIITKLDGHAKGGGALSAVAATESPIIFIG SGRHKQEDA LFDEMK LI YDAVQPDEVVFVMDSHI GQACYDQA SAFNKAVDVGSVIITKLDGHAKGGGALSAVSATNSPIIFIG SGRHKQEDA LFDEMK LIYDAVQPDEVV FVMDSHI GQACYDQAAAFNKAVDVGSVIITKLDGHAKGGGALSAVSATNSPIIFIG SGRHMQEEALFAEMKA LAAAVNPHEII FVMDGTIGQAAYDQALGFKNAVGVGSIIITKLDSNAKGGGALSAVAATNSPI SFIG SGRHKQESALFE EMK QVQQAVK PNDIVFVMSATDGQGI EEQARQFKEKVPIGSVIVTK LDGQAKGGGALAAVAMTKSPIVFIG SGRHKQESALFE EMK QVQEAVKPNDIVFVMSATDGQGIREQARQFKEKVPVGSVIITK LDGHAKGGGALAAVAMTKSPIVFIG SGRHKQEQELFDEMREI DTAVT PDLTI MV LDANI GQAA EAQSRAFKQAAGYGAI IVTK LDGHAKGGGAI SAVAATKTPIMFIG SGRHRQES ELFTEMVDIGAAVK PDSTI MV LDASI GQAAEPQSRAFKDASDFGSII LTKMDGHAKGGGAISAVAATNTPIIFIG

Anopheles_gambiae/1-3220
Arabidopsis_thaliana/1-3286
Ashbya_gossypii/1-3281
A fumigatus/1-3241
Aspergillus_niger/1-3300
Bombyx_mori/1-2389
Bos_taurus/1-3273
_briggsae/1-3231
Caenorhabditis elegans/1-3289
Candida_albicans/1-3266
Candida_glabrata/1-3282
Canis_familiaris/1-3286
Ciona intestinalis/1-2712
Cryptococcus_neoformans/1-3301 Cryptosporidium_hominis/1-3020 C_parvum/1-3281
Danio rerio/1-3286
Debaryomyces hansenii/1-3166
Dictyostelium_discoideum /1-3275 D_melanogaster/1-3287
Drosophila_pseudoobscura/1-3286
Encephalitozoon cuniculi/1-3199
Entamoeba_histolytica/1-3271
Gallus_gallus/1-3286
Homo_sapiens/1-3286
Kluyveromyces_lactis/1-3281
Leishmania_major/1-3286
Macaca mulatta/1-3182
Magnaporthe_grisea/1-3286
Methanosarcina_acetivorans/1-3043
Monodelphis_domestica/1-3196
Mus musculus/1-3286
Neurospora_crassa/1-3264
Oryza_sativa/1-3218
Ostreococcus_lucimarinus/1-3287
Pan_troglodyte/1-3286
Paramecium_tetraurelia/1-3282
Pichia_stipitis/1-3290
_falciparum/1-3285
P_knowlesi/1-3284
Plasmodium_yoelii/1-3285
Populus_trichocarpa/1-3286
Rattus_norvegicus/1-3286
Saccharomyces_cerevisiae/1-3284
Schizosaccharomyces_pombe/1-3284
Strongylocentrotus_purpuratus/1-3252
Takifugu_rubripes/1-2930
Tetrahymena_thermophila/1-3270
T_annulata/1-3248
Theileria parva/1-3273
Trichomonas vaginalis/1-3198
T_brucei/1-3264
rypanosoma_cruzi/1-3266
stilago_maydis/1-3291
Yarrowia_lipolytica/1-3278

1830
1840
1850
1860

Anopheles_gambiae/1-3220
Arabidopsis_thaliana/1-3286
Ashbya gossypii/1-3281
A fumigatus/1-3241
Aspergillus_niger/1-3300 Bombyx_mori/1-2389
Bos_taurus/1-3273
C_briggsae/1-3231
Caenorhabditis_elegans/1-3289
Candida_albicans/1-3266 Candida_glabrata/1-3282
Canis_familiaris/1-3286
Ciona intestinalis/1-2712
Cryptococcus_neoformans/1-3301 Cryptosporidium_hominis/1-3020 C_parvum/1-3281
Danio_rerio/1-3286
Debaryomyces hansenii/1-3166
Dictyostelium_discoideum /1-3275 D_melanogaster/1-3287
Drosophila_pseudoobscura/1-3286 Encephalitozoon cuniculi/1-3199 Entamoeba_histolytica/1-3271 Gallus_gallus/1-3286
Homo_sapiens/1-3286
Kluyveromyces_lactis/1-3281
Leishmania_major/1-3286
Macaca_mulatta/1-3182
Magnaporthe_grisea/1-3286
Methanosarcina_acetivorans/1-3043
Monodelphis_domestica/1-3196
Mus musculus/1-3286
Neurospora_crassa/1-3264
Oryza_sativa/1-3218
Ostreococcus_lucimarinus/1-3287
Pan troglodyte/1-3286
Paramecium_tetraurelia/1-3282
Pichia_stipitis/1-3290
_falciparum/1-3285
P_knowlesi/1-3284
Plasmodium_yoelii/1-3285
Populus_trichocarpa/1-3286
Rattus_norvegicus/1-3286
Saccharomyces_cerevisiae/1-3284
Schizosaccharomyces_pombe/1-3284
Strongylocentrotus_purpuratus/1-3252
Takifugu_rubripes/1-2930
Tetrahymena_thermophila/1-3270
T_annulata/1-3248
Theileria parva/1-3273
Trichomonas vaginalis/1-3198
T_brucei/1-3264
rypanosoma_cruzi/1-3266
Ustilago maydis/1-3291
Yarrowia_lipolytica/1-3278

## ------- -

 -GFSQD FMTK G -GEQE SMAR I KR LMT MMD SMSDG E LDRVTRVAQGSGVMER EVRDLI GGMNG L'QNMMDYR SQMEL$\qquad$
$\qquad$ SA MMPKG-H EK E SQAK I KRYMT MMD SMTND E LDRMMR I ARGSGRQVR EVMEMLGGMGGLQSLMDFT-DLEL - . . -- ------- P-GFGTDFMSKG-NEQESMAR LKKLMTI MDSMNDQELDRIQRVARGSGVSTRDVQELLGGMAGLQSMMDYR LQMPL -------P-GFGPEFMNKG-NEQESVNR LKRMMTVMDSMSDK ELDRVARVARGSGC FQQEVRDLLGGMGGLQNMMDYRKDMPI ------- - - GFGSEFMTKG-NEQESVNR LKRMMTVMD SMSDK ELDRVARVARGSGSHQQEVRDLLGGMGGLQNMMDYRKDMPL - - - - - - P-GLSN-IMSQV-GDEETSKKIKNMI YI MDSMTIKELERIVRVARGSGCAVVEVEMI LGG---LAGMMDFS-YLKL ------- P-GMGN-MMSQV-GEEETSQKMKKMVYV LDSMTK EELERLVRVARGSGTSVFDVEMI LGGMPDMNEMMDFS-Y LKL ------ P-GFGTDFMSKG-NEQESMAR LKK LMTI MD SMNDQELDRI QRVARGSGVSTRDVQELLGGMAGLQSMMDYR LQMPL
$\qquad$ .... _ _ _ GMG PNI LAK E-DEQAGI ER LKK FMVI MD SMT ESELDRVERI SRGSGT SNQDVQELLGGAQNMMNMMDYSSELKL- - . - . - -------- ---------- - P

-     - . - . - -GFGTDFMSKG-NEQESMAR LKK LMTI MD SMNDQE LDRI QRVARGSGVATRDVQELLGGMAG LQSMMDYR LQMLLGMSN-I MSQV-GEEETSSKIKNMI YI MDSMTTKELDRI IRVARGAGCSAVEVEMVLGGMPDMSSMMDFS-Y LKL-GMNQ--LPQL-QGNEGGLKLKAYINI LDSLSEKELDRIITIAQGSGRHPNEVVELLGGMPSMGDLADYSKRCILG F SQD FMTKG-G EQE SMARVKRMMT MMD SMSDN E LDRCVRVAQGAGVMEREVKELIGGVGGI QNMMDYR SQMQLG F SQD FMTK G-GEA ESMARI KR MMT MMD SMSDNE LDRCTRVAQGAGV LER EAK ELI GGVGG LQNMMDYRGQMQLGFSG------LSLPDEDT FKKLI YVFDSLSRGELDRIMRVARGSGTSVQGVVEI LGSMF
$\qquad$ -PYE-ELV
 -G LNHPMFQGG-NI E---KK FKV FMVI LDSMTDRELDRIRRLARGSGRDIREVNELF-NQAQLQQLMDY--DMRL -------P-GFGTDFMSKG-NEQESMAR LKKLMTIMDSMNDQELDRI QRVARGSGVSTRDVQELLGGMAGLQSMMDYR LQMPL -------P-GMGN-MLNQF-SEEETSKKMKTMVYI FDSMTKKELERLVRVAKGSGTTVFDI EMLLGGMPNMQDIMDFS-YLKL
$\qquad$
$\qquad$
$\qquad$
$\qquad$ M -GMSN -MMQGM-DDEEGTGK LKRMIYI CDSMTDK E LDRMTRVARGSGTHVREVEDLLGGM-DMAAMMDYS -Y LK L - - - - - - PMGMGGMK F SDE - MFQAT SDKMKNYKVI MD SMT EEEMTRIKRISKGSGCSSEDVRELLGGKFNIQKMMEISFK-------- - - GFGTDFMSKG-NEQESMAR LKK LMT I MD SMNDQE LDR I QRVARGSGVSTRDVQELLGGMAGLQSMMDYR LQMPL
$\qquad$
$\qquad$
$\qquad$
------- $P$
------- - P
------- -- - - - - - - P FGTDFMSKG NEQESMAR LKKLMTI MD SMNDQELDRI QPVARGSGVSRDVQELLCGMACLOSMMDYRQQMP GMSN -MMQNM - DD E EGS LK LKRMI Y I CD SMT DK E LDR MTRVARGSGTTVREVEDLLGGM-DMNAMMDFS -Y LNL GFS S ELMPKG-HEK ESQAK I KRYMT MMD SMT DG ELDR I LRIARGSGR PVRDVVDMLGGMGGLQSLMDFT-KLEL GMAD -MI PKG-GE EQGTKRMK SMMV LMD SMTDA ELDRMERVCRGAGRLPSEMITLLGGQQGMMNMMDFS-ELEL GFGTDFMSKG-NEQESMAR LKK LMT I MD SMNDQE LDR I QRVARGSGVSTRDVQELLGGMAGLQSMMDYR LQMPL -GMGS SV LSKG -NEK ESI KRI QR FLCI MN SMTADELDRIVRIAKGSGTSI EEVHI LLGGMDNVMNMMDYR---NI GLSN-I MSQV -GEEETSKKI KNMVYI MDSMTT EELERI LRVAKGSGCAAVEI EMVLGGMPNMANMMDYS-Y LKL -GFGNNLI SKG-TEK EGIDKI KK FMVI MD SMTNEELDRCLRIVKGSGTR LQDIKELLGGANNMVNI LDY SKDMKL
$\qquad$
$\qquad$
$\qquad$ GFGNNI I SKG-TEKEGI EKIKKFMVI MD SMTNEELDRC LR GFGTNLI SKG-TEKEGIDKIKKYMVI MDSMTNEELDRCIR GFAELMPK HEK SQAKIKRYMTMMD SMTNEELDR GFSA ELMPKG-HEK E SQAK I KRYMT MMD SMTNEELDRMMRIARGAGRPIRDVMEI LGGMGGLQNLMDFS-KLEL -GFGTDFMSKG -NEQESMAR LKK LMT I MD SMNDQE LDR I QRVARGSGVSTRDVQELLGGMAG LQSMMDYRQQMPL
$\qquad$ -GMSN -MMNQV -GEEETSQKMKKMVYV LDSMTK EELERMVRVAKGSGT SVFEVEMI LGGMPDMNEMMDFS-Y LRL
$\qquad$ -GMSN -MMNGM-NDEEGS LRMKRMLYIVDSMTEQELDRVLRVARGSGTSVLEVEETIGGM-DFSGMLDFSNLLGL
$\qquad$ -GFS SDFMTKG-NEQESMAR LKK LMTMMD SMKDEE LDRVQRVARGSGVSVR EVQELLGGMNGLQSMMDYRGQMEL
$\qquad$ -GFGTDFMSKG - NEQESMAR LKK LMT I MD SMNDQE LDR I QRVARGSGVATRDVQELLGGMAGLQSMMDYR SQMQM -GI PPELLQAG-R EQEGVDRI KRFMI I MD SMTDEELDRIMRIAKGSGSSPHEINFLIGGAGNI MK LMDY SN - LKL -GI PPELLQAG-R EQEGVDRIKR FMI I MDSMTDEELDRIMRIAKGSGSSPHEI SFLIGGAGNIMK LMDYSN -LKL
------
TIAQGIK GMKN --GMSGFTGNAG-DPEHTKDTVKRIVVIDAMSTSEI ERIAR LCSGSGMPPPFVGYVIGGIEGLEASMKYFKDLYI -------P-GMSGFTGNAG---DAGDVTLKTFI HMMDSMTAAELDRIHRIARGSGHTI LEVHNLIGG LTGLQDIM


MGQMLSGAGGDD EAACSKMKRMMFIDAMTAE LDR ARRVARGSGT SVK EVEEFLGGMPDSOLADFT-KMPL GLSG-MASSI - SDEEGTRRIKRMIYI LDSMNQKELDRITRVARGSGTSIREVEEVLGGMPDMGQMMDFS-YLKL

Anopheles_gambiae/1-3220 Arabidopsis_thaliana/1-3286
Ashbya_gossypii/1-3281
A fumigatus/1-3241
Aspergillus_niger/1-3300 Bombyx_mori/1-2389
Bos_taurus/1-3273
C_briggsae/1-3231
Caenorhabditis_elegans/1-3289
Candida_albicans/1-3266 Candida_glabrata/1-3282
Canis_familiaris/1-3286
Ciona intestinalis/1-2712
Cryptococcus_neoformans/1-3301 Cryptosporidium_hominis/1-3020 C_parvum/1-3281
Danio_rerio/1-3286
Debaryomyces hansenii/1-3166
Dictyostelium_discoideum /1-3275 D_melanogaster/1-3287
Drosophila_pseudoobscura/1-3286 Encephalitozoon cuniculi/1-3199 Entamoeba_histolytica/1-3271 Gallus_gallus/1-3286
Homo_sapiens/1-3286
Kluyveromyces_lactis/1-3281
Leishmania_major/1-3286
Macaca_mulatta/1-3182
Magnaporthe_grisea/1-3286
Methanosarcina_acetivorans/1-3043
Monodelphis_domestica/1-3196
Mus musculus/1-3286
Neurospora_crassa/1-3264
Oryza_sativa/1-3218
Ostreococcus_lucimarinus/1-3287
Pan troglodyte/1-3286
Paramecium_tetraurelia/1-3282
Pichia_stipitis/1-3290
_falciparum/1-3285
P_knowlesi/1-3284
Plasmodium_yoelii/1-3285
Populus_trichocarpa/1-3286
Rattus_norvegicus/1-3286
Saccharomyces_cerevisiae/1-3284
Schizosaccharomyces_pombe/1-3284
Strongylocentrotus purpuratus/1-3252
Takifugu_rubripes/1-2930
Tetrahymena_thermophila/1-3270
T_annulata/1-3248
heileria parva/1-3273
Trichomonas_vaginalis/1-3198
T_brucei/1-3264
rypanosoma_cruzi/1-3266
Ustilago maydis/1-3291
Yarrowia_lipolytica/1-3278

# 2000 

2040
2060


## 2080

Anopheles_gambiae/1-3220
Arabidopsis_thaliana/1-3286
Ashbya_gossypii/1-3281
A fumigatus/1-3241
Aspergillus_niger/1-3300 Bombyx_mori/1-2389
Bos_taurus/1-3273
_briggsae/1-3231
Caenorhabditis elegans/1-3289
Candida_albicans/1-3266
Candida_glabrata/1-3282
Canis_familiaris/1-3286
Ciona intestinalis/1-2712
Cryptococcus_neoformans/1-3301 Cryptosporidium_hominis/1-3020 C_parvum/1-3281
Danio rerio/1-3286
Debaryomyces hansenii/1-3166
Dictyostelium_discoideum /1-3275 D_melanogaster/1-3287 Drosophila_pseudoobscura/1-3286 Encephalitozoon_cuniculi/1-3199 Entamoeba histolytica/1-3271 Gallus_gallus/1-3286
Homo_sapiens/1-3286
Kluyveromyces_lactis/1-3281
Leishmania_major/1-3286
Macaca mulatta/1-3182
Magnaporthe_grisea/1-3286
Methanosarcina_acetivorans/1-3043
Monodelphis domestica/1-3196
Mus musculus/1-3286
Neurospora_crassa/1-3264
Oryza_sativa/1-3218
Ostreococcus_lucimarinus/1-3287
Pan troglodyte/1-3286
Paramecium tetraurelia/1-3282
Pichia_stipitis/1-3290
_falciparum/1-3285
P_knowlesi/1-3284
Plasmodium_yoelii/1-3285
Populus_trichocarpa/1-3286
Rattus_norvegicus/1-3286
Saccharomyces_cerevisiae/1-3284
Schizosaccharomyces_pombe/1-3284
Strongylocentrotus purpuratus/1-3252
Takifugu_rubripes/1-2930
Tetrahymena_thermophila/1-3270
T_annulata/1-3248
Theileria parva/1-3273
Trichomonas_vaginalis/1-3198
T_brucei/1-3264
rypanosoma_cruzi/1-3266
Ustilago maydis/1-3291
Yarrowia_lipolytica/1-3278

2090
2100
2120
2130
2140
2150
SKIHA LSYGKKLVLKHNKYFVEFEVNQEKI-EV LQKRCI--EI EFPLLAEYDFRNDTIN----ADINIDL SKIKGATVSYGKKLVIKKNRYFIEFEIDPALV-ENVKQRCLPNALNYPMLEEYDFRNDNVN----PDLDMELKPH---AQPR SKIIDFTKSFGKKVVLKHRRFFVEFEI PNEAV-ESVKARCQ--AMGCPALEEYDFRNDEIN----PTLDIDLKPN---ARIRS SKI LDFTK SYGKKVV LKHNRFFVEFEI PNEAV-EPVKARCQ--AMGCPALEEYDFRNDEIN----PTLDIDLKPA---ARIRS ------------------Y LVEFEVDPDKI-EVI QKRCI--ELEHPLLAEYDFRNDSIN----PDINIDLKPT---AVLRP SKIKLCTVSYGKKLVLKHNRYFVEFEVKQEMI-EELQKRCI--HLEYPLLAEYDFRNDSVN----PDINIDLKPT---AVLRP SKI QMCTV SYGKKLVLKHNRYFVEFEIKQETI-ETVQRRCI--ELEYPLLAEYDFRNDTMN----PNLGIDLKPS---TTLRP SKVQMCTV SYGKKLVLKHNRYFVEFEIKQETI-ETVQKRCI--ELEYPLLAEYDFRNDTLN----PNLGIDLKPS---TTLRP SKIKAATVSYGKKLVLKHNRYFVEFEIAHDSV-EIVKRRCQ--DIEYPVLEEYDFRHDARN----PDLEIDLKPS---TQIRP SKIKGATI SYGKKLVIKHNRYFVEFEIANSSV-EIVKRRCQ--EIDYPVLEEYDFRNDNRN----PDLEIDLKPS---TQIRP SKIKLCTVSYGKKLVLKHNRYFVEFEVKQEMI-EELOKRCI--HLEYPLLAEYDFRNDSVN----PDINIDLKPT---AVLRP

SKIEEWTASFGKRLVLKDNRYFLEFEVSGERM-EDVRRRCK--DIDLPALEEYDFRNDTIN----PNLDIQLKPM---TVIRP SKIRGHCKLFGKKIVLLEGRYFVEFEISGDKV-DIVTMASF-VSLHRPLLSEYDFRSDIKN----PNLDISLKHT---TQIRY SKIRGHCKLFGKKIVLLEGRYFVEFEI SGDKV-DIVTMASF-VSLHRPLLSEYDFRSDIKN----PNLDISLKHT---TQIRY SKIKLCTV SYGKKLVLKHNRYFVEFEIRQEMI-EELQKRCI--QLEYPLLAEYDFRNDTVN----PDINMDLKPT---AVLRP SKIRGATI SYGKKLVLKHRYYVEFEIANESV-EIVKRRCQ--EI EYPVLEEYDFRNDDRN----PDLDIDLKPS---TQIRP SKVR QCTQSYGKKLVLQKNKYFVEFEIDPQQV-EEVKKRCI--QLDYPVLEEYDFRNDTVN----PNLNIDLKPT---TMIRP SKIRLCTLSYGKKLVLKHNKYFIEFEVAQEKI-EVI QKRCI--EI EHPLLAEYDFRNDTNN----PDINIDLKPA---AVLRP SKIRLCTLSYGKKLVLKHNKYFIEFEVSQEKI-EVI QKRCI--EI EHPLLAEYDFRNDTNN----PDINIDLKPA---AVLRP SKITECTLSYGKKLVMKESSFFLELSIEVEEV-ELVKKRCI--EIDYPLIEEYDFRNDKVL----RSLQIDLKPT---TIIRS SKIRQAGEKKQNRLVLINGKYYLQI EVKQTSV-FKLKKKCK--KKKVRVYEEYHFLRDKQ-----KELPIQLRKD----CLRP SKIKLCTV SYGKKLVLKRNRYFVEFEVKQEMI-EELQKRCI--HLDYPLLAEYDFRNDSVN----PDINIDLKPT---AVLRP SKIKLCTV SYGKKLVLKHNRYFVEFEVKQEMI-EELQKRCI--HLEYPLLAEYDFRNDSVN----PDINIDLKPT---AVLRP SKIK SATVSYGKKLVIKHNRYFVEFEIDNASV-EIVKKRCQ--ELDYPVLEEYDFRNDRRN----PDLDIDLKPS---TQIRP AEI ESCMKRYNLRIIIDAERTLVQFLLQSRAMSKVVAAQCV--VLGLPI QQQYDFENDTSV----RTAHISLRTQ---TKPRR SKIKLCTV SYGKKLVLKHNRYFVEFEVKQEMI-EELQKRCI--HLEYPLLAEYDFRNDSVN----PDINIDLKPT---AVLRP LKIEVSTKSYGKK LV LKNTQYFVEFQIEDEGV-EIVQKRCL--ELNYPI LEEYDFRNDT FN----PVLDIDLRPN---TQVRP
SKIKLCTV SYGKKLVLKHNRYFVEFEVKQEMI-EELOKRCI--HLEYPLLAEYDFRNDSVN----PDINIDLKPT---AVIR SKIKLCTV SYGKKLVLKHNRYFVEFEVKQEMI-EELQKRCI--CLEYPLLAEYDFRNDTLN----PDINIDLKPT---AVLRP LKIESCTKSYGKKLVLNNNKYFVEFEIPETAV-EIVQRRCL--DLGFPILEEYDFRNDSNN----ADLEIDLRPN---TQIRP SKIHASTANYGKKLVLKKNRYFVEFEIDPSQV-ENVK QRCLPNALNFPMLEEYDFRNDTVN----PDLEMELKPQ---ARPRP SKVHECTENYGKKLVLQRNKFYLEFEI EARQV-EHVK QRCLPGNLGYPTLEEYDFRNDTRN----PDLGIELKPM---TRIRP SKIKLCTVSYGKKLVLKHNRYFVEFEVKQEMI-EELQKRCI--HLEYPLLAEYDFRNDSVN----PDINIDLKPT---AVLRP SKIRHHTNNIGQKFFLQDK SYYIDFRIVGDYF--DVAQALI--RSSVPLIQEYDFTKEK------QKLDINLKPS---TKPRL SKIKSATI SYGKKLVLKHNRYFVEFEIANESV-EIVKRRCQ--DIDYPVLEEYDFRNDARN----PDLEIDLKPS---TQIRP SKITKSAESFGKKLVLRENKYYI EFEVNCDKI-EEVKQEAL-QTMQRPLLMEYDFRRDKKN----PNLICSLKSH---VQIRY SKITKSAESFGKKLVLRENKYYIEFEVNCDKL-EEVKQEAL-QTMQRPLLMEYDFRRDKKN----PNLICSLKSH---VQIRY SKITKSAESFGKKLVLRENKYYIEFEVNCDKI-EEVKQEAL-QTMQRPLLMEYDFRRDKKN----PNLNCSLKSH---VQIRY SKIHGSTANYGKKLVLKKNRYFIEFEVDPSQV-ENVK QRCLPNALNYPMLEEYDFRNDTVN----PDLNMELKPH---AQPRP SKIKLCTV SYGKKLVLKHNRYFVEFEVKQEMI-EELQKRCI--CLEYPLLAEYDFRNDSLN----PDINIDLKPT---AVLRP SKIKGATI SYGKKLVIKHNRYFVEFEIANESV-EVVKKRCQ--EIDYPVLEEYDFRNDHRN----PDLDIDLKPS---TQIRP SKIRACTV SYGKKLVLKKNRYFIEFEIKHSSV-ETIKKRCA--EIDYPLLEEYDFRNDNIN----PDLPIDLKPS---TQIRP SKIKLCTLSYGKKLVLKHNRYFVEFEVVQDEI-ENLQKRCI--ELEYPLLAEYDFRNDTRN----PDLSIDLKPT---AVLRP SkIKv
SKITENTQNYGARLFLDDSSYYLDIQIIGDHF--EVTKAVI--NCSVPLIQEYDFENKSF-----KQLEIELKPK ---IKVRY SKILNTSSAFGKKLVLRDSRYWIEFEVQQEKI-EDLKREAL-QTMRRPLVMEYDFRKDNNS----PSLNCCIRSN---IKIRY SKILNTSSAFGKKLVLRDSRYWIEFEVQQEKI-EELKREAL-QNMRRPLVMEYDFRKDNNS----PSLNCCIRSN---IKIRY SKIELCCLSVGKKSVLRNTKYYIEFQIKTESV-REIRQYAV--DHNLFI SDEYDFMNDKTI----DNLGIQLKNT---TRIRP EQAKVSANGDVK--VKKEETK EVASQVMDGKM-RNVRERLY-KELSVRADLFYDYVQDHSL----HVCDLELSEN---VRLRP EQTHESEESQGKSLVKQEATEETASQVKDGRL-RNVRERLF-KELGVRADLFYDYVQDGTL----DVRDLALAEH---VRLRP SKIREYTASFGKKLVLKQNKYFVEFEIAEEYI-EQVKKRCN--EIGYPMLEEYDFRNDQLN----ADLEIDLKPI---THIRP SKIHSCTKSYGKKLVLKHRYYFEFEIAPDSV-ETVKKRCQ--EIDYPVLEEYDFRNDHGN----PDLDIDLKSS---TQIRP

2160
2170
2180
2190
2200
2210
2220
2230

Anopheles_gambiae/1-3220 Arabidopsis_thaliana/1-3286 Ashbya_gossypii/1-3281
A fumigatus/1-3241 Aspergillus_niger/1-3300 Bombyx_mori/1-2389
Bos_taurus/1-3273
C_briggsae/1-3231
Caenorhabditis_elegans/1-3289
Candida_albicans/1-3266
Candida_glabrata/1-3282
Canis_familiaris/1-3286
Ciona intestinalis/1-2712
Cryptococcus_neoformans/1-3301 Cryptosporidium_hominis/1-3020 C_parvum/1-3281
Danio_rerio/1-3286
Debaryomyces_hansenii/1-3166
Dictyostelium_discoideum /1-3275 D_melanogaster/1-3287
Drosophila_pseudoobscura/1-3286 Encephalitozoon cuniculi/1-3199 Entamoeba_histolytica/1-3271 Gallus_gallus/1-3286
Homo_sapiens/1-3286
Kluyveromyces_lactis/1-3281 Leishmania_major/1-3286 Macaca_mulatta/1-3182 Magnaporthe_grisea/1-3286 Methanosarcina_acetivorans/1-3043 Monodelphis_domestica/1-3196
Mus musculus/1-3286
Neurospora_crassa/1-3264
Oryza_sativa/1-3218
Ostreococcus_lucimarinus/1-3287
Pan_troglodyte/1-3286
Paramecium_tetraurelia/1-3282
Pichia_stipitis/1-3290
_falciparum/1-3285
P_knowlesi/1-3284
Plasmodium_yoelii/1-3285
Populus_trichocarpa/1-3286
Rattus_norvegicus/1-3286
Saccharomyces_cerevisiae/1-3284
Schizosaccharomyces_pombe/1-3284
Strongylocentrotus_purpuratus/1-3252
Takifugu_rubripes/1-2930
Tetrahymena_thermophila/1-3270
T_annulata/1-3248
Theileria parva/1-3273
Trichomonas vaginalis/1-3198
T_brucei/1-3264
rypanosoma_cruzi/1-3266
stilago maydis/1-3291
Yarrowia_lipolytica/1-3278
 YQEKSLSKMFGNGRARSGIIVLPCGAGKSLVGVSAAARIKKSCLCLATNAVSVDQWAYQFWGLLLMDEVHVVPAHMFRKVISI YQEK SLSKMFGNGRARSGIIVLPCGAGKTLVGITAACTIKKSVIVLCTSSVSVMQWRQQFWGFILLDEVHVVPAAMFRRVVST YQEKSLSKMFGNGRAKSGIIVLPCGAGKTLVGITAACTIKKGTIVLCT SSMSVVQWRNEFWGLMI LDEVHVVPASMFRKVTSA YQEK SLSKMFGNGRAK SGIIVLPCGAGKT LVGITAGCTIKKGTIVLCT SSMSVVQWRNEFWGLMI LDEVHVVPASMFRKVTSA YQEKSLRKMFGNGRARSGVIVLPCGAGKSLVGVTAVCTVRKRALVLCNSGVSVEQWKQQFWGLVVLDEVHTI PAKMFRRVLTI YQEK SLRKMFGNGRAR SGVIVLPCGAGK SLVGVTAACTVRKRCLVLGNSAVSVEQWKAQFWGLMI LDEVHTI PAKMFRRVLTI YQEK SLRKMFGNSRAR SGVIVLPCGAGKTLVGVTAVTTVNKRCLVLANSNV SVEQWRAQFWGLLLLDEVHTI PAKMFRRVLTI YQEKSLRKMFGNSRARSGVIVLPCGAGKTLVGVTAVITVNKRCLVLANSNV SVEQWRAQFWGLLLLDEVHTI PAKMFRRVLTI YQEKSLSKMFGNGRARSGIIVLPCGAGKTLVGITAACTIRKSVIVLCTSSVSVMQWRQQFWGFIILDEVHVVPAQMFRRVVTT YQEKSLSKMFGNGRARSGIIVLPCGAGKTLVGITAACTIKKSVIVLCTSSVSVMOWRQQFWGFIILDEVHVVPAAMFRRVVST YQEK SLRKMFGNGRAR SGVIVLPCGAGK SLVGVTAACTVRKRCLVLGNSAV SVEOWKAQFWGLMI LDEVHTI PAKMFRRVLTI

YQEMSLAKMFGNGRARSGIIVLPCGAGKTLVGITAACTIKKSALVLCT SAVSVAQWKQQFWGFLLLDEVHVTPADMFRKCINN YQEQALRMMF SNGRAR SGIIVLPCGAGKTLTGITAACTMRKSILI LTTTSAVAVSQWKFQFWGLLIFDEVQFAPAPAFRRINGI YQEQALRMMFSNGRARSGIIVLPCGAGKTLTGITAACTMRKSVLI LTTTSAVAVSQWKFQFWGLLIFDEVQFAPAPAFRRINGI YQEKSLRKMFGNGRARSGVIVLPCGAGKSLVGVTAACTVRKRCLVLGNSSVSVEQWKAQFWGLII LDEVHTI PAKMFRRVLTI YQEKSLSKMFGNGRARSGIIVLPCGAGKTLVGITAACTIRKSVIVLCTSSVSVMQWRQQFWGFIILDEVHVVPAAMFRRVVTT YQEKSLSKMFGNGRARSGIIVLPCGAGKSLSGITAACTVKKSI LVLCTSAVSVEQWKYOFWGLVLLDEVHVVPAAMFRKVLTV YQEK SLRKMFGNGRARSGVIVLPCGAGKSLVGVTACCTVRKRALVLCNSGVSVEQWKQQFWGIMVLDEVHTI PAKMFRRVLTI YQEKSLRKMFGNGRARSGVIVLPCGAGKSLVGVTACCTVRKRALVLCNSGVSVEQWKQAFWGIMVLDEVHTI PAKMFRRVLTI YQEICLNKMFGNGRARSGIIVLPCGSGKTIVGITAISTIKKNCLVLCTSAVSVEQWKQQTWGLLVLDEVHVVPAMMFRRVLSL HQERALQQI FDNEMAR SGIVVLPCGAGKTLTAIAACSKIKRSTIVLTHTTQSVFQWKEEFWGFII FDEVHGSTTDNI EK FVCK YQEK SLRKMFGNGRAR SGVIVLPCGAGK SLVGVTAACTVRKRCLVLGNSAVSVEQWKAQFWGLMI LDEVHTI PAKMFRRVLTI YQEK SLRKMFGNGRAR SGVIVLPCGAGK SLVGVTAACTVRKRCLVLGNSAV SVEQWKAQFWGLMI LDEVHTI PAKMFRRVLTI YQEKSLSKMFGNGRARSGIIVLPCGAGKTLVGITAACTIKKSVIVLCTSSVSVMQWRQQFWGFIILDEVHVVPAAMFRRVVST YQIEAVDAAIHDGTLNSGCLLLPCGAGKTLLGIMLMCKVKKPTLVLCAGSV SVEQWKSQIYGLLI LDEVHVMPAESFRGSLGF YQEK SLRKMFGGGRARSGVIVLPCGAGKSLVGVTAACTVRKRCLVLGNSAV SVEQWKAQFWGLMI LDEVHTI PAKMFRRVLTI YQEKSLSKMFGNGRAKSGIIVLPCGAGKTLVGITAACTIKRGVIVLCTSTMSVVQWRDEFWGLMI LDEVHVAPAKMFRRVTSA YQAEALVAWSEN--EKWGVLVLPTGSGKTLLGIRAIAGCNTPALVIVPTLDLLEQWKTQLFGLLVFDEVHHLPAAGYRSIAEF YQEK SLRKMFGNGRARSGVIVLPCGAGK SLVGVTAACTVRKRCLVLGNSAVSVEQWKAQFWGLMI LDEVHTI PAKMFRRVLTI YQEK SLRKMFGNGRARSGVIVLPCGAGKSLVGVTAACTVRKRCLVLGNSAV SVEQWKAQFWGLMI LDEVHTI PARMFRRVLTI YQEQSLSKMFGNGRAKSGIIVLPCGAGKTLVGITAACTIKKGVIVLCTSSMSVVQWRQEFWGLMLLDEVHVVPADVFRRVISS YQEKSLSKMFGNGRARSGIIVLPCGAGKSLVGVSAACRIKKSCLCLATNAV SVDQWAFQFWGLLLMDEVHVVPAHMFRKVISI YQEKSLSKMFGNGRARSGIIVLPCGAGKSLTGIAAAARIRKSCLCLCTSSVSVDQWAQFWGCMLLDEVHVVPAAMFRKVIGI YQEK SLRKMFGNGRAR SGVIVLPCGAGK SLVGVTAACTVRKRCLVLGNSAV SV EQWKAQFWG LMI LDEVHTI PGKQAGAELRV YQLRAAKTVIMGDYAKSGLIVLPCGAGKTLVGVLCMSLIKSSTVIICDSNVSVEQWKREIWGICIVDEVHRLPAVQFQNVLKQ YQEKSLSKMFGNGRARSGIIVLPCGAGKTLVGITAACTIRKSVIVLCTSSVSVMQWRQQFWGFIILDEVHVVPAAMFRRVVTT YQEKALRKMFSNGR SRSGIIVLPCGVGKTLTGITAASTIKKSALFLTT SAVAVEQWKKQFWGLLVFDEVQFAPAPSFRRINDI YQEKALRKMFSNGRSRSGIIVLPCGVGKTLTGITAASTIKKSSLFLTTSAVAVEQWKKQFWGLLVFDEVQFAPAPSFRRINDI YQEKALRKMFSNGRSRSGIIVLPCGVGKTLTGITAASTIKKSSLFLTT SAVAVEQWKKQFWGLLVFDEVQFAPAPSFRRINDI YQEKSLSKMFGNGRARSGIIVLPCGAGKSLVGVSAACRIKKSCLCLAINAVSVDQWAFQFWGLLLMDEVHVVPAHMFRKVISI YQEKSLRKMFGNGRARSGVIVLPCGAGKSLVGVTAACTVRKRCLVLGNSAVSVEQWKAQFWGLMI LDEVHTI PAKMFRRVLTI YQEKSLSKMFGNGRARSGIIVLPCGAGKTLVGITAACTIKKSVIVLCTSSVSVMQWRQQFWGFIILDEVHVVPAAMFRRVVST YQEKSLSKMFGNGRARSGIIVLPCGAGKTLVGITAACTIKKSVIVLCTSSVSVMQWRQQFWGFILLDEVHVVPAAMFRRVVTT位

YQERALKNIFIQKKARSGLIILPCGAGKTIVGVIAIERIKQSTVIICDSDVSVDQWRDELWGVCVIDEVHKLPANTFQNVLKQ YQERALRRMFSNGRARSGIIVLPCGAGKTLTGIVAACTVRKSI FVLTTSAVAVEQWIKQFWGMLIFDEVQFVPAPAFRRINEI YQERALRRMFSNGRARSGIIVLPCGAGKTLTGIVAACTVRKSIFVLTTSAVAVEQWIKQFWGMLIFDEVQFVPAPAFRRINEI YQEKALTKMFSGGRSISGIIVLPCGAGKTLVGIAALATINKPTVIVCNNRLTVKQWYNQIWGLLI LDEVQDSAANTFRNVTDI YQVASLERFRSGNKAHQGVIVLPCGAGKTLTGIGAAATVKKRTIVMCINVMSVLQWOREFWGLLLLDEVHTALAHNFQEVLNK YQVASLERFRCGNKAHQGVIVLPCGAGKTLIGIGAATI LKKRTIVMCINVISVLQWQREFWGLLLLDEVHAALAHHFQEV LNK YQEKSLAKMFGNGRARSGIIVLPCGAGKTLVGITAACTIKKSCLVLCTSSVSVMQWRQQFWGFILLDEVHVVPASMFRRVLTK YQEKSLSKMFGNGRARSGIIVLPCGAGKTLVGITAACTIRKSVIVLCT SSVSVMQWRQQFWGFII LDEVHVVPAAMFRKVVTN

# 2250 <br> 2260 <br> 2270 <br> 2280 <br> 2290 <br> 2300 <br> 2310 

Anopheles_gambiae/1-3220
Arabidopsis_thaliana/1-3286
Ashbya_gossypii/1-3281
A fumigatus/1-3241
Aspergillus_niger/1-3300 Bombyx_mori/1-2389
Bos_taurus/1-3273
C_briggsae/1-3231
Caenorhabditis_elegans/1-3289
Candida_albicans/1-3266
Candida_glabrata/1-3282
Canis_familiaris/1-3286
Ciona intestinalis/1-2712
Cryptococcus_neoformans/1-3301 Cryptosporidium_hominis/1-3020 C_parvum/1-3281
Danio_rerio/1-3286
Debaryomyces_hansenii/1-3166
Dictyostelium_discoideum /1-3275 D_melanogaster/1-3287
Drosophila_pseudoobscura/1-3286 Encephalitozoon_cuniculi/1-3199 Entamoeba_histolytica/1-3271 Gallus_gallus/1-3286
Homo_sapiens/1-3286
Kluyveromyces_lactis/1-3281 Leishmania_major/1-3286 Macaca_mulatta/1-3182 Magnaporthe_grisea/1-3286 Methanosarcina_acetivorans/1-3043 Monodelphis_domestica/1-3196
Mus musculus/1-3286
Neurospora_crassa/1-3264
Oryza_sativa/1-3218
Ostreococcus_lucimarinus/1-3287
Pan troglodyte/1-3286
Paramecium_tetraurelia/1-3282
Pichia_stipitis/1-3290
_falciparum/1-3285
P_knowlesi/1-3284
Plasmodium_yoelii/1-3285
Populus_trichocarpa/1-3286
Rattus_norvegicus/1-3286
Saccharomyces_cerevisiae/1-3284
Schizosaccharomyces_pombe/1-3284
Strongylocentrotus_purpuratus/1-3252 Takifugu_rubripes/1-2930
Tetrahymena_thermophila/1-3270
T_annulata/1-3248
Theileria parva/1-3273
Trichomonas_vaginalis/1-3198
T_brucei/1-3264
Trypanosoma_cruzi/1-3266
stilago_maydis/1-3291
Yarrowia_lipolytica/1-3278

VQSHCKLGLTATLLREDDKIADLNFLIGPKLYEANWLELQKRGYIARVQCAEVWCPMAPEFYREYLYVMNPAKFRACQYLIRY TK SHCK LGLTAT LVREDEKITDLNFLIGPK LY EANWLDLVKGGFIANVQCAEVWCPMTKEFFAEY LYVMNPNKFRACEFLIRF I AAHAK LGLTAT LVREDDKI SDLNFLIGPK LY EANWMELSQKGHIANVQCAEVWCPMTAEFYQEYLYIMNPTKFQACQFLI QY I ATQSKLGLTATLLREDDK I KDLNFLIGPKLY EANWMELAEQGHIAKVQCAEVWCPMTTEFYTEYLYIMNPRKFQACQFLIDY I ACQSK LGLTAT LLREDDK IKDLNFLIGPK LY EANWME LA EQGHI AKVQCAEVWC PMTTEFYSEYLYI MNPRKFQACQFLIDY VHSHAKL
VQAHCK LGLTAT LVREDDKIVDLNFLIGPKLY EANWME LQN SGYI AKVQCAEVWCPMSPEFYREYLYTMNPNKFRACQFLIKF VQAHCK LGLTAT LVREDDK I TDLNFLIGPKI YEANWMELQKAGHI AKVQCAEVWCPMT SAFY SYY LAVMNPNKFRICQFLIKF VQAHCK LGLTAT LVREDDK I TDLNFLIGPKIY EANWME LQKAGHI AKVQCAEVWCPMT SAFY SYY LAVMNPNKFRICQFLIKF I AAHAK LGLTAT LVREDDK I DD LNFLIGPK LY EANWMD LAQKGHIANVQCA EVWC PMTAEFYQEYLYIMNPTKFQACQFLIHY I AAHAK LGLTAT LVREDDK I SDLNFLIGPK LY EANWMELSQKGHIANVQCAEVWCPMTAEFYQEYLYIMNPTKFQACQFLIQY VQAHCK LGLTAT LVREDDK IVDLNFLIGPK LY EANWME LQNNGYI AKVQCAEVWCPMSPEFYREY LYTMNPNKFRACQFLIKF

FK VHAK LGLTAT LVREDDRIGDLGYLIGPK LY EANWMD LAKNGHI ATVQCAEVWCPMTPEFYREY LHAMNPNKI QACQFLINY VKAHCK LGLTAT LVREDDLI QD LQWLIGPK LY EANWME LQDRGY LAKA LC SEVWC PMTA SYYR EY LWVCNPNKLRVCEFLIRW VKAHCK LGLTAT LVREDDLI QDLQWLIGPK LY EANWME LQDRGY LAKALCSEVWCPMTA SYYR EY LWVCNPNKLRVCEFLI HW VQAHCK LGLTAT LVREDDK IVDLNFLIGPK LY EANWME LQNNGYI AKVQCA EVWC PMSPEFYR EY LYTMNPNKFRACQFLIRF I AAHAK LGLTAT LVR EDDK I DDLNFLIGPK LY EANWMD LAQKGHI ANVQCAEVWC PMT SEFYQEY LYIMNPTKFQACQFLI HY TKAHCK LGLTAT LLREDEKI QDLNFLIGPK LY EANWLD LQKAGFLANVSCSEVWCPMTAEFYKEY LYTMNPNKFRACEYLIRF VQSHCK LGLTAT LLREDDK I ADLNFLIGPKLY EANWLELQKKGYIARVQCAEVWCPMSPEFYREYLYVMNPSKFRSCQFLIKY VQSHCK LG LTAT LLREDDK I ADLNFLIGPK LY EANWLELQKKGYI ARVQCA EVWC PMSPEFYREYLYVMNPSKFR SCQFLI KY V SHHCKLGLTAT LVR EDDK I EDLNFLIGPKLY EADWQDLSAKGHI ARVSCI EVWCGMTGDFYREYLSIMNPTKFQVCEYLINK I KAQCK LGLTATLIREDDRIRDLEFMIGPMLYEASWQELAKQGYIANAKCFEVICPMTKTYYSAY LAQLNPNKIDACKYLLEQ VQAHCKLELTAT LVREDDK I VDLNFLIGPK LY EANWME LQNSGYI AKVQCAEVWCPMSPEFYR EY LYTMNPNKFRACQFLIKF VQAHCK LGLTAT LVREDDK IVDLNFLIGPK LY EANWME LQNNGYI AKVQCAEVWCPMSPEFYREY LYTMNPNKFRACQFLIKF I AAHAK LGLTAT LVR EDDK I SDLNFLIGPK LY EANWMELSQKGHI ANVQCAEVWCPMTAEFYQEY LYIMNPTKFQACQFLI QY VDAKGVIGLTATYVREDHK I LDLFHLVGPK LYDI SMET LA SQGY LAKVHCVEVRTPMTKEFGLEY LAAANPNKMMCVRELVRQ VQAHCKLGLTAT LVR EDDK IVDLNFLIGPK LY EANWME LQNNGYI AKVQCA EVWC PMSPEFYREY LYTMNPNKFRACQFLIKF LK SH SK LG LTAT LLR EDDK I SD LNFLIGPK LY EANWME LS LGGHI ARVQCA EVWC PMPTEFYR EY LY I MNPMK FQACQYLI NY SAAPCR LGLTATY ER EDG LHT ELNR LVGGKVY EKKVSELA-GGHLAPYTIKRFAVTLTEKEQREY LAFNSNSKI EKLREI LEQ VQAHCKLGLTAT LVREDDK IVDLNFLIGPK LY EANWME LQNNGYI AKVQCA EVWC PMSPEFYREY LYTMNPNKFRACQFLIKF VQAHCK LGLTAT LVR EDDK I VDLNFLIGPK LY EANWME LQNNGYI AKVQCA EVWC PMSPEFYREYLYTMNPNKFRACQFLIKF I K SHSKLGLTAT LLREDDK I SHLNFLIGPK LY EANWMELSEKGHI AKVQCA EVWC PMPT EFYDEY LYAMNPRK FQACQYLINY TK SHCK LGLTAT LVREDERITDLNFLIGPK LY EANWLDLVKGGFIANVQCAEVWCPMTKEFFAEY LYAMNPNKFRACEFLIRF TKAHCK LG LTAT LVR EDDKVDH LNFLIGPK LY EANWLD LQRDGHI ANVQCV EVWC PMTAEFFRKY LYCMNPNKFMACQFLMQF I LAHCNLR LLATAFGHHDPV LDFLFTHLQSI FEWAWWLT PNNGYI AKVQCVEVWCPMSPEFYREY LYTMNPNKFRACQFLIKF I KCAIKIGLTAT LLREDQK LDNLYFMIGPKLYEENLIDLMTQGFLAKPHII EI QCDMPPI FLQEY LHTGNPGKYKALQFLIKN I AAHAK LGLTAT LVR EDDK I HD LNFLIGPK LY EANWMD LAQKGHI ANVQCA EVWC PMT SEFYQEY LYI MNPTKFQACQFLI HY VK SHCK LGLTAT LVREDLLIRDLHWI I GPK LY EANWV ELQNKGFLAKA LCKEI WCSMPC SFYKYY LYTCNPRKLMMCEYLIKY VK SHCKLGLTAT LVREDLLIRDLQWI IGPKLY EANWVELQNKGFLAKALCKEI WCSMPSSFYKYYLYTCNPRKLMMCEYLIKY VK SHCK LG LTAT LVREDLLIRDLQWI I GPK LY EANWV ELQNKGF LAKALCKEI WCSMPSSFYKYY LYTCNPRKLMMCEYLIKY TK SHCK LGLTAT LVREDERITDLNFLIGPK LY EANWLDLVKGGFIANVQCAEVWCPMTKEFFAEY LYVMNPNKFRACEFLIRF VQAHCK LGLTAT LVREDDK IVDLNFLIGPK LY EANWME LQNNGYI AKVQCAEVWCPMSPEFYREYLYTMNPNKFRACQFLIKF I AAHAK LGLTAT LVREDDKIGDLNFLIGPK LY EANWMELSQKGHIANVQCAEVWCPMTAEFYQEYLYIMNPTKFQACQFLI QY I AAHTK LGLTAT LVREDDK I DD LNFLIGPKMY EANWMD LAQKGHI AKVQCAEVWCAMTTEFYNEYLYIMNPKKFQACQFLIDY VQAHCK LGLTAT LVR EDDKIADLNFLIGPK LY EANWME LQNKGFIARVQCAEVWC PMAPEFFREY LYVMNPNKFRACQFLVRF

YK FHFK LGLTAT PYREDEKI INLFYMI GPK LY E ENWYD LV SQGFLAKPYCVEIRCEMSQLWMSEYIHTSNPRKFKTLEYLIKV I R SHCK LGLTAT LVREDDLIRDLQWLIGPKLY EANWLELQQKGY LAKVICKEI WCPMTAPFYREY LWSCNPVKLITCEYLLRF I R SHCK LGLTAT LVREDDLIRDLQWLIGPK LY EANWLELQEKGY LAKVICKEI WC PMTAPFYREY LWSCNPVKLITCEYLLKF AKAHTRLGLTATLIREDDKI SDLRYLVGPK LY EANWLELSEQGYLARVKCFEVTVPMTASFYKYYLCSSNPNKIRTVAGIIKF VKYKCVIGLSAT LLREDDKIGDLRHLVGPK LY EANWLDLTRAGFLARVECAEI QCPLPKAFLTEYVVCLNPYKLWCTQALLEF VKYKCVVGLSAT LLREDDKIGDLRHLVGPK LY EANWLELTRAGFLARVECAEVQCPLPLPFFREYVVCFNPYKLWCTQALLEF I KAHSKLGLTAT LVREDEKI DELNFLVGPK LY EANWMD LAAKGHI ATVQCA EVWC PMT PEFYR EY LYCMNPNKFQACQFLI DY I AAHAK LG LTAT LVREDDK I DD LNFLIGPK LY EANWMD LAQKGHI ANVQCAEVWCPMTSEFYQEY LYI MNPSKFQAAQFLINY

Anopheles_gambiae/1-3220
Arabidopsis_thaliana/1-3286
Ashbya_gossypii/1-3281
A fumigatus/1-3241
Aspergillus_niger/1-3300 Bombyx_mori/1-2389
Bos_taurus/1-3273
C_briggsae/1-3231
Caenorhabditis_elegans/1-3289
Candida_albicans/1-3266
Candida_glabrata/1-3282
Canis_familiaris/1-3286
iona_intestinalis/1-2712
Cryptococcus_neoformans/1-3301 Cryptosporidium_hominis/1-3020 C_parvum/1-3281
Danio_rerio/1-3286
Debaryomyces hansenii/1-3166
Dictyostelium_discoideum /1-3275 D_melanogaster/1-3287
Drosophila_pseudoobscura/1-3286 Encephalitozoon_cuniculi/1-3199 Entamoeba histolytica/1-3271 Gallus_gallus/1-3286
Homo_sapiens/1-3286
Kluyveromyces_lactis/1-3281 Leishmania_major/1-3286 Macaca_mulatta/1-3182 Magnaporthe_grisea/1-3286 Methanosarcina_acetivorans/1-3043 Monodelphis_domestica/1-3196
Mus musculus/1-3286
Neurospora_crassa/1-3264
Oryza_sativa/1-3218
Ostreococcus_lucimarinus/1-3287
Pan_troglodyte/1-3286
Paramecium_tetraurelia/1-3282
Pichia_stipitis/1-3290
_falciparum/1-3285
P_knowlesi/1-3284
Plasmodium_yoelii/1-3285
Populus_trichocarpa/1-3286 Rattus_norvegicus/1-3286
Saccharomyces_cerevisiae/1-3284
Schizosaccharomyces_pombe/1-3284
Strongylocentrotus purpuratus/1-3252
Takifugu_rubripes/1-2930
Tetrahymena_thermophila/1-3270
T_annulata/1-3248
Theileria parva/1-3273
Trichomonas vaginalis/1-3198
T_brucei/1-3264
rypanosoma_cruzi/1-3266
Ustilago maydis/1-3291
Yarrowia_lipolytica/1-3278

HE-KRG--DKTIVFSDNVFALKHY--AIKMNKPYI YGPT SQNERIQ'L LQNFKFNPKVNTI FVSKVADTSFDLPEANVLIQISS HEQQRG--DKI IVFADNLFALTEY--AMK LRKPMIYGATSHIERTKI LEAFKTSKTVNTVFLSKVGDNSIDIPEANVIIQISS HE-KRG--DKI IVFSDNVYALQEY--ALKLGKPFIYGSTPQQERMNI LQNFQYNDQI STIFLSKVGDTSIDLPEATCLIQISS HE-KRG--DKVIVFSDNVYALERY--ALK LNKAYI YGGTPQNERMRI LENFQHNEQVNTIFLSKIGDTSLDLPEATCLIQISS HE-KRG--DKVIVFSDNVYALQRY--ALKLNKAYI YGGT PQNERMRILENFQHNEQVNTIFLSKIGDTSLDLPEATCLIQISS


HE-ERG--DKI LV FCDRPMIIDYY--GNILKYPVIYGDVSQDERKKIFNLFKVSNQINTIFLSRVGDTAIDLPQANVGIQIGM HE-SRG--DKVIVFSDNLFALLHA--AKLLNRPFIYGKVSSAERIVILNKFKNETTFNTIFLSKVGDNALDI PCANVVIQISF HE-SRG--DKVIVFSDNLFALLHA--AKLLNRPFIYGKVSSAERIIILNKFKNETTFNTIFLSKVGDNALDIPCANVVIQISF HE-RRG--DKVLVFCDIIHILIHL--AGLLHCPEIHGETPENVRSSIFHEFKNGSKVNTLILSSVGDKAIDLPSASVVVQVCS HR-NRSPPDKVIIFCDQIDGIQYY--AQHLHVPFMDGKTSDMERENLLOYFQHSDNINAII LSRVGDVALDIPCASVVIQISC HR -NR SPPDKVII FCDDLEGVQYY--ARHLNVPFMDGKTTEVERENLLQYFQHSNDINAII LSRVGDVALDI PCASVII QVSG HE-NRG--DKIIVFSDNVYALVAY--AHK LKKPFIHGGTAHLERMRILQNFQHNPLVNTIFLSKVGDTSIDLPEATCLIQISS HE-KRG--DKIIVFSDNVHALKAY--ALK LGKFFI FGGTPQQERMKI LKNFQYNDQVNTIFLSKVGDTSIDLPEATCLIQISS

Anopheles_gambiae/1-3220
Arabidopsis_thaliana/1-3286
Ashbya_gossypii/1-3281
A fumigatus/1-3241
Aspergillus_niger/1-3300
Bombyx_mori/1-2389
Bos_taurus/1-3273
C_briggsae/1-3231
Caenorhabditis_elegans/1-3289
Candida_albicans/1-3266
Candida_glabrata/1-3282
Canis_familiaris/1-3286
Ciona intestinalis/1-2712
Cryptococcus_neoformans/1-3301 Cryptosporidium_hominis/1-3020 C_parvum/1-3281
Danio rerio/1-3286
Debaryomyces hansenii/1-3166
Dictyostelium_discoideum /1-3275 D_melanogaster/1-3287
Drosophila_pseudoobscura/1-3286
Encephalitozoon cuniculi/1-3199
Entamoeba_histolytica/1-3271
Gallus_gallus/1-3286
Homo_sapiens/1-3286
Kluyveromyces_lactis/1-3281
eishmania_major/1-3286
Macaca mulatta/1-3182
Magnaporthe_grisea/1-3286
Methanosarcina_acetivorans/1-3043
Monodelphis_domestica/1-3196
Mus musculus/1-3286
Neurospora_crassa/1-3264
Oryza_sativa/1-3218
Ostreococcus_lucimarinus/1-3287
Pan troglodyte/1-3286
Paramecium tetraurelia/1-3282
Pichia_stipitis/1-3290
_falciparum/1-3285
P_knowlesi/1-3284
Plasmodium_yoelii/1-3285
Populus_trichocarpa/1-3286
Rattus_norvegicus/1-3286
Saccharomyces_cerevisiae/1-3284
schizosaccharomyces_pombe/1-3284
Strongylocentrotus_purpuratus/1-3252
Takifugu_rubripes/1-2930
Tetrahymena_thermophila/1-3270
T_annulata/1-3248
Theileria parva/1-3273
Trichomonas_vaginalis/1-3198
T_brucei/1-3264
rypanosoma_cruzi/1-3266
Ustilago maydis/1-3291
Yarrowia_lipolytica/1-3278

HGG SRRQEAQR LGRVLRAKAFFY SLV SQDTEMAY STKR QR FLVDQ HGG SRRQEAQR LGRI LRAKAFFY SLV SQDTEMGY SRKRQRFLVNQ HGG SRRQEAQR LGRI LRAKAFFY SLV SQDT EMGY SRKRQRFLVNQ HYGSRQEAQR LGRI LRAKAFFY SLV SKDTEMYY STKRQAFLVDQ HYGSRRQEAQRLGRI LRAKAFFY SLVSKDTEMYY STKRQAFLVDQ HGG SRRQEAQR LGRVLRAKAFFY SLV SQDT EMAY STKR QR FLVDQ HGG SRRQEAQR LGRI LRAKAFFY SLV SQDTEVAY STKRQRFLVDQ HFGSRRQEAQRLGRILRAKAFFYSLV SKDT EMFY SSKRQGFLIDQ NFASRRQEAQRLGRI LRPKAFFY SLLSKDTEMEYADKRQQFIIDQ NFA SRRQEAQRLGRI LR PKAFFY SLLSKDT EMEYADKRQQFIIDQ HGG SRRQEAQR LGRVLRAKAY FY SLV SQDT EMAY STKRQRFLVDQ HYGSRQEAQRLGRILRAKAFFY SLV SKDT EMYY STKRQAFLVDQ HYGSRRQEAQR LGRILRPKAFFY SLV SKDTEMYY STKRQQFLIDQ HGG SRRQEAQR LGRI LRAKAFFYT LV SQDT EMSY SRKRQRFLVNQ HGG SRRQEAQR LGRILRAKAFFYTLV SQDTEMSY SRKRQRFLVNQ HFG SRRQEAQRLGRILRAKVYFY SLV SKDT EMFY SSKRQQFLIDQ SSG SRRQEAQR LGRI LRAKAYFYT LT SKDT EMY FSQRRQRVMRQN HGG SRRQEAQR LGRVLRAKAFFY SLVSQDTEMAY STKRQRFLVDQ hGG SRRQEAQR LGRVLRAKAFFY SLV SQDT EMAY STKR QRFLVDQ HYGSRRQEAQRLGRI LRAKAFFY SLVSKDTEMYY STKRQAFLVDQ HGGSRRQEAQR LGRI LR PKAWFYSII STDTEINYAAHRTAFLVDO hgG SRRQEAQR LGRVLRAKAFFY SLV SQDT EMAY STKR QR FLVDQ HYGSRQEAQRLGRI LRAKAFFY SLVSKDT EMYY SAKRQAFLVDQ GTGSKRAYVQR LGRI LRKKAVLYEIIAGETETGTARRRKEALSSG HGG SRRQEAQR LGRVLRAKAFFY SLV SQDT EMAY STKRQRFLVDQ HGG SRRQEAQR LGRVLRAKAFFY SLV SQDTEMAY STKRQRFLVDQ
HFGSRRQEAQRLGRI LRAKAFFY SLVSKDTEMYY SSKRQAFLVDQ
HAG SRRQEAQRLGRI LR PKAFFYSLVSTDTEMYY STKRQQFLIQQ HGG SRROEAOR LGRVLRAKAFFY SLV SODT EMAY STKROR FLVDO LGG SRRQKV QR LGRVMR PKAFFY SLA SKDT ESEY SYKRQKYITEQ HYGSRRQEAQRLGRI LRAKAFFY SLV SKDT EMYY STKRQAFLVDQ NFASRRQEAQRLGRIIRPKSFFY SLV SKDT EMCY SDKRQRFLINQ NFASRRQEAQRLGRIIRPK SFFY SLV SKDT EMCY SDKRQRFLINQ NFA SRRQEAQRLGRIIRPK SFFY SLV SKDTEMCY SDKRQRFLINQ HAG SRRQEAQR LGRILRAKAFFY SLV STDTEMYY STKRQQFLIDQ HGG SRRQEAQR LGRVLRAKAFFY SLVSQDT EMAY STKRQRFLVDQ HYGSRRQEAQR LGRI LRAKAFFY SLVSKDT EMYY YTKRQAFLVDQ HYGSRRQEAQRLGRILRAKAFFY SLV SKDT EMYY SSKRQAFLIDQ HGG SRRQEAQR LGRI LRAKAFFYTLV SQDT EMFY SLKRQRFLVNQ
HFK SRRQEVQRLGRIMRAKAFWYTLVSKGTETSYCLAROKCLINQ-GFKYEYYKIR SQAIEELKGA--GE-----DPYPHKF NFA SRRQEAQR LGRI LRPKAFFY SLVSKDT EMV FADKRQQFIIDO NFASRRQEAQR LGRILRPKAFFY SLV SKDT EMV FADKRQQFIIDO NYGARMQESQR LGRVLR PKAFFY SCI SDMTDLKY SARRQQFLVDQ LGA SRRQEAQR LGRI LRPK SY FYT LV SQDTEI SQSY ERQSWLRDQ LGA SRRQEAQR LGRILRPKSYFYTLV SQDTEVQQSYGRQSWLRDQ HFGSRRQEAQRLGRI LRAKAFFY SLVSKDTEMFY STKRQQFLIDQ HYGSRRQEAQR LGRI LRAKAFFY SLVSKDTEMYY STKRQAFLVDQ


Anopheles_gambiae/1-3220
Arabidopsis_thaliana/1-3286
Ashbya_gossypii/1-3281
A fumigatus/1-3241
Aspergillus_niger/1-3300 Bombyx_mori/1-2389
Bos_taurus/1-3273
C_briggsae/1-3231
Caenorhabditis_elegans/1-3289
Candida_albicans/1-3266 Candida_glabrata/1-3282
Canis_familiaris/1-3286
Ciona intestinalis/1-2712
Cryptococcus_neoformans/1-3301 Cryptosporidium_hominis/1-3020 C_parvum/1-3281
Danio_rerio/1-3286
Debaryomyces_hansenii/1-3166
Dictyostelium_discoideum /1-3275 D_melanogaster/1-3287 Drosophila_pseudoobscura/1-3286 Encephalitozoon_cuniculi/1-3199 Entamoeba histolytica/1-3271 Gallus_gallus/1-3286
Homo_sapiens/1-3286
Kluyveromyces_lactis/1-3281
Leishmania_major/1-3286
Macaca_mulatta/1-3182
Magnaporthe_grisea/1-3286
Methanosarcina_acetivorans/1-3043
Monodelphis_domestica/1-3196
Mus musculus/1-3286
Neurospora_crassa/1-3264
Oryza_sativa/1-3218
Ostreococcus_lucimarinus/1-3287
Pan troglodyte/1-3286
Paramecium tetraurelia/1-3282
Pichia_stipitis/1-3290
_falciparum/1-3285
P_knowlesi/1-3284
Plasmodium_yoelii/1-3285
Populus_trichocarpa/1-3286
Rattus_norvegicus/1-3286
Saccharomyces_cerevisiae/1-3284
schizosaccharomyces_pombe/1-3284
Strongylocentrotus purpuratus/1-3252
Takifugu_rubripes/1-2930
Tetrahymena_thermophila/1-3270
T_annulata/1-3248
Theileria parva/1-3273
Trichomonas_vaginalis/1-3198
T_brucei/1-3264
Trypanosoma_cruzi/1-3266
stilago_maydis/1-3291
Yarrowia_lipolytica/1-3278
 AVSMSI PKYI ETY---GS-LNNGDVVE-NAEESLAGRIMSKRSSSSKLFFYDLH------GDDFKVQVFLKLHSNAKRGDIVG NVTIGLPAFLNKY---AH-LQRGETLP-EERVSIAGRIHAKRESGSKLRFYVLH------ADGVEVQVYEQDHGLLKRGDIVG QVTDDLREY LKTY---DG-LAKGEQKP-DVTVRIAGRIYTKRSSGSKLFFYDIR------AEGVKVQVFEAQHEHLRRGDIVG QVIDDLRKYLTDY---EG-LAKGEQKP-EVAVRIAGRIYTKRASGAKLI FYDIR------AEGVKVQVFEAQHEHLRRGDIVG HVDI SLTHFIQEY---SH-LQPGDHLT-DITLKVAGRIHAKRASGGKLIFYDLR-----GEGVKLQVFIRINNKLRRGDIIG NVTISLTDFIAKY---SP-LQNEQ-VA-DEIVSVAGRIHSKRESGSKLVFYDIH-----GEGTHIQIFVTLHDRIKRGDIVG NVTISLTDFITKY---TP-LEKEQ-VV-EEIVSVAGRIHSKRESGSKLVFYDIH-----GEGTHIQIFVTLHDRIKRGDIVG NVTTKIPEFVEKY---AH-LQRGET LK-DVTVSVSGRIMTKRESGSKLKFYVLK-----GDGVEVQIFESMHEI LRRGDIIG QVSI SNPEFLAKY---AH-LKRGETLP-NEIVSIAGRIHAKRESGSKLKFYVLH------GDGVEVQVYENDHDLIKRGDIVG hVDI SLTHFIEEY---GH-LQPGDHLT-DITLKVAGRIHAKRASGGKLI FYDLR------GEGVKLQVFIHINNKLRRGDIIG HVSI SLSDYVEKY---NN-I EVGSHLN-DQQVSIAGRIHAKREAGPKLI FYDVR------GDGVKLQVYQEINERTRRGDIIG NVTHAVPK FVEEWGK EGK -LEKGETAQLNEPI SLAGRVYTIRESSSKLRFYDLK------ADGVKVQIYLDTHDRIRRGDIIG KI SMSLPAYALKY---GN-VENGYIDK-DTTLSLSGRVTIRSSSSKLIFYDIF-----CEEQKVQIFSVSHSEIRRGDVVG KI SMSLPAYALKY---GN-VENGYIDK -DTTLSLSGRVTIRSSSSKLIFYDIF-----CEEQKVQIFSVSHSEIRRGDVVG HVDLSLTEFIERY---NH-LQPGDHLT -DVV LNLSGRVHAKRASGAKLLFYDLR-----GEGVKLQVFVHINNKLRRGDIIG hV Si QLPAFAEKY---KD-LKKGESLK-DVEVKVSGRIMGKRESGSKLKFYVLK------GDGVQI QIYEKMHEYLRRGDIIG EV SHQLPKFVEEF---SV-LEKDGEPS-TQVVSIAGRVLSKRAAGSGLVFYDIT-----GEFNKVQVYVKINGLLRRGDIIG hVSSSLEDFIAKY--ENS-LKEGETLE-NVKLSVAGRVHAIRESGAKLIFYDLR-----GEGVKVQVFEIDTSKLRRGDIIG NVSISLENFIEQY---SG-LTDGETLE-KVSLSVAGRVHAIRESGAKLI FYDLR------GEGVKLQVFETDTAKLRRGDIIG MSMR LH-
 NVSITVPEFIAKY---SG-LEK SQ-VS-DDIVSVAGRVLSKRSS-SALMFIDLH--.--SNGESLQREQMAKFLRRGVVG HVDLSLSDFIERY---SH-LQPGDHLT-DITVSVAGRIHAKRASGGKLIFYDLR-----GEGVKLQVYFRINNKLRRGDIIG HVDI SLTDFIQKY ---SH-LQPGDHLT-DITLKVAGRIHAKRASGGKLIFYDLR-----GEGVKLQVFIHINNKLRRGDIIG QVSVTLPEFLSKY---AN-LKRGETLP-EEKVSIAGRIHAKRESGSKLKFYVLH------GDGVEVQVYEDDHSLIKRGDIVG HRQYTI PQYRRKY---APLLTEPDTSL-DETVTIAGRIINKRSSGSKLHFITIQ------GDMEIVQVFAEIHSKLKRGDIIG HVDISLTDFIQKY---SH-LQPGDHLT-DITLKVAGRIHAKRASGGKLIFYDLR------GEGVKLQVFIHINNKLRRGDIIG LVDYDPSQFDKDF---KH-LK SGDVDK-TREIRIAGRIFTKRSSGNKLIFYDIKTGSDTTTTGSKMQIFEQQHEHLGRGDVIG EKNGDICEI LVKF---ED-FEKNEGLS----VRTAGRLYNIRKHG-KMI FADLG------DQTGRI QVFATFKNLMDSGDIIG HVDTSLTHFIEQY---NN-LQPGDHLT-DITVRVAGRIHAKRASGGKLI FYDLR------GEGVKLQVFFPINNKLPRGDIFG HVDISLTQFIQEY---SH-LQPGDHLT-DVTLKVAGRIHAKRASGGKLI FYDLR------GEGVKLQVFVHINNKLRRGDIIG QVNYDDSNFVEEF---GS-LKTGETLP-EKELRIAGRIYNIRTAGSKLIFYDIRTSADTKSIGTRMQVFEKQHAHLRRGDIIG LANITVADYI EKY---KS-MNVGDKLV-DVTECLAGRIMTKRAQSSKLLFYDLY-----GGGEKVQVFIKFHSTLKRGDIVG hVDTRVGEFIEKY---SG-LADGTTAE-GESASVAGRIMSKRASGKKLYFYDLI------ADGKKI QVFQKIHSATRRGDIVG HVDI SLTDFIQKY---SH-LQPGDHLT-DITLKVAGRIHAKRASGGKLI FYDLR------GEGVKLQVFIHINNKLRRGDIIG QVDLTIAQFRDKY---GPLCTEKGKIH-EDFVSVAGRVVTIRSMGAKLMFYDLQ------GEGTKI QVFEKVHTLIKRGDIIG HRNITLPEFAEKY---SS-LTRGETLQ-DVEVKVTGRIMTKRESGAKLRFYVLK-----GDGVEVQIYEKMHEY LRRGDVIG ERTISIPEFIEKY---KD-LGNGEHLE-DTILNITGRIMRVSASGQKLRFFDLV-----GDGEKIQVFAECYDKIRRGDIVG ERTITVPEFVEKY---QN-LASGEHLE-NTVLNVTGRIMRVSASGQKLRFFDLV------GDGAKI QVFAEAYDKIRRGDIVG ERTITIPDFIEKY---KD-LQNGEHLE-ETILNMTGRIMRVSSSGQKLRFFDLV------GDGKRIQVFVECYDKIKRGDIVG FVILSIPEYIDKY---GG-LSNGEHLE-DVSVSLAGRIMSKRSSSSKLFFYDLH------GLGAKVQVFSKLHSSVKRGDIVG HVDI SLTQFIQEY---SH-LQPGDHLT-DITLKVAGRIHAKRASGGKLIFYDLR-----GEGVKLQVFVHINNKLRRGDIIG HV SI SNPEFLAKY---AH-LKKGETLP-EEKVSIAGRIHAKRESGSKLKFYVLH------GDGVEVQLYEKDHDLLKRGDIVG QVTITLPEFIAKY---EG-LARGETKP-EVEVAVAGRVLGLRTAGNKLRFYEIH------ADGKKLQVFAAQHEHLRRGDIIG HVTI SLTDFLEKY ---DY -LKAED-IA-DEVLSLSGRVHAKRASGAKLIFYDLR------GEGVKLQVFTRLNEKIRRGDIIG HVDVSLTEFIEKY---KN-LQPGDQLT-D-AVKVAGRVHAKRVSGAKLLFYDLR-----GEGVKLQVFVAINNKLRRGDIIG DVSHSISQFIEEF---DPKLTENGQTI-DTIVTIGARITSFRASGKALIFYQVQ------QEGKKLQVFEEINSLFKRGDIIG ---MSLKEYVDKY---EH-LEAGEHLE-NELVSIAGRVSRIASSSSKLRFLDIK------SEGTKLQVFNDTYNNIKRGDIIG HVNMSLKEFVGKY---DH-LEAGAHLE-NELVSIAGRVSRIASSSSKLRFLDIK------SEGTKLQVFNDTYNNIKRGDIIG NVSHTFKQFYAQF---EH-LKAGEELP-DVKVSVASRIAQLRAHG-NLYFFEMY------ESTFKLQLFKEEVSSFHLGDIVG HRDYTLPAFRECF---KPMLQEKGQRL-DKVVTIAGRIVVKRSSSSKLHFLALQ------GDGEVLQVFADIHSKIKRGDIIG DRQYTI PAFKARF---APQLSEKGQRV-EEVVAIAGRIVNKRSSGSKLNFLT LQ------GDADTVQVFAAVHGRIRRGDIIG HVSI SLSEFISKY---EGKLEAGQHLD-QEEVSIAGRLHNMR SSGQKLRFYDLH------GEGVKVQVFFAIHELLRRGDVVG NVTTKVDEFVEKY---KG-LARGEIKK-DEEVSVAGRVHT LRAAGSKLRFYVLH------QEGKTVQIDWGIHDLIRRGDVIG

Anopheles_gambiae/1-3220
Arabidopsis_thaliana/1-3286
Ashbya_gossypii/1-3281
A_fumigatus/1-3241
Aspergillus_niger/1-3300
Bombyx_mori/1-2389
Bos_taurus/1-3273
C_briggsae/1-3231
Caenorhabditis_elegans/1-3289
Candida_albicans/1-3266
Candida_glabrata/1-3282
Canis_familiaris/1-3286
Ciona_intestinalis/1-2712
Cryptococcus_neoformans/1-3301 Cryptosporidium_hominis/1-3020 C_parvum/1-3281
Danio_rerio/1-3286
Debaryomyces_hansenii/1-3166
Dictyostelium_discoideum /1-3275 D_melanogaster/1-3287
Drosophila_pseudoobscura/1-3286 Encephalitozoon_cuniculi/1-3199 Entamoeba_histolytica/1-3271 Gallus_gallus/1-3286
Homo_sapiens/1-3286
Kluyveromyces_lactis/1-3281
Leishmania_major/1-3286
Macaca_mulatta/1-3182
Magnaporthe_grisea/1-3286
Methanosarcina_acetivorans/1-3043
Monodelphis_domestica/1-3196
Mus_musculus/1-3286
Neurospora_crassa/1-3264
Oryza_sativa/1-3218
Ostreococcus_lucimarinus/1-3287
Pan_troglodyte/1-3286
Paramecium_tetraurelia/1-3282
Pichia_stipitis/1-3290
P_falciparum /1-3285
P_knowlesi/1-3284
Plasmodium_yoelii/1-3285
Populus_trichocarpa/1-3286
Rattus_norvegicus/1-3286
Saccharomyces_cerevisiae/1-3284
Schizosaccharomyces_pombe/1-3284
Strongylocentrotus_purpuratus/1-3252
Takifugu_rubripes/1-2930
Tetrahymena_thermophila/1-3270
T_annulata/1-3248
Theileria_parva/1-3273
Trichomonas_vaginalis/1-3198
T_brucei/1-3264
Trypanosoma_cruzi/1-3266
Ustilago_maydis/1-3291
Yarrowia_lipolytica/1-3278

VTGVPGELSAMARRIKLLSPCLHMLPGLKDK ETRFRQRYLDLI LNNNVRNI FVTRALI I SYVRRFFDNLGFLEVETPMMNMIP VIGFPGELSIFPRSFI LLSHCLHMMPVLKDQESRYRQRHLDMI LNVEVRQIFRTRAKIISYVRRFLDNKNFLEVETPMMNMIA VEGYVGEISVFVSRIQLLTPCLHMLPGFKDQETRYRKRYLDLIMNKDARGRFITRSKIITYIRKFLDNRDFIEVETPMMNVIA VVGFPGELSIFATEVVLLAPCLHAI PGFQDKEQRFR QRY LDLIMNER SRNVFVTRSKIVRYVRNFFDSRDFI EVETPMMNAIA IVGFPGELSI FATEVVLLSPCLHAI PGLQDKEQRFR QRYLDLIMNDK SRNVFVTRSKIVRYIRNFFDNRDFVEVETPMMNAIA VQGNPGELSIIPYEITLLSPCLHMLPGLKDKETRYRQRYLDLILNDFVRQKFIIRSKIITYIRSFLDELGFLEVETPTPMMNIIP FTGQAGELSLI PKEVLQLTPCLHMLPGLKDKELRFRKRYLDLILNPRVKDNFVIRSKIITFLRRYLDNLGFLEVETPIMNQIA FTGRAGELSLIPNEI LQLTPCLHMLPGLKDKELRFRKRYLDLILNPRVKDNFVIRSKIITFLRRYLDNLGFLEVETPIMNQIA VTGYPGELSVFATKVQLLTPCLHMLPGFKDQEARYRKRYLDLIMNDSSRERFRVRSKII OYIRKFLDNRDFVEVETPI LNVIA VEGYVGEI SVFVKRIELLTPCLHMLPGFKDQETRYRKRYLDLIMNKDSRKRFITRSKIIKYIRKFLDNRDFIEVETPMMNVIA VKGNPGELSIIPYEITLLSPCLHMLPGLKDKETRYRQRYLDLI LNDFVRQKFIIRSKIITYIRSFLDELGFLEIETPMMNIIP VIGHPGELSIVPNTIEILSPCLHMLPGLKDK ETRYRQRYLDLIMNDQTRQKFITRAKIISYIRSFFDQMGFLEVETPMMNMVA VTGIPGELSLSISSIQLLSPCLHLLPGVVDLETRYRKRYLDLIMNPSTRDIFVTRSKVINYIRKYLDAQGFLEVETPMMSMIA FTGFPGELSLFSK SVVLLSPCYHMLPGLKDQEVRYRQRYLDLMLNEESRKVFKLRSRAIKYIRNYFDRLGFLEVETPMLNMIY FTGFPGELSLFSK SVVLLSPCYHMLPGLKDQEVRYR QRY LDLMLNEESRKVFKLRSRAIKYIRNYFDRLGFLEVETPMLNMIY VRGNPGELSIIPVEMTLLSPCLHMLPGLKDK ETRFRQRYLDLI LNDFVRQKFVTRSKIITYLRSFLDQLGFLEI ETPMMNLIP VTGYPGEVSVFAT SVQLLTPCLHMLPGFKDQEARYRKRYLDLIMNESTRDRFKVRSQIISFIRKFLDTRDFTEVETPMMNVIA AKGTPGELSLFATEVILLSPCLHMLPGLTDPETRFRQRYLDMICNESVKKNFIIRSKVIQGVRRYLDNLGFIEVETPMMNMIA VVGHPGELSVMPSEIKLLSPCLHMLPGLKDKETRYRQRYLDLI LNNNVREKFQIRAKIISYVRQFLDRLGFLEI ETPMMNMIA VKGHPGELSIMPTEIKLLSPCLHMLPGLKDKETRYRQRYLDLI LNNKVRENFQIRAKIISYVRQFLDRLGFLEI ETPMMNMVA FTGNPLEASVFATDIIVLTPCLRTI PGLKDPETIYRKRYMDLLINRESRNRFQKRAQIIGYIRSFLDSRGFLEVETPMMNLIP FTGHPGELSLIPISGMILSPCLHMLPGLGDQETRFRKRYLDLIVNPESVKNFVLRTKVVKAVRKYLDDKGFLEVETPILNTIP VVGNPGELSIIPYEITLLSPCLHMLPGLKDKETRYRQRYLDLILNDYVRQKFITRAKIVTYIRSFLDELGFLEIETPMMNIIP VQGNPGELSIIPYEITLLSPCLHMLPGLKDKETRYRQRYLDLI LNDFVRQKFIIRSKIITYIRSFLDELGFLEIETPMMNIIP VEGYVGEI SVFVSRIQLLTPCLHMLPGFKDQELRYRKRYLDLIMNKDARNRFITRSKIISYVRKFLDTRNFIEVETPMMNVIA IAGKPNEFSLKATEITLLSTCYHMLPGLSSFEQRFRQRYLDFIVNRDNIKTFI QRANIIKYIRKFFDERDFVEVETPVLNQIA VQGNPGELSIIPYEITLLSPCLHMLPGLKDKETRYRORYLDLILNDFVRQKFIIRSKMVTYIRSFLDELGFLEIETPMMNIIP IVGFPGELSLFATEVVQLSPSLHLLPGFTDGEKRFRMRYLDFMFNDK SREVLWQRSRIVKYIRDFFHDRRFIEVETPMMTSIA I QGELGENSISVSEFSLLSKSLCALPGLKDVETRYRKRYLDLIVNAEKREIFVMRSKLISEIRRFLADREFLEFETPI LQTVY VPGNPGELSLI PHEITLLSPCLHMLPGLKDKETRYRQRYLDLI LNDFVRQKFITRSKIITYIRSFLDELGFLEI ETPMMNVIP VEGNPGELSIIPQEITLLSPCLHMLPGLKDK ETRYRQRYLDLILNDFVRQKFIVRSKIITYIRSFLDELGFLEIETPMMNIIP IVGFPGELSVFATEVQLLSPC LHMLPPFADA EQRARMRY LDMLWNDR SRET LWQR SRMVRYIRDFFHERRFI EVETPMMHAIA VCGYPGELSIFPKKIVVLSPCLHMMPVLRDQETRYRQRYLDLMVNHEVRHIFKTRSKVVSFIRKFLDGLDFLEVETPMMNMIA VKGTPGELSLFPSNFEILTPCLKMLPGLKDVETRFRMRFLDLMMNNEVRDTFYIRSNI IRYIRKYLDDRDFLEVETPMMNMIA VQGNPGELSIIPYEITLLSPCLHMLPGLKDKETRYRQRYLDLI LNDFVRQKFIIRSKIITYIRSFLDELGFLEIETPMMNIIP VKGNPGELSIAPGFIQLLSPTLHMLPGFKDHEQRYRMR LDLIMNKKVRDIFLTRSSVIKQLREYFDGKGFIEVETPSLNVIQ VTGYPGEVSVFAT SVQLLTPCLHMLPGFKDQEARYRKRYLDLIMNDATRDRFKVRSKIIGYIRKFLDNRDFVEVETPILNVIA IVGFPGELSIFPKETILLSACLHMLPGLKDTEIRYRQRYLDLLINESSRHTFVTRTKIINFLRNFLNERGFFEVETPMMNLIA IVGFPGELSIFPKETIILSPCLHMLPGLKDTEIRSRQRYLDLMINESTRSTFITRTKIINYLRNFLNDRGFIEVETPTMNLVA IIGFPGELSIFPKETIVLSPCLHMLPGLKDTEIRYRQRYLDLLINESTRNVFITRTKIINFLRNFLNNQGFIEVETPSMNLMA ITGFPGELSIFPTSFMVLSHC LHMMPI LKDQETRYRQRYLDLMLNSEVRQIFKTRSKIIKYIQNFLDDLDFLEVETPMMNMIP VEGNPGELSIVPREMTLLSPCLHMLPGLKDKETRYRQRYLDLILNDFVRQKFIIRSKIITYIRSFLDELGFLEIETPMMNIIP VEGYVGEVSVFVSRVQLLTPCLHMLPGFKDQETRYRKRYLDLIMNKDARNRFITRSEIIRYIRRFLDQRKFIEVETPMMNVIA IRGYPGELSIFAR QCVLLSPCLRMLPGLKDLEIRHR QRYLDLIMNRSTRDRFVMR SRIIQYIRHFFDSRDFMEVETPMMNMIA VKGRPGELSILPSEITLLSPCLHMLPGVTNK ETRFRQRYLDLIMNDYVRDKFITRSKIVSYLRRFFDELGFLEVETPMMNMIA VCGNPGELSIIPKEMILLSPCLHMLPGLKDKETRYRQRYLDLI LNDSVRQKFITRSKIITYLRSFLDQMGFLEIETPMMNIVP ITGKPGELSIAPTKLQLLSPCLHMLPGLKDMETRYRKRYLDLIMNNS SRNNFITRTKIISYIRRYLDDRNFLEVETPQMNMIP LTGFPGELSVFPKSVKILSPCLHMLPGLKDNDVRFR QRY LDLMMNDDSLKVMK LRSRIIDYLRKFLTSRGFFEVETPMLKTTS LTGFPGELSVFPK SVQI LSPCLHMLPGLKDNDVRFR QRYLDLMMNDDSLKVMK LRSRIIDYLRKFLTSRGFFEVETPMLKITS AEGFPGELSVVVTKLVLLAPCLFQMPKLEDLEVRYRQRFFDLIVNRENRQIFETRCKVVKMIRGFLDDLDFTEVETPIMWKTA VRGVPGEFSMSAYEITLLSTCFHMLPGLSSVEQRFRQRYLDLIVNRENAKTFILRSKIISYIRSFFDQKDFLEVETPMLNQIA VKGVAGEFSMNAFEITLLSTCYHMLPGLSSI EQRFRQRYLDFIVNRENIQTFVTRSKVIRYIRNFFEDLNFLEVETPVLNQIA VTGVPGELSIFPSSIKLLSPSLKMLPGFTDTEQRHRKRYLDLIMNNHVRDIFVKRAKIINYVRRFLDNLGFLEVETPMMNQIA IRGYPGELSVFCKELVLLTPSLHMLPGFKDVETRFRQRYLDLIMNDSTRERFIVRSKII QYIRKFLDNKDFIEVETPMMNIIA

Anopheles_gambiae/1-3220 Arabidopsis_thaliana/1-3286
Ashbya gossypii/1-3281
A fumigatus/1-3241
Aspergillus_niger/1-3300 Bombyx_mori/1-2389
Bos_taurus/1-3273
_briggsae/1-3231
Caenorhabditis elegans/1-3289
Candida_albicans/1-3266
Candida_glabrata/1-3282
Canis_familiaris/1-3286
iona intestinalis/1-2712
Cryptococcus_neoformans/1-3301 Cryptosporidium_hominis/1-3020 _parvum/1-3281
Danio rerio/1-3286
Debaryomyces hansenii/1-3166
Dictyostelium_discoideum /1-3275
D_melanogaster/1-3287
Drosophila_pseudoobscura/1-3286
ncephalitozoon cuniculi/1-3199
Entamoeba_histolytica/1-3271
Gallus_gallus/1-3286
Homo_sapiens/1-3286
Kluyveromyces_lactis/1-3281 eishmania_major/1-3286 Macaca_mulatta/1-3182 Magnaporthe_grisea/1-3286 Methanosarcina_acetivorans/1-3043 Monodelphis_domestica/1-3196
Mus musculus/1-3286
Neurospora_crassa/1-3264
Oryza_sativa/1-3218
Ostreococcus_lucimarinus/1-3287
Pan troglodyte/1-3286
Paramecium tetraurelia/1-3282
Pichia_stipitis/1-3290
_falciparum/1-3285
knowlesi/1-3284
Plasmodium_yoelii/1-3285
Populus_trichocarpa/1-3286
attus_norvegicus/1-3286
Saccharomyces_cerevisiae/1-3284
Schizosaccharomyces_pombe/1-3284
Strongylocentrotus_purpuratus/1-3252
Takifugu_rubripes/1-2930
Tetrahymena_thermophila/1-3270
_annulata/1-3248
Theileria parva/1-3273
Trichomonas vaginalis/1-3198
T_brucei/1-3264
Trypanosoma_cruzi/1-3266
stilago maydis/1-3291
Yarrowia_lipolytica/1-3278

GGATAKPFITHHNDLNMDLFLRIAPELYLKMLTVGGLDRVYEIGRQFRNEGIDLTHNPEFTTCEFYMAYADYNDIIDITQQLL GGAAARPFVTHHNDLDMR LYMRIAPELYLKQLIVGGLERVYEIGKQFRNEGIDLTHNPEFTTC EFYMAFADYNDLMEMT EVML GGATAKPFVTHHNDLDMQMYMRIAPELFLKQLVVGGMDRVYEIGRQFRNEGIDMTHNPEFTTCEFYQAYADVYDLMDMT ELLF GGATAKPFITHHNDLDMNLFMRVAPELYLKMLIVGGLERVYELGRQFRNEGIDLTHNPEFTTCEFYWAYADVYDVMNLTEELI GGATAKPFVTHHNDLDMNLFMRVAPELYLKMLIVGGLERVYELGRQFRNEGIDLTHNPEFTTCEFYWAYADVYDVMNLTEELV GGATAKPFITHHNELNMDLYMRIAPELYHKMLVVGGLDRVYEIGRQFRNEGIDLTHNPEFTTCEFYMAYADYNDLMTITESIL GGAVAKPFITYHNELDMNLYMRIAPELYHKMLVVGGIDRVYEIGRQFRNEGIDLTHNPEFTTCEFYMAYADYHDLMEITEKMI GGATAKPFITHHNDLDMNLFLRVAPELYHKMLVVGGIDRVY EVGRLFRNEGIDLTHNPEFTTCEFYMAYADY EDVI QLT EDLL GGATAKPFITHHNDLDMNLFLRVAPELYHKMLVVGGIDRVYEVGRLFRNEGIDLTHNPEFTTCEFYMAYADYEDVIQLTEDLL GGATAKPFTTHHNDLNMEMFMRIAPELFLKELVVGGMDRVYEIGRQFRNEGIDMTHNPEFTTCEFYQAYADVYDLMDMT ELMF GGATAKPFVTHHNDLDMDMFMRIAPELFLKELVVGGMDRVYEIGRQFRNEGIDMTHNPEFTTCEFYQAYADVYDLMDMTELLF GGAVAKPFITYHNELDMNLYMRIAPELYHKI LVVGGIDRVYEIGRQFRNEGIDLTHNPEFTTCEFYMAYADYHDLMEITEKMI GGATAK PFITHHNDLDMDLYMRVAPELY LKMLVVGGLDRVYEIGR LFRNEGIDMTHNPEFT SC EFYMAYADYEDLMKISETLI GGATAKPFVTHHNDLK LDLFMRIAPELYLKELVVGGLDRVFEIGRVFRNEQIDMTHNPEFSICEFYMAYADMYDIMDMTEELI GGAAARPFITYHNELETQLYMRIAPELYLKQLIVGGLDKVYEIGKNFRNEGIDLTHNPEFTAMEFYMAYADYYDLMDLTEELI GGAAARPFITYHNELETQLYMRIAPELYLKQLIVGGLDKVYEIGKNFRNEGIDLTHNPEFTAMEFYMAYADYYDLMDLTEELI GGAVAKPFITYHNDLNMNLYMRIAPELYHKMLVVGGIDRVYEIGRQFRNEGIDLTHNPEFTTCEFYMAYADYHDLMEITEKLL GGATAKPFVTHHND NMDMFM GGAAAKPFLTHHNALNMDLFMR GGATAKPFVTHHNDLKMDLFMR GGATAKPFVTHHNELKMDLFMR GGAAAKPFITHHNELKLDLYMR GGATARPFITHHNQLDI QMY MR GGAMAKPFITYHNELDMK LYMR GGAVAKPFITYHNELDMNLYMR GGATAK PFVTHHNDLDMDMYMR GGAAARPFVTHHNDLNQTMFLR GGAVAKPFITYHNELDMNLYMR GGATALPFVTHHNEYDLDMFMR GGANARPFKTFHNCLGQNLFLR GGAVAKPFITYHNELDMNLYMR GGAVAKPFITYHNELDMNLYMR GGATALP FVTHHNDLDMDMFMR GGAAARPFVTHHNELNMRLYMR GGATARPFITHHNDLNMT LYMR GGAVAKPFITYHNELDMNLYMR GGATAKPFKT FHNSLHRDLFMR GGATAKPFITHHNDLSMDMFMR GGANARPFITHHNDLDLDLYLR GGANAKPFITHHNDLDLDLYLR GGASARPFITHHNDLDLDLYLR GGAAAR PFKTHHNDLNMK LYMR GGAVAKPFITYHNELDMNLYMR GGATAKPFITHHNDLDMDMYMR GGATAK PFVTHHNDLDMDLYMR GGATAKPFITHHNDLNMDLFMR GGAVARPFVTYHNDLDMNLYMR GGAAARPFVTHHNDLNMDIFMR TGASAKPFITHHNELDLDLFMR TGASAKPFITHHNELDLDLFMR GGATAKPFITHHNALDIDLWLR GGAAARPFITHHNELNQTMYLR GGAAARPFITHHNELNQRMYLR GGATAK PFVTYHNDLKLDLEMR GATAKPFVTHHNDLNLDMY

## APELFLK

 APELY LKOLVVGGMDRVY EIGKDFRNEDIDHTHNPEFTTC EFYMAYADYNDLYTMT ELL APELYHKMLVVGGLDRVYEIGRQFRNEGIDLTHNPEFTTCEFYMAYADYADIMDITEQLV APELYHKMLVVGGLDRVYEIGRQFRNEGIDLTHNPEFTTCEFYMAYADYADVMDITEQLI SPELYLKKLVVGGLERVYEIGKQFRNEGIDLTHNPEFT SCEFYMAYADYNDLMEMTEELI APELY LKELVVGGINRVYEIGRLFRNEGIDQTHNPEFTTCEFYMAYADYNDIMKMTEELL APELYHKMLVVGGLDRVYEIGRQFRNEGIDLTHNPEFTTC EFYMAYADYRDLMEITEKLL APELYHKMLVVGGIDRVYEIGRQFRNEGIDLTHNPEFTTC EFYMAYADYHDLMEIT EKMV APELFLKELVVGGMDRVYEIGRQFRNEGIDMTHNPEFTTCEFY QAYADVYDLMDMTEIMI APELY LKELVVGGMDRVY EIGKQFRNEGIDLTHNPEFT SC EAY WAY MDYHDWMTAT EDLL APELYHKI LVVGGIDRVYEIGRQFRNEGIDLTHNPEFTTCEFYMAYADYHDVMEITEKMV APELYLKMLVVGGYNKVFEIGKNFRNEGCDLTHNPEFTTI EAYAAYYDMYDVMDYTEELV APELYLKRLVVGGY EKVFEI SKNFRNEDIDTTHNPEFTMI EVY EAYRDYNDMMDLTEALI APELYHKMLVVGGIDRVYEI GR QFRNEGIDLTHNPEFTTCEFYMAYADYHDLMEITEKMI APELYHKMLVVGGIDRVYEIGRQFRNEGIDLTHNPEFTTCEFYMAYADYHDLMEITEKML VAPELFLKKMIVGQFGKVFEMGKNFRNEGIDLTHNPEFTSIEFYWAYADVYDLMSITEELV APELYLKELVVGGLDRVYEIGKQFRNEGIDLTHNPEFTTCEFYMAYADYNDLI ELTETML APELYLKQLVVGGI ERVYEIGRQFRNEGIDMTHNPEFTTCEFYQAYADYDDLMQMTEEMI APELYHKMLVVGGIDRVYEIGRQFRNEGIDLTHNPEFTTC EFYMAYADYHDLMEITEKMV APELYLKMLIVGGLDRVYEIGKNFRNEGIDQTHNPEFTAMEFYWAYCDYNDLMTVTEEVL APELFLKELVVGGMDRVYEIGRQFRNEGIDMTHNPEFTTC EFY QAYADVYDLMDMT ELMF ATELPLKMLIVGGIDKVYEIGKVFRNEGIDNTHNPEFTSCEFYWAYADYNDLIKWSEDFF ATELPLKMLIVGGIDKVYEIGKVFRNEGIDNTHNPEFTSCEFYWAYADFYDLIKWSEDFF ATELPLKMLIVGGLDRVYEIGKVFRNEGIDNTHNPEFTSCEFYWAYADYYDLIKWSEEFF APELY LKELVVGGLDRVYEIGKQFRNEGIDLTHNPEFTTCEFYMAYADYNDLMELTEKML IAPELYHKMLVVGGIDRVY EI GRQFRNEGIDLTHNP EFTTCEFYMAYADYHDLMEITEKML APELFLKQLVVGGLDRVYEIGRQFRNEGIDMTHNPEFTTCEFYQAYADVYDLMDMTELMF APELY LKMLVVGGLDRVYEI GRQFRNEGADLTHNPEFTSI EFY QAYADYYDLMDTTEELL APELY LKMLVVGGLQRVYEIGRQFRNEGIDLTHNPEFTTLEFYMAYADYNDLMDIAERLL APELYHKMLVVGGIDRVYEIGRQFRNEGIDMTHNPEFTTCEFYMAYADYHDLMEITEKLL APELYLKNLVVGGFERVYEIGKQFRNEGIDRTHNPEFTSI ELYQAYADYEDMMK LTEDLL APELPLKLIIIGGFEKVFEIGKCFRNEGIDPTHNPEFT SCEFYWAYADYHDLMKLTEELL APELPLKLIIIGGFEKVFEIGKCFRNEGIDPTHNPEFTSCEFYWAYADYHDLMKLTEEFL VAPELFLKMLVVGGMNRVYELGKQFRNEGIDLTHNPEFT SCEFYMAYADYNDLMDLTEKLY APELYLKKLVVGG LDRVYEIGKORNEGIDLTHNPEFTS EEYY WAYADYNDWMETTEELL APELYLKELVVGGMDRVYELGKQFRNEGIDLTHNPEFTSVEAY WAYADYNDWMRTTEDLF APELFLKELVVGGLDRVYEIGRVFRNESIDQTHNPEFSICEFYMAYADMYDLMDITESMIQLVVGGMERVYEI GRQFRNEGI DQTHNPEFTTC EFY EAYADVYDLMETTELLF

Anopheles_gambiae/1-3220 Arabidopsis_thaliana/1-3286
Ashbya_gossypii/1-3281
A fumigatus/1-3241
Aspergillus_niger/1-3300
Bombyx_mori/1-2389
Bos_taurus/1-3273
C briggsae/1-3231
Caenorhabditis_elegans/1-3289
Candida_albicans/1-3266
Candida_glabrata/1-3282
Canis_familiaris/1-3286
Ciona_intestinalis/1-2712
Cryptococcus_neoformans/1-3301 Cryptosporidium_hominis/1-3020 C_parvum/1-3281
Danio_rerio/1-3286
Debaryomyces_hansenii/1-3166
Dictyostelium_discoideum /1-3275 D_melanogaster/1-3287
Drosophila_pseudoobscura/1-3286 Encephalitozoon_cuniculi/1-3199 Entamoeba_histolytica/1-3271 Gallus_gallus/1-3286
Homo_sapiens/1-3286
Kluyveromyces_lactis/1-3281 Leishmania_major/1-3286 Macaca_mulatta/1-3182 Magnaporthe_grisea/1-3286 Methanosarcina_acetivorans/1-3043 Monodelphis_domestica/1-3196
Mus musculus/1-3286
Neurospora_crassa/1-3264
Oryza_sativa/1-3218
Ostreococcus_lucimarinus/1-3287
Pan troglodyte/1-3286
Paramecium_tetraurelia/1-3282
Pichia_stipitis/1-3290
_falciparum/1-3285
P_knowlesi/1-3284
Plasmodium_yoelii/1-3285
Populus_trichocarpa/1-3286
Rattus_norvegicus/1-3286
Saccharomyces_cerevisiae/1-3284
Schizosaccharomyces_pombe/1-3284
Strongylocentrotus_purpuratus/1-3252
Takifugu_rubripes/1-2930
Tetrahymena_thermophila/1-3270
T_annulata/1-3248
Theileria parva/1-3273
Trichomonas_vaginalis/1-3198
T_brucei/1-3264
rypanosoma_cruzi/1-3266
Ustilago_maydis/1-3291
Yarrowia_lipolytica/1-3278


Anopheles_gambiae/1-3220 Arabidopsis_thaliana/1-3286
Ashbya_gossypii/1-3281
A fumigatus/1-3241
Aspergillus_niger/1-3300 Bombyx_mori/1-2389
Bos_taurus/1-3273
_briggsae/1-3231
Caenorhabditis_elegans/1-3289
Candida_albicans/1-3266
Candida_glabrata/1-3282
Canis_familiaris/1-3286
Ciona intestinalis/1-2712
Cryptococcus_neoformans/1-3301 Cryptosporidium_hominis/1-3020 C_parvum/1-3281
Danio rerio/1-3286
Debaryomyces hansenii/1-3166
Dictyostelium_discoideum /1-3275 D_melanogaster/1-3287
Drosophila_pseudoobscura/1-3286 Encephalitozoon_cuniculi/1-3199 Entamoeba_histolytica/1-3271 Gallus_gallus/1-3286
Homo_sapiens/1-3286
Kluyveromyces_lactis/1-3281 Leishmania_major/1-3286 Macaca_mulatta/1-3182 Magnaporthe_grisea/1-3286 Methanosarcina_acetivorans/1-3043 Monodelphis_domestica/1-3196
Mus musculus/1-3286
Neurospora_crassa/1-3264
Oryza_sativa/1-3218
Ostreococcus_lucimarinus/1-3287
Pan_troglodyte/1-3286
Paramecium tetraurelia/1-3282
Pichia_stipitis/1-3290
_falciparum/1-3285
P_knowlesi/1-3284
Plasmodium_yoelii/1-3285
Populus_trichocarpa/1-3286
Rattus_norvegicus/1-3286
Saccharomyces_cerevisiae/1-3284
Schizosaccharomyces_pombe/1-3284
Strongylocentrotus_purpuratus/1-3252
Takifugu_rubripes/1-2930
etrahymena_thermophila/1-3270
T_annulata/1-3248
Theileria parva/1-3273
Trichomonas vaginalis/1-3198
T_brucei/1-3264
Trypanosoma_cruzi/1-3266
stilago maydis/1-3291
Yarrowia_lipolytica/1-3278

YHR SATGLTERFELFVMRKEVCNAYTELNDPAV'QRER FEQQAADKAAGDDEAQLVDENFCTALEYGLPPTGGWGMGIDRLTMF WHR SK SGLTERFELFINKHELCNAYTELNDPVVQRQRFADQLKDRQSGDDEAMALDETFCNALEYGLAPTGGWGLGIDRLSML HSRDQPGLCERFEVFVATKEI CNAYTELNDPFDQRARFEEQARQKDQGDDEAQLIDETFCNALEYGLPPTGGWGCGVDRLAMF YHRQHAGLCERFEAFVCKKEIVNAYTELNDPFDQRLRFEEQARQKDQGDDEAQLIDENFCTSLEYGLPPTGGWGMGIDRLVMF YHRQNVGLCER FEAFVCKKEIVNAYTELNDPFDQR LR FEEQARQKDQGDDEAQI IDENFCTSLEYGLPPTGGWGMGIDRLVMF YHRDVPGLTER FEVYVAKKEICNAYT ELNDPATQR ER FEEQAKNRAAGDDETPPTDEAFCTALEYGLPPTGGWGLGVDRLTMF WHR SK EGLTERFELFVMKKEICNAYTELNDPVRQRQLFEEQAKAKAAGDDEAMFIDETFCTALEYGLPPTGGWGMGIDRVTMF WHR SI PGLTERFELFAVTREIANAYTELNDPITQRQRFEQQAKDKDAGDDEAQMIDETFCNALEYGLPPTGGWGMGIDRLSMI WHR SI PGLTERFELFAVTREI ANAYTELNDPITQRQRFEQQAKDKDAGDDEAQMI DETFCNALEYGLPPTGGWGMGIDRLSMI KDRNI PGLCER FEVFVATKEICNAYT ELNDPFDQRAR FEEQARQKAQGDDEAQMVDETFCNALEYGLPPTAGWGCGIDRLAMF Y SRDQPGLCER FEVFVATKEICNAYTELNDPFDQRAR FEEQARQKDQGDDEAQLIDETFCNALEYGLPPTGGWGCGIDRLAMF WHR SK EGLTER FELFVMKKEI CNAYTELNDPVRQRQLFEEQAKAKAAGDDEAMFIDENFCTALEYGLPPTAGWGMGIDRVTMF WHR SIKGLTERFELFVNKKEICNAYTELNDPMI QRQR FEQQA LDKAAGDDEAQMVDENFCTALEYGLPPTGGWGMGIDRLTMF YDR SR PG LC ER F EAF LCTK EI CNAYT ELNDPFDQR ER FMEQVRQK EQGDEEAQGVDETFLDALEYGLPPTGGWGLGIDR LVMF WHR EK P EMT ER FELFVLGK ELCNAYT ELNEPLQQRK FF EQQADAKASGDVEACPIDETFCLALEHGLPPTGGWGLGIDRLIMF WHR EK PEMT ER FELFVLGKELCNAYTELNEPLQQRKFFEQQADAKA SGDVEACPIDETFCLALEHGLPPTGGWGLGIDRLIMF WHR SQK LT ER FELFVMKKEICNAYT ELNDPIRQR ELFEQQAKAKA EGDDEAMFIDETFCTALEYGLPPTAGWGMGIDRLTMF KDRDNVGLCER FEVFVATKEI CNAYTELNDPFDQRQRFEEQARQKAQGDDEAQMVDETFCNALEYGLPPTAGWGCGIDRLAMF YHR EK PQLTER FELFVNTKEI CNAYTELNNPFVQI ERFAEQAKAKAAGDDESMLIDKVFTTSLEYGLPPTGGFGLGIDRFAML YHR SI PGLTERFELFVAKKEI CNAYTELNDPVVQRERFEQQA SDKAAGDDEAQLVDENFCTSLEYGLPPTGGFGMGIDRLAMF YHR SAPGLTERFELFVAKKEICNAYTELNDPVVQRERFEQQASDKAAGDDEAQMVDENFCTALEYGLPPTGGFGMGIDRLTMF NHR SKAGLTERFELFINCKEICNAYTELNNPFEQRERFLQQTQDLNAGDDEAMMNDEDFCTALEYGLPPTGGWGIGIDRLVMY YHR SEPELTERFELFI LKREIANAYTELNNPIVQR SNFEQQAKDKAAGDDEAQLVDEVFLDAI EHAFPPTGGWGLGIDRLAML WHRI HRGLTER FELFVMKK EVCNAYT ELNDPFQQRQLFEDQAKAKAAGDDEAMFIDENFCTALEYGLPPTAGWGMGIDRFTMF WHR SK EG LT ER FELFVMKKEI CNAYT ELNDPMRQRQLFEEQAKAKAAGDDEAMFIDENFCTALEYGLPPTAGWGMGIDRVAMF Y SRDQPGLCERFEVFVATKEICNAYTELNDPFDQRARFEEQANQKAQGDDEAQLVDETFCNALEYGLPPTGGWGCGIDRLAMF WHRNDPR LT ER FELFVNKK ELANAYTELNNPIVQREEFLKQVRNR WHR SK EGLTERFELFVMKKEICNAYT ELNDPVRQRQLFEEQAKAK YHRTEKGI SERFEGFVCKKEICNAYTELNNPFDQR LRFEEQARQK NHR EK EGFVERFELFLNGWELANGY SELNDPLEQEKRFEEQDKK WHR SKNGLTER FELFVMKKEI CNAYT ELNDPVRQRQLFEEQAKA WHR SK EGLTERFELFVMKKEICNAYTELNDPVRQRQLFEEQAKA YHR SKNG LCERFEAFVCKKEIANAYT ELNNPFDQR LR FEEQARQ WHR SR PGLTERFELFVNKHEVCNAYTELNDPVVQRQRFEEQLKD YHRNI PGMT ERFELFVNTKELCNAYTELNDPI DQR ERFDEQAKA WHR SK EGLTERFELFVMKKEI CNAYTELNDPMRQRQLFEEQAKA VHRQYPGLTERFELFVNYHELCNAYTELNDPFVQKALFQKQVEDA KDRNI PGLCERFEVFVATKEICNAYTELNDPFDQRQR FEEQARQ YHRTK PGLTERLEMFICGKEVLNAYTELNDPFKQK EC FK LQQKDR YHR SK PG LT ER LEMFICGKEV LNAYTELNDPFKQK EC F SAQQKD YHR SK PG LT ER LEMFI CGKEV LNAYTELNDPFKQK ECFA SQQKD WHR SK PG LT ER FELFVNKHELCNAYTELNDPVVQRQR FEAQLKD WHRCKEGLTERFELFVMKKEICNAYTELNDPVRQRQLFEEQAKAK Y SRDQPGLCERFEVFVATKEICNAYTELNDPFDQRARFEEQARQK YHR SDAGLCER FEAFVATKEI CNAYTELNDI FDQRARFEEQARQ WHR S I PGLTERFELFVARKEI CNAYTELNDPMVQRERFATQAKD WHR SEKGLTERFELFVMKKEICNSYTELNDSVRQRELFEQQAKA YHR SK PNVTER FELFVNYYELCNAFTELNDPFKQRKI FVQQI EEK WHR SK ENVCERFELFICGKELINSYTELNDPITQRECFKQQQKA WHR SK ENVCERFELFICGKELINSYTELNDPITQRDCFKQQQKA WHRTKPGIVERFEVFINGLEYANAYT ELNC PMVQR ELFLDQLKAK WHRKDLRLSERFELFINKKEICNSYTELNSPLVQREEFERQLRDR WHRNDPQLTERFELFLNKKELCNAYTELNNPIVQREEFMKQLRN RHRDI PGLCERFEVFVATKEICNAYTELNDPWVQRANFEEQSRQ YSRDR PGICER FEVFVATKEICNAYTELNDPFDOROR FEEQARQKDADDEAQGIDHVFIDALEHGLPPTGGWGLGIDRLVMF

Anopheles_gambiae/1-3220
Arabidopsis_thaliana/1-3286
Ashbya_gossypii/1-3281
A fumigatus/1-3241
Aspergillus_niger/1-3300 Bombyx_mori/1-2389
Bos_taurus/1-3273
_briggsae/1-3231
Caenorhabditis_elegans/1-3289
Candida_albicans/1-3266
Candida_glabrata/1-3282
Canis_familiaris/1-3286
Ciona intestinalis/1-2712
Cryptococcus_neoformans/1-3301 Cryptosporidium_hominis/1-3020 C_parvum/1-3281
Danio_rerio/1-3286
Debaryomyces hansenii/1-3166
Dictyostelium_discoideum /1-3275 D_melanogaster/1-3287
Drosophila_pseudoobscura/1-3286 Encephalitozoon cuniculi/1-3199 Entamoeba_histolytica/1-3271 Gallus_gallus/1-3286
Homo_sapiens/1-3286
Kluyveromyces_lactis/1-3281 Leishmania_major/1-3286 Macaca_mulatta/1-3182 Magnaporthe_grisea/1-3286 Methanosarcina_acetivorans/1-3043 Monodelphis_domestica/1-3196
Mus musculus/1-3286
Neurospora_crassa/1-3264
Oryza_sativa/1-3218
Ostreococcus_lucimarinus/1-3287
Pan_troglodyte/1-3286
Paramecium_tetraurelia/1-3282
Pichia_stipitis/1-3290
_falciparum/1-3285
P_knowlesi/1-3284
Plasmodium_yoelii/1-3285
Populus_trichocarpa/1-3286
Rattus_norvegicus/1-3286
Saccharomyces_cerevisiae/1-3284
Schizosaccharomyces_pombe/1-3284
Strongylocentrotus_purpuratus/1-3252
Takifugu_rubripes/1-2930
Tetrahymena_thermophila/1-3270
T_annulata/1-3248
Theileria parva/1-3273
Trichomonas vaginalis/1-3198
T_brucei/1-3264
Trypanosoma_cruzi/1-3266
stilago maydis/1-3291
Yarrowia_lipolytica/1-3278

LTDSNNIKEVLLFPAMKPRQAA EAHRQTRQYMQ-KWI KPGMTMI EICEELENTARGLAFPTGCSRNHCAAHYTPNAGD-PTVL LTDSLNIKEVLFFPAMR PRRAA EVHRQVRKYVR-SIVKPGMLMTDICETLENTVRGIAFPTGCSLNWVAAHWTPNSGD-KTVL LTDSNTIREVLLFPTLKPRKGAEI HRRVRRHLQ-NR LR PGQTLTEVVELVENATRGIGFPTGVSLNHCAAHFTPNAGD-TTVL LTDNY SIKEV LAFPFMKERQAA EVHRQVRQYAQ-KTIK PGQTLTEIAEGI EDAVRGMGFPCGLSINHCAAHYTPNAGN-KMVL LTDNY SI K EV LAF PFMK ERQAA EVHRQVRQYAQ-KTIK PGQTLTEIAEGI EESVRGMGFPCGLSINHCAAHYTPNAGN-KMVL LTD SNNIK EV LLFPAMK PRHAA EAHRQTRKHIR-NWIKPGMTMIDICEELEKTARGLAFPTGCSRNHCAAHYTPNTGD-TTVL LTD SNNIK EV LLFPAMK PR EAA EAHRQVRKYVM-SWI K PGMTMI EI CEKMEDCSRG LAFPTGCSLNNCAAHYTPNAGD-TTVL LT DNNNIK EV LLFPAMR PRR SA EAHRQVRQYVK - SWI K PGMSMI EI CERLETTSRG LAFPTGCSLNHCAAHYTPNAGD-TTVI LT DNNNIK EV LLFPAMR PRR SA EAHRQVRKYVK-SWI KPGMTMI EI CER LETTSRGLAFPTGCSLNHCAAHYTPNAGD-TTVL LTD SNTI REVLLFPTLKPRKGAEI HRRVRHKAQ-SSIRPGMTMI EI ANLI EDSVRGIGFPTGLSLNHVAAHYTPNTGD-KLI L LTDSNTIREVLLFPTLKPRKGAEI HRRVRKNVQ-NK LKPGMLLTEVADII ENATRGIGFPTGLSLNHCAAHYTPNTGD-KTVL LTD SNNIK EV LLFPAMK PR EAA EAHRQVRKYVM-SWI KPGMTMI EICEKLEDCSRGLAFPTGCSLNNCAAHYTPNAGD-TTVL LTDSNNIK EV LLFPAMK PRQAA ETHRQVRHHVQ-EFIKPGLSMI EICERLEQASRGLAFPTGCSLNNCAAHYTPNAGD-KTVL LTDCSNIKEVLLFPAMR PRRAGEVHRQVRAYAQ-KAIKPGMTMTEIANLI EDGTRGIGFPTGLSVNEVAAHYTPNPGD-KQVL LADKNNIKEVI LFPAMRNRRAAEVHRQVRKYMQ-SII RPEMKLI DMCNI LESKVKGWGFPTGCSLNHCAAHYTPNPHD-FTKL LADKNNIKEVI LFPAMRNRRAAEVHRQVRKYMQ-SIIRPEMKLIDMCNI LESKVKGWGFPTGCSLNHCAAHYTPNPHD-FTKL LTD SNNIK EV LLFPAMK PRQAA EAHRQVRKYVQ-SWI K PGMTMI EI CEK LEDC SRG LAFPTGCSLNHCAAHYTPNAGD-PTVL LTDSNTIREVLLFPT LKPRKGAEI HRRVRHKAQ-SSIRPGMNMTEIADLI ENSVRGIGFPTGLSLNHVAAHYTPNAGD-KTVL MSDTYNIKEVI LFPAMK PRRAA EVHRQVRKYVQ-GIVKPGLGLTELVESLENASRGIAFPTGVSLNHIAAHFTPNTGD-KTVL LTD SNNIK EV L LF PAMK PR QAA EAHRQTRQYMQ-RYIKPGMTMI QICEELENTARGLAFPTGCSLNHCAAHYTPNAGD-PTVL LTDSNNIK EV LLFPAMK PRQAA EAHRQTRQYMQ-RFIKPGMTMI QI CEELENTARGLAFPTGCSLNHCAAHYTPNAGD-PTVL LTDAANIRDVI FFPTMK PRRAAEAHRRARYRVQ-SIVRPGITLLEIVRSI EDSTRGIGFPAGMSMNSCAAHYTVNPGEQDIVL LADVDNIKEVI LFPTMR PRKAAAI HK SVRQWAQ-QWI KPGMSDLFVAENI ERKVRGMAFPCGLSVNSCAAHFTPNPNDPLSFY LTDSSNIK EVLLFPAMK PR EAA EAHRQVRKYVM-SWI K PGMTMI EICEK LEDCSRGLAFPTGCSLNNCAAHYTPNAGD-PTVL LTD SNNIK EV LLFPAMK PR EAA EAHRQVRKYVM-SWI KPGMTMI EI CEK LEDC SRGLAFPTGCSLNNCAAHYTPNAGD-TTVL LADSNTIREVLLFPTLKPRKGAEI HRRVRESVR-NKIKPGMTLTEIANLVEDGTRGIAFPTGLSLNHCAAHFTPNAGD-KTVL LT SQ SNI K EV L LF PAMK PR EAA EVHRQVRTWAQ-SWI K PGLSLMLMTDRI EKK LNGQAFPTGCS LNHVAAHYT PNTGDEKVVL LTDSNNIK EV LLFPAMK PR EAA EAHRQVRKYVM-SWI K PGMTMI EI CEKLEDCSRGLAFPTGCSLNNCAAHYTPNAGD-TTVL LTNNATIREV LAFPFMRDRHGAEAHRQARRWAH-KHVKPGMSLTDIANGI EDSVRGMGFPTGLSINHCAAHYTPNAGN-KMVL LAGLESIKEVI LFPQMKRREAGRI LKIVRTEAA -DMI RVGNSLLEVAEFVEKKTI - -AFPCNI SRNQEAAHATPKAGD-QDVF LT D SNNIK EV LLFPAMK PR EAA EAHRQVRKYVM-SWI K PGMTMI EICEK LEDC SRG LAFPTGCSLNNCAAHYTPNAGD-PTVL LTD SNNIK EV LLFPAMK PR EAA EAHRQVRKYVM-SWIK PGMTMI EICEK LEDCSRGLAFPTGCSLNNCAAHYTPNAGD-TTVL LTDNY SIR EV LAFPFLRERHAA EVHRQVRQWAQ-KSIKPGQTLTEIAENI EDSVRGMGFPTGLSINHCAAHYTPNAGN-KMVL LTD SQNIK EV LLFPAMK PRRAA EVHRQVRKHMR -SI LK PGMLMI DLCETLENMVRGIAFPTGCSLNWVAAHWT PNSGD-KTVL LADKNNIK EV LLFPAMK PRQCA EVHR EVRQYIS-DWVKPGMKYI DVCET LENSVRGVAFPTGCSKNHVAAHWT PNGGC-ESVI LTD SNNIK EV LLFPAMK PR EAA EAHRQVRKYVM-SWI K PGMTMI EI CEK LEDC SRG LAFPTGCSLNNCAAHYTPNAGD-TTVL LTDTQNI QEV LLFPAMK PRKAA ECHRQVRQYAQAK L LK PGNKLIDICEKLEDMNRGI AFPTGCSLNFCAAHYTPNNGD-NTI L LTDSNTIREVLLFPTLKPRKGAEI HRRVRHKAQ-SSI RAGMSMTEIADLI ENSVRGIGFPTGLSLNHVAAHYTPNTGD-KLSL LTNKNSIKDVI LFPTMR PRKAA ECHRQVRKHMQ-AFIKPGKKMIDIAQETERKTKGWGFPTGCSLNHCAAHYTPNYGD-ETVL LTNKNCIKDVI LFPTMR PRKAAECHRQVRKYIQ-AYVQPGRKMIDIVKETEKKTKGWGFPTGCSLNHCAAHYTPNYGD-ETVL LTNKNSIKDVI LFPTMK PRKAAECHRQVRKYIQ-SYIKPGRKMIDIVQKTEQKTKGWGFPTGCSLNNCAAHYTPNYGD-ETVL LTDSQNIK EV LLFPAMK PRQAA EVHRQVRKYMK - SI LK PGMLMMDLCET LENTVRGIAFPTGCSLNWVAAHWTPNSGD-KTVL LTD SNNIK EV LLFPAMK PR EAA EAHRQVRKYVM-SWIKPGMTMI EICEKLEDCSRGLAFPTGCSLNNCAAHYTPNAGD-TTVL LTD SNTI REVLLFPT LKPRKGAEI HRRVRRAIK-DRIVPGMK LMDI ADMI ENTTRGIGFPTGLSLNHCAAHFTPNAGD-KTVL LTD SNT I REV L LF PHMK PRRAA EVHRQARQYAQ-SVIK PGMSMMDVVNTI ENTTRGIGFPTGVSLNHCAAHYTPNAGD-TTI L LTD SNNIK EV LFFPAMK PRQAA EAHRQVRKHVQ-GFIKPGMTMIDICERLETASRGLAFPTGCSRNHCAAHYTPNAGD-TTVL LTD SNNIKEVLLFPAMK PRQAA EAHRQVRAYVR-SWIK PGMTMIDICEKLEDCSRGLAFPTGCSINHCAAHYTPNAGD-PTVL LTDNIYI QEV LLFPAMKPRKAA ECHRQVRKYCQ-QLIRPGKKLIDICESI EEMNRGIAFPTGCSLNHVAAHYTPNNGD-FTTI LSDKNNI KVFIGVIIIV-RRAA EVHRQVRRYI Q-SVIR PGVSCLDIVQAVESKTKGWGFPTGCSLNSCAAHYTPNYGD-KTVF LADKNNIKEVI FFPTMR PRKAA EVHRQARRYIQ-SVIKPGLSCLDIVQALEFKTKGWGFPTGCSLNSCAAHYTPNHGD-KTIF LTNQV SI R EV LLFPLMK PR EGAEI HRRVRRWAMENVIKPGVK LYDMCAQI EEAVRGLAFPCGCSINNCAAHYTPMYNTDQRVL LT SQNNIK EV LLFPAMK PRCAA EVHRQVRRYAQ-SFIKPGI SLLSMTDRI EKKLEGQAFPTGCSLNHVAAHYTPNTGD-KCVL LT SQANIK EV LLFPAMK PRHAA EVHRQVRRYAQ-SFIKPGI SLI SMTDRI ERKVEGQAFPTGCSLNHVAAHYTPNTGD-KTVL LTDSNSIKEVLAFPANK PRRAA EVHRQVRQYAQ-SAIKPGMTMTEIAELVEDGTRGIGFPTGVSVNECAAHYTPNAGD-KRVL LTNSNTIKEVLLFPAMK PRKGAEI HRVVRKYAR-DNI KAGMTMTSIAEMI EDSVRGQGFPTGVSLNHCAAHYTPNAGD-KIVL
2990
3000

Anopheles_gambiae/1-3220
Arabidopsis_thaliana/1-3286
Ashbya_gossypii/1-3281
A fumigatus/1-3241
Aspergillus_niger/1-3300
Bombyx_mori/1-2389
Bos_taurus/1-3273
C_briggsae/1-3231
Caenorhabditis_elegans/1-3289
Candida_albicans/1-3266
Candida_glabrata/1-3282
Canis_familiaris/1-3286
Ciona intestinalis/1-2712
Cryptococcus_neoformans/1-3301 Cryptosporidium_hominis/1-3020 C_parvum/1-3281
Danio_rerio/1-3286
Debaryomyces hansenii/1-3166
Dictyostelium_discoideum /1-3275
D_melanogaster/1-3287
Drosophila_pseudoobscura/1-3286
Encephalitozoon_cuniculi/1-3199
Entamoeba_histolytica/1-3271
Gallus_gallus/1-3286
Homo_sapiens/1-3286
Kluyveromyces_lactis/1-3281 Leishmania_major/1-3286
Macaca_mulatta/1-3182
Magnaporthe_grisea/1-3286
Methanosarcina_acetivorans/1-3043
Monodelphis_domestica/1-3196
Mus musculus/1-3286
Neurospora_crassa/1-3264
Oryza_sativa/1-3218
Ostreococcus_lucimarinus/1-3287
Pan_troglodyte/1-3286
Paramecium_tetraurelia/1-3282
Pichia_stipitis/1-3290
_falciparum/1-3285
P_knowlesi/1-3284
Plasmodium_yoelii/1-3285
Populus_trichocarpa/1-3286
Rattus_norvegicus/1-3286
Saccharomyces_cerevisiae/1-3284
Schizosaccharomyces_pombe/1-3284
Strongylocentrotus_purpuratus/1-3252
Takifugu_rubripes/1-2930
etrahymena_thermophila/1-3270
T_annulata/1-3248
Theileria parva/1-3273
Trichomonas vaginalis/1-3198
T_brucei/1-3264
rypanosoma_cruzi/1-3266
stilago maydis/1-3291
Yarrowia_lipolytica/1-3278

LYDDVTKI DFGTHIKGRI I DCAFTLSFNP--KYDKLLEAVKEATETGIREAGIDVR LCDIGAAIQEVMESYEVELDGKTYQVK QY DDVMK LDFGTHI DGHI I DCAFTVAFNP--MFDPLLAASREATYTGIKEAGI DVRLCDIGAAIQEVMESYEVEINGKVFQVK RHEDVMKVDFGVQVNGHI I DSAWTVTFDP--RYDPLLEAVREATYTGIREAGIDVRLTDIGEAIQEVMESYEVTLGGQTYQVR QQGDVMKVDFGAHINGRIVDSAFTMTFDP--VYDPLLEAVKDATNTGIREAGIDVRMSDIGAAIQEAMESYEVELNGTMYPVK QQGDVMKVDFGAHINGRIVDSAFTVAFDP--VYDPLLAAVKDATNTGIREAGI DVRMSDIGAAI QEAMESYEVEINGTMYPVK EYDDVVKI DFGTHINGRI I DCAFTLHFNP--RYDPLVKGVQEATEAGIKASGVDVR LCDVGAAVQEVMESHEVELDGQMY QYDDICKI DFGTHI SGRI I DCAFTVTFNP--KYDTLLKAVKDATNTGIKCAGIDVRLCDVGEAIQEVMESYEVEIDGKTYQVK QYGDVCKI DYGI HVRGRLI DSAFTVHFDP--KFDPLVEAVKEATNAGIRESGIDVRLCDVGEVVEEVMTSHEVELEGKTYVVK QYGDVCKI DYGI HVRGRLIDSAFTVHFDP--KFDPLVEAVREATNAGIKESGIDVRLCDVGEIVEEVMTSHEVELDGKSYVVK KKDDI MKVDI GVHVNGRICDSAFTMT FNEDGKYDTI MQAVKEATYTGIKESGIDVR LNDIGAAIQEVMESYEMEENGKTYPIK KY EDVMKVDI GVQVNGHIVDSAWTVSFDP--QYDNL LAAVKDATYTGIKEAGI DVRLTDIGEAIQEVMESYEVEIKGKTYQVK QYDDICKIDFGTHI SGRI I DCAFTVTFNP--KYDTLLKAVKDATNTGIKCAGI DVRLCDVGEAIQEVMESYEVEIDGKTYQVK SYDDVCKIDFGTHINGRI I DCAFTVSFNP--KYDR LLEAVKDATNTGIKNAGIDVRLCDVGAAIQETMESYEVEIDGKTYQVR QQHDVMKVDFGVHVNGRIVDSAFTMSFEP--TWDKLLEAVKDATNTGIREAGIDVRMCDIGEAIQEVMESYEVEVNGKVYPVK TQDDI CK LDFGVQVNGMI I DCAFTVAFND--VFDPLI QSTLDATNTGLKVAGIDVMFSEIGSAI EEVIKSYEFEYKSKVYNIK TQDDI CK LDFGVQVNGMI I DCAFTVAFND--VFDPLIQSTLDATNTGLKVAGIDVMFSEIGSAI EEVIKSYEFEYKSKVYNIK QYDDVCKIDFGTHINGRI I DCAFTVTFNP--KYDKLLEAVKDATNTGIKCAGIDVRLCDIGESIQEVMESYEVDLDGKTYQVK NY EDVMKVDIGVHVNGHIVDSAFTLTFDD--KYDSLLKAVKEATNTGVKEAGI DVR LNDIGEAIQEVMESYEMELNGKTYPIK KKDDVLKI DFGTHVNGYI I DCAFTVTFDE--KYDK LKDAVREATNTGIYHAGIDARLGEIGAAIQEVMESHEIELNGKTYPIR QYDDVCKIDFGTHIKGRI I DCAFTLTFNN--KYDKLLQAVKEATNTGIREAGIDVRLCDIGAAIQEVMESYEIELDGKTYPIK QYDDVCKIDFGTHIKGRI I DCAFTLTFNN - -KYDK LLQAVKEATNTGIKEAGI DVRLCDIGAAIQEVMESYEVELDGKTYPIK K EDDV LK I DFGTHSDGRIMDSAFTVAFKE--NLEPLLVAAREGTETGIKSLGVDVRVCDIGRDINEVISSYEVEIGGRMWPIR KTDDVVKI DFGVHVNGHLI DSAFTMTWDP--ALQPI LDCSKDATNTGIKNIGVDVRLCDIGDAI EEVMSSYEVEIKGKTYQLQ HYDDICKI DFGTYYSGRI I DCAFTVTFNP--KYDRLLEAVKDATNTGIKCAGIDVRLCDVGEAIQEVMESYEVEIDGKTYQVK QYDDICKIDFGTHI SGRII DCAFTVTFNP--KYDTLLKAVKDATNTGIKCAGIDVR LCDVGEAIQEVMESYEVEIDGKTYQVK K FEDVMKVDFGVHVNGYI I DSAFTIAFDP--QYDNLLAAVKDATNTGIKEAGIDVRLTDIGEAIQEVMESYEVEINGETHQVK TYDDVMKVDFGTHINGRI I DCAWTVAFNP--MFDPLLQAVKEATYEGIKQAGIDVR LGDIGAAI E EVMESHEVEINGKVHQVK QYDDICKIDFGTHI SGRI I DCAFTVTFNP--KYDTLLKAVKDATNTGIKCAGIDVR LCDVGEAIQEVMESYEVEIDGKTYQVK EHDDV LKVDIGVHVNGRIVDSAFTVAFNP--RYDNLLAAVKDATNTGIREAGIDARLGEIGEAIQETMESYEVEIDGETYPVK G-NDMVK LDLGVHVDGYI ADSAVTVDLSG--NSD-IVKASEEALAAAIDLMK PGVSTGEIGAAIEERIHS--------YGLK QYDDICKIDFGTHI SGRI I DCAFTVTFNP--KYDTLLKAVKDATNTGIKCAGIDVR LCDVGEAIQEVMESYEVEIDGKTYQVK QYDDICKIDFGTHI SGRII DCAFTVTFNP--KYDI LLTAVKDATNTGIKCAGIDVRLCDVGEAIQEVMESYEVEIDGKTYQVK QEDDVMKVDFGVHVNGRIVDSAFTVAFNP--RYDPLLEAVKAATNAGIKEAGI DVRVGDIGAAIQEVMESYEVEINGQMLPVK QYDDVMK LDFGTHI DGYIVDCAFTVAFNP--MFDSLLQASKDATNTGVKEAGI DAR LCDVGAAIQEVMESYEVEINGKVFQIK DKDDVIKFDFGVQVKGRI I DCAFTKTFND--MYDPLLKAVNEATETGIRSAGIDVRLCDIGEAVQEVMESHTVEI HGKEYQVK QYDDICKIDFGTHI SGRIIDCAFTVTFNP--KYDTLLKAVKDATNTGIKCAGIDVRLCDVGEAIQEVMESYEVEIDGKTYQVK TYDDVCKI DFGTQVDGWI I DCAFTVAFNP--VYDTLLQAAKDATDTGIRNSGIDVR LGDVGAAIQETMESYEVEIGGKVYKVK GKDD LMKVDI GVHVNGHI CDSAFTMT LNDTGKYDSI MKAVKDATNTGVKEAGI DVR LNDIGEAIQEVMESYEMELDGKTYPVK KYDDVCK LDFGVHVNGYI I DCAFTIAFNE--KYDNLIKATQDGTNTGIKEAGIDARMCDIGEAIQEAIESYEIELNQKIYPIK KYDDVCK LDFGVHVNGYI I DCAFTIAFNE--KYDNLIKATQDGTNTGIREAGIDARMCDIGEAIQEAIESYEI ELNKKIYPIK K EDDVCKLDFGVHVNGYI I DCAFTIAFND--KYDNLIKATQDGTNTGIKEAGIDARMCDIGEAIQEAIESYEIELNQKVYPIK QY DDVMK LDFGTHI DGHIVDCAFTVAFNP--MFDPLLEASREATNTGIKESGI DVR LCDVGAAIQEVMESYEVEINGKVFQVK QYDDICKI DFGTHI SGRII DCAFTVTFNP--KYDI LLKAVKDATNTGIKCAGIDIRLCDVGEAIQEVMESYEVEIDGKTYQVK KY EDVMKVDYGVQVNGNI I DSAFTVSFDP--QYDNLLAAVKDATYTGIKEAGIDVRLTDIGEAIQEVMESYEVEINGETYQVK K EK DVMKVDI GVHVNGRIVDSAFTMSFDP--QYDNLLAAVKAATNKGI EEAGIDAR LNEIGEAIQEVMESYEVEINGKTHQVK EYDDVCKI DFGTHINGRII DCAFTVTFNP--KYDQLLAAVKDATNTGIKEAGIDVRLCDVGERIQEVMESYEVELDGKTYQVK RYDDVCKI DFGTHINGRI I DCAFTVTFNP--KFDGLLEAVRDATNTGIKFAGIDVRLCDVGETIQEVMESYEVEIDGKTYQVK EYDDVCKI DFGTQVEGRI I DCAFTVAFNP--KYDKLLEAVKEATNTGIKEAGIDVRIPDVGAAIQEVMESYEVEI EGKTYPVK EKDDI MK LDFGTHVNGYI I DSAFTIAFDE--KYDPLI ESTKEATNTGLKLAGIDARTSELGEAIEEVIESFEITLKNRTHKIK HKNDVMK LDFGTHVNGYI I DSAFTIAFDE--KYDPLI ESTKEATNTGVKLAGIDARTSELGEAIQEVIESYEITLKNKTHKIK GK SDVMKI DFGVAI NGNI I DSAFTVCFDP--KFEPLLEAAKTATNTGVKIAGIDARMNEIGDAIQEVFDASSIDIDGKHYDIK MY DDVMKVDFGTQI NGRI I DCAWTVAFKD--EYEPLLTAVKEATY EGVKQAGI DVR LCDVGAAIQEVMESYEVELNGKVYPVK TYDDVMKVDFGTQI NGRIVDCAWTVAFND--EYAPLLEAVKSATYEGIKQAGIDVRLCDIGEAIQEVMESYEVEIKGKVYPVK QATDV LKVDFGVHVKGRIVDSAFT LNFEP--TWDPLLAAVKAATNAGIKEAGIDARLGEIGASIQEVMESHEFEANGKTHRVK K EDDV LKVDFGVHVNGKI I DSAFTHVOND - -KWQGLLDAVKAATETGIREAGIDVRLGDIGEAIOETMESHEVEVDGKVYOVK

Anopheles_gambiae/1-3220
Arabidopsis_thaliana/1-3286
Ashbya_gossypii/1-3281
A fumigatus/1-3241
Aspergillus_niger/1-3300
Bombyx_mori/1-2389
Bos_taurus/1-3273
C_briggsae/1-3231
Caenorhabditis_elegans/1-3289
Candida_albicans/1-3266
Candida_glabrata/1-3282
Canis_familiaris/1-3286
Ciona intestinalis/1-2712
Cryptococcus_neoformans/1-3301 Cryptosporidium_hominis/1-3020 C_parvum/1-3281
Danio rerio/1-3286
Debaryomyces hansenii/1-3166
Dictyostelium_discoideum/1-3275
D_melanogaster/1-3287
Drosophila_pseudoobscura/1-3286
Encephalitozoon_cuniculi/1-3199
Entamoeba_histolytica/1-3271
Gallus_gallus/1-3286
Homo_sapiens/1-3286
Kluyveromyces_lactis/1-3281
Leishmania_major/1-3286
Macaca_mulatta/1-3182
Magnaporthe_grisea/1-3286
Methanosarcina_acetivorans/1-3043
Monodelphis domestica/1-3196
Mus musculus/1-3286
Neurospora_crassa/1-3264
Oryza_sativa/1-3218
Ostreococcus_lucimarinus/1-3287
Pan troglodyte/1-3286
Paramecium tetraurelia/1-3282
Pichia_stipitis/1-3290
_falciparum/1-3285
P_knowlesi/1-3284
Plasmodium_yoelii/1-3285
Populus_trichocarpa/1-3286
Rattus_norvegicus/1-3286
Saccharomyces_cerevisiae/1-3284
Schizosaccharomyces_pombe/1-3284
Strongylocentrotus purpuratus/1-3252
Takifugu_rubripes/1-2930
Tetrahymena_thermophila/1-3270
T_annulata/1-3248
Theileria parva/1-3273
Trichomonas_vaginalis/1-3198
T_brucei/1-3264
Trypanosoma_cruzi/1-3266
stilago maydis/1-3291
Yarrowia_lipolytica/1-3278

3080
3090
3100
AI RNLNGHSI SPYRI HA

3120
-MEENEFYAI ETFGS MEEGEFYAI ETFGS MEEGEHFAI ETFGT MEEGETFAI ETFGS MEEGEVFAI ETFGS MEENEIYAI ETFGS MEEGEVYAI ETFGS MEENEIYAI ETFG MEENEIYAI ETFGS MEEGETFAI ETFGS MEEGEHFAI ETFGT MEEGEVYAI ETFGS MEEGEVFAI ETFG MEEGEYFAI ETFGS ME ENEIYAI ET FAT MEENEIYAI ETFAT MEEGEVYAI ETFG MEEGETFAI ETFGS MEEGEFYAI ETFGS MEEDEFYAI ETFGS MEEDEFYAI ETFGS KGGSFYAVETFAT MEEGEVYALETFAT MEEGEVYAI ETFGS MEEGEVYAI ETFGS MEENEHFAI ETFG MEEGEVFAI ETFGS MEEGEVYAI ETFGS MEEGEIYAI ETFG LKEGDVLAI EPFAT MEEGEVYAI ETFGS MEEGEVYAI ETFGS MEEGDVFAI ETFGS MEEGEFYAI ETFGS MEEGEYYAI ETFGT MEEGEVYAI ETFG MKEGELYAI ETFGS MEEGETFAI ETFGS MEEGELFAI ETFAS MEEGELFAI ETFAS MEEGDLFAI ETFAS MEEGEFFAI ETFAS MEEGEVYAI ETFG MEEGEHFAI ETFGS MEEGEI FAI ETFG MEENEFYAI ETFGS MEEGDVYAI ETFGS MEEGEQYAI ET FGVI MEEGDVFAI ETFAT MEEGEVFAI ET FAT MEEGELFACETFGS MEEGELFAI ETFGS MEEGELFAI ETFGS MEEGEYYAI ETFGS MEEGETFAI ETFGS

31303140

TGRGLV SHY MKD FDA P - - -KV PLRL TGKGYV SHYMKNFDAG - - -HV PLRL TGRGYV SHYAKNPGAL-- - PAPTL TGKGYV SHYALI PDAP - - - SV PLRL TGKGYV SHYALI PDHS -- -QV PLRL TGRGQV SHY MKN FDQQ - - - FV PLRL TGKGVV SHYMKNFDVG---HVPIRL GKGYV SHY MKNFELAD - EK I PLR TGKGYV SHY MKNF ELAD - EK I PLRL TGNGYV SHYAMNKGVE---HLKPPS TGRGYV SHY AR LPSDG---LPQPNL TGKGVVSHYMKNFDVG---HVPIRL TGKGYV SHY MKN FD LAN - QHV PLR L TGRGRV SHYA LN SAA P - - EKYQGHH TGRGYV SHYMKY YDN P F LN EN STRL TGRGYV SHYMKYYDNPF LN ENSTRL TGKGMV SHYMKNFEVG---HVPIRL TGNGYV SHYAKNPGTD---DIVVPG TGRAQV SHY MKTDY - ---- QTTVRL TGRGLVSHY MKNFDLP---FVPLRL TGRGLVSHYMKNFDLP---YVPLRL TGKGSI SHFV LNTYKS-------RK TGRGRV SHY MVDANAF---DY PVRD TGKGVV SHYMKNFDVG---HVPIRL TGKGVV SHY MKNFDVG---HVPIRL TGRGYV SHYAKK PGSH--- PTPSL TGRG FV SHY MMQ PGAE---VMQLRS TGKGVV SHY MKNFDVG---HVPIRL TGLGYV SHYAKRADAP---NVALRL NGTGLVEIYSLIK------KKPVRL TGKGVV SHYMKNFDVG---HVPIR TGKGVV SHY MKNFDVG---HVPIRL TGNGYV SHYAKRGDAA ---KVDLRL TGKGFV SHY MKN FDVG---HV PLRV TGRGYVSHYMKNFDVG---HV PLRL TGKGVV SHY MKN FDVG---HVPIRL TGKGYV SHYMKDFYAK - - - PTAVRV TGRGYV SHY SRNQNI D---GIRVPS TGKGYV SHYMRNPEKQ---FVPIRL TGKGYV SHY MRNPDKQ - - - FVPIR L TGKGFVSHYMRNRDVQ---YAPIRL TGKGYV SHY MKNFDVG---HI PLRL TGKGVV SHY MK N FDVE---HVPIRL TGRGYV SHYAR SA EDH - - -QVMPTL TGRGVV SHYAKI PDAG---HI PLRL GKG FV SHY MKNFEVG-- - HV P LRM TGRGAV SHY MKN FNVG---HVPIRL NGKASI SHYMKDFNKE---MV PLRQ TGSGTV SHYMKNPNSI - - YAPIRL TGSGMV SHY MKNPNSI - - -YAPIRL TGKGIV SH FMV ARN PP-----TPRT TGRG FV SHY MMV PGGE---KTQVR S TGRGVV SHY MMV PGGD---KTQLRS TGRGYV SHYARKKNLP--KSI PIRV

Anopheles_gambiae/1-3220
Arabidopsis_thaliana/1-3286
Ashbya_gossypii/1-3281
A fumigatus/1-3241
Aspergillus_niger/1-3300 Bombyx_mori/1-2389
Bos_taurus/1-3273
C_briggsae/1-3231
Caenorhabditis_elegans/1-3289
Candida_albicans/1-3266
Candida_glabrata/1-3282
Canis_familiaris/1-3286
Ciona intestinalis/1-2712
Cryptococcus_neoformans/1-3301 Cryptosporidium_hominis/1-3020 C_parvum/1-3281
Danio_rerio/1-3286
Debaryomyces hansenii/1-3166
Dictyostelium_discoideum/1-3275 D_melanogaster/1-3287
Drosophila_pseudoobscura/1-3286 Encephalitozoon cuniculi/1-3199 Entamoeba_histolytica/1-3271 Gallus_gallus/1-3286
Homo_sapiens/1-3286
Kluyveromyces_lactis/1-3281
Leishmania_major/1-3286
Macaca_mulatta/1-3182
Magnaporthe_grisea/1-3286
Methanosarcina_acetivorans/1-3043
Monodelphis_domestica/1-3196
Mus musculus/1-3286
Neurospora_crassa/1-3264
Oryza_sativa/1-3218
Ostreococcus_lucimarinus/1-3287
Pan_troglodyte/1-3286
Paramecium_tetraurelia/1-3282
Pichia_stipitis/1-3290
_falciparum/1-3285
P_knowlesi/1-3284
Plasmodium_yoelii/1-3285
Populus_trichocarpa/1-3286
Rattus_norvegicus/1-3286
Saccharomyces_cerevisiae/1-3284
Schizosaccharomyces_pombe/1-3284
Strongylocentrotus_purpuratus/1-3252
Takifugu_rubripes/1-2930
Tetrahymena_thermophila/1-3270
T_annulata/1-3248
Theileria parva/1-3273
Trichomonas_vaginalis/1-3198
T_brucei/1-3264
rypanosoma_cruzi/1-3266
stilago_maydis/1-3291
Yarrowia_lipolytica/1-3278


#### Abstract

'QSSKSLLGL-'-INRNFGT LA FCKRWLDRAGATK PRAKQLLAT ---I NKNFST LAFCRRY LDRIGETK SRAKALLRT---I DANFGTLPWCRRY LDR LGEDK S SAKNLLNV ---I NKNFGTLPFCRRY LDR LGQEK S SAKNLLNV ---I NKNFGT LPFCRRY LDR LGQDK QSSKQLLNV ---I NKNFGT LAFCKRWLERAGASR PRTKHLLNV---INENFGTLAFCRRWLDR LGESK QK SKGLLSL---I DKNFSTLAFCRRW- QK SKGLLNL---I DKNFAT LAFCRRWI DR LGETK ER SKQLLET -- - IKQNFGTLPWCRRYLERTGEEK A SAKQLLKV -- - I DDHFGT L PWCRRY LDR LGQDK PRTKHLLNV ---I NENFGTLAFCRRWLDR LGESK PRAKRLLHV---I NENFGTLAFCRRWI DRIGETK QSAKSLLAS---VKRNFGTLPFCRRYLDHVGEKN NSAK I LLGG---I NTHFGTLAFCRRWLDQLGFNK NSAK I LLGG---I NTHFGT LAFCRRWLDQLGFNK PRAKHLLNV ---VNENFGTLAFCRRWLDR LGETK DKAKSLLNV---I NENFGTLPWCRRY LDR LGQDK PKAKQLLQY ---I NKNYDT LCFCRRWLDRAGEDK QS SKQLLGT---I NKNFGTLAFCKRWLDRAGATK QSSKQLLGT---I NKNFGT LAFCKRWLDRAGATK LFNKDLIKVYEFVKDSLGTLPFSPRHLDYYGLVK GNAKR LLHA---LDANFKTLAFCRRYVDKIGFAK PRAKHLLNV---I NENFGT LAFCRRWLDRLGESK PRTKHLLNV -- - I NENFGTLAFCRRWLDR LGESK SSAKNLLKV---I DENFGTI PFCRRY LDR LGEDK EKAQQLLKH---I HK SYST LAFCRKWLDRDGFDR PRTKHLLNV---I NENFGTLAFCRRWLDR LGESK T SAQK I LNV---I NKNFGT LPFCRRY LDR LGQDK PAVRNVLKQ---V-EEYRELPFAKRWLE---SDK PRAKHLLNV---I NENFGTLAFCRRWLDRLGES PRTKHLLNV---I NENFGTLAFCRRWLDR LGESK SSAK SLLNV ---ITKNFGTLPFCRRYI DR LGQDK AKAKQLLGT---I NNNFGTLAFCRRY LDR LGETK PRAKQLLGV---I DRNFGTLAFCKRYLDRIGEQR PRTKHLLNV---I NENFGTLAFCRRWLDR LGESK PKAK SLLTH---I DNHYDT LAFCRR FLDRDGQSN ERAKTLLNS---ITSNFGTLPWCRRYLERTGEEK NSAKTLLKV---I NDNFDT LPFCNRWLDDLGQTR NSAKT LLKV---I NDNFDT LPFCHRWLDD LGQK NSAKTLLKV---I NDK FDT LPFCNRWLDD LGQT PRAKQLLAT---I NKNFST LAFCRRY LDR LGETK PRTKHLLNV---I NENFDT LAFCRRWLDR LGESK D SAKNLLKT---I DRNFGTLPFCRRY LDR LGQEK PRAKALLNT---I TQNFGTLPFCRRYLDRIGESK QR SKA LLKV---I NNNFGT LAFCRRWLDR LGETK PRAKHLLNV ---I NENFGT LAFCRRWLDR LGESK PKAKNLLKF---I DNNFGT LAFCRRWLDRGGQT K SAR ESLNV---I NREFSTLPFCKRWLDDLTNKR K SAR EA LNV---I NREFSTLPFCKRWLDDLTNR PAARKLLKT---LQENFSTLAFSQRFIDRIGEKK DKAQQLLRH---I HKTYNT LA FARKWLDR DGHDR EK AQHLLKH---I NKTYGT LA FARKWLDRDGYDR H SAHGLLRT---I NKHFDSLPFCRRY LDRVGEKN NKAKQLLAT - - -I DKNFGTLPFCRRYLDR LGEEK

3190 Y LMA LKN LCDSGI VQPY P P LCDVK YMFA LNH LVK QGI VQDY PPLVDVE Y LLG LNN LV S SGI VQDY PPLCDVK Y LLG LNN LV S SGI VQDY PPLCDIK YAMA LKD LCDKGVVDAY PPLCDIK Y LMA LKN LCDLGIVDPYPPLCDIK Y LMA LKD LCDKGI VDPY PPLCDVK LFA LNQLVRHGIVEEYPPIVDKR-GCYTAQWEHTI LMR PTV俭 Y LFA LN S LVKQGHVQDY PPLNDVI -GSYTAQY EHTI LLHPHKK LMA LKNLCDLGIVDPYPPLCDIK-GSYTAQFEHTI LLRPTCK LMA LKN LCDSGVVDAYPPLCDIK-GCYTAQFEHTI LMR PNCK Y L LA LNT LVREDFI ADY PPLVDPQPGAMTAQFEHTI LLRPTCK HA LA LK SLVDSEI IR PYPPLNDI P-GSFS SQMEHTI LLRPSCK HA LA LK SLVDSEI I R PYPPLNDI P-GSFSSQMEHTI LLRPSCK LMA LKNLCDLGI I DPYPPLCDTK-GCYTAQFEHTI LLR PTCK Y L LA LNQLVRAGIVQDYPPIVDIK-GSYTAQFEHTI LLHPHKK HI LA LNN LCDLGI I QRHAPLVDSK-GSYVAQY EHT LLLKPTAK Y QMA LKD LCDKGIVEAY PPLCDIK Y QMA LKD LCDKGIVEAY PPLCDIK GSLK SVNLLTMMGLLTPYPPLNDID WQMPFKFLVDDGCVNAYPPLSDCH Y LMA LKN LCDLGIVDPY PPLCDIK Y LMA LKN LCDLGIVDPY PPLCDIK HVYA LNT LVRQGIVEDYPPLNDIK HLMN LNR LVDEGAVNKY PPLVDVK Y LMA LKN LCDLGI VDPY PPLCDIK Y LLG LNN LV SNGIV EAY PPLVDKK LEFSLIQLEKAGI LHSYPVLVESA Y LMA LKN LCDLGIVDPY PPLCDIK Y LMA LKN LCDLGI VDPYPPLCDIK Y LLG------GIVEAY PPLVDKK Y LMA LKN LCDVGIVQPYPPLCDVR Y SMA LKN LCDNGI VQPY PPLCDIK Y LMA LKNLCDLGIVDPY PPLCDIK Y L LG LKNLCDLGIVNPYPPLCDIR Y LFALNQLVRAGIVEEYPPLVDIK HFMA LKTLIDLNIVEPYPPLCDIK HFMA LKT LVDLNIVEPYPPLCDVK HFMA LKT LVDLNIVEPYPPLCDIK Y LMA LKN LCDSGI I QPY PPLCDVK Y LMA LKN LCDLGIVDPYPPLCDIK Y LFA LNN LVRHGLVQDY PPLNDI P LLA LNNLVSAGIVQDY PPLCDIR Y LMA LKN LCDTG LVDPY PPLCDVK Y LMA LKN LCDLGI I DPYPPLCDIK HI LSLKQLCDAGI VVPYPPLVDVR GSLVLRNLVDAGI IVPYPPLCDNN G SMV LR S LVDAGIVVPYPPLSDNN YQLN LRHLVECRAVHDY PSLSDVK HLLN LNQLVEAGAVNKY PPLCDIR -HLLN LNQLVEAGAVNRYPPLCDVK Y L LG LKH LV S LGVVQDY P P LCDI A Y L LA LKN LVQ SGVVQDY PPLVDQK G SYTAQY EHTI MLR PTCK GSYVSQFEHTI LLRPTCK GSYTAQF EHTI LLHPHKK GSYTAQFEHTI LLRPTVK GSYTAQY EHTIVLRPNVK GCYTAQF EHTI LLRPTCK GSYTAQFEHTI LLR PTCK


Danio_rerio/1-3286
Debaryomyces_hansenii/1-3166
Dictyostelium_discoideum /1-3275 D_melanogaster/1-3287 Drosophila_pseudoobscura/1-3286 Encephalitozoon cuniculi/1-3199 Entamoeba_histolytica/1-3271 Gallus_gallus/1-3286
Homo_sapiens/1-3286
Kluyveromyces_lactis/1-3281
Leishmania_major/1-3286
Macaca_mulatta/1-3182
Magnaporthe_grisea/1-3286
Methanosarcina_acetivorans/1-3043
Monodelphis_domestica/1-3196
Mus musculus/1-3286
Neurospora_crassa/1-3264
Oryza_sativa/1-3218
Ostreococcus_lucimarinus/1-3287
Pan troglodyte/1-3286
Paramecium_tetraurelia/1-3282
Pichia_stipitis/1-3290
_falciparum/1-3285
P_knowlesi/1-3284
Plasmodium_yoelii/1-3285
Populus_trichocarpa/1-3286
Rattus_norvegicus/1-3286
Saccharomyces_cerevisiae/1-3284
Schizosaccharomyces_pombe/1-3284
Strongylocentrotus_purpuratus/1-3252
Takifugu_rubripes/1-2930
Tetrahymena_thermophila/1-3270
T_annulata/1-3248
Theileria parva/1-3273
Trichomonas_vaginalis/1-3198
T_brucei/1-3264
rypanosoma_cruzi/1-3266
stilago maydis/1-3291
Yarrowia_lipolytica/1-3278

|  |
| :---: |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |

Anopheles_gambiae/1-3220 Arabidopsis_thaliana/1-3286
Ashbya_gossypii/1-3281
A fumigatus/1-3241
Aspergillus_niger/1-3300 Bombyx_mori/1-2389
Bos_taurus/1-3273
C_briggsae/1-3231
Caenorhabditis_elegans/1-3289
Candida_albicans/1-3266 Candida_glabrata/1-3282
Canis_familiaris/1-3286
Ciona intestinalis/1-2712
Cryptococcus_neoformans/1-3301 Cryptosporidium_hominis/1-3020 C_parvum/1-3281
Danio rerio/1-3286
Debaryomyces hansenii/1-3166
Dictyostelium_discoideum /1-3275 D_melanogaster/1-3287
Drosophila_pseudoobscura/1-3286 Encephalitozoon_cuniculi/1-3199 Entamoeba_histolytica/1-3271 Gallus_gallus/1-3286
Homo_sapiens/1-3286
Kluyveromyces_lactis/1-3281 Leishmania_major/1-3286 Macaca_mulatta/1-3182 Magnaporthe_grisea/1-3286 Methanosarcina_acetivorans/1-3043 Monodelphis_domestica/1-3196
Mus musculus/1-3286
Neurospora_crassa/1-3264
Oryza_sativa/1-3218
Ostreococcus_lucimarinus/1-3287
Pan_troglodyte/1-3286
Paramecium tetraurelia/1-3282
Pichia_stipitis/1-3290
__falciparum/1-3285
P_knowlesi/1-3284
Plasmodium_yoelii/1-3285
Populus_trichocarpa/1-3286
Rattus_norvegicus/1-3286
Saccharomyces_cerevisiae/1-3284
Schizosaccharomyces_pombe/1-3284
Strongylocentrotus_purpuratus/1-3252
Takifugu_rubripes/1-2930
etrahymena_thermophila/1-3270
T_annulata/1-3248
Theileria parva/1-3273
Trichomonas vaginalis/1-3198
T_brucei/1-3264
rypanosoma_cruzi/1-3266
Ustilago maydis/1-3291
Yarrowia_lipolytica/1-3278

- -MK 'EAAFS LA EAK FASG-DFNQVVLQNV'KAQIKIRTKK'DNVAGVTLPVFGLAKGGQQLQKLKKNYQSAVKLLVELASLQTS - -MKT S SFALTEVKYVAGDNVKHVVLENVK EAT LKVR SRTENI AGVKLPKFGLARGGQQVRACRVAYVKAI EVLVELASLQTS - - MQTASFS LA EVTYATGENI GYQVQESVANAR FKVGARQENVSGVYLPQFGLGRGGQQVQRAKNIYTKVVESLVQLASLQTA - -MQIAAFSLAEVSYAVGGDIGYQVQESAKQARFRVRAKQENVSGVFLPQFGLGKGGQQVQRCRETYARAVETLVELA
- -MQIAAFSLAEVSYAVGGDIGYQI QESAKQARFRVRAKQENVSGVLLPHFGLGKGGMQVQRCRETYARAVETLVELASLQTA
- -MK EAAF S LA EAK FTTG -D FNQVVLQNVTKAQIKIR SKKDNVAGVTLPIFGLARGGQQLAKLKKNFQSAVKLLVELASLQTS
- -MR EAAFS LA EAK FTAG-DFSTTVI QNVNKAQVKIRAKKDNVAGVTLPVFGLARGGEQLAKLKRNYAKAVELLVELASLQTS
-     - MK EAAF SLA EAK FTAG -DF SHTVI QNV SQAQYRVRMKK ENVVGVLLPVFGLGKGGANI AR LKKNYNKAIELLVELATLQTC
-     - MK EAAF SLA EAK FTAG -DF SHTVI QNV SQAQYRVRMKK ENVVGVFLPVFGLGKGGANIARLKKNYNKAI ELLVELATLQTC
- -MQTAAFS LA EVQYATGDNI SYQVQESVQKAR FTVKAKQENVSGVFLPTFALARGGQQVQKAKLIYSKAVETLVELASLQTA
-     - MQTAAFSLAEVSYATGENIGYQVQESV LNAR FKVKARQENVSGVYLPQFGLGRGGQQVQRAKDIYSKAVETLVELASLQTA
- -MR EAAFSLAEAK FTAG-DFSTTVIQNVNKAQVKIRAKKDNVAGVTLPVFGLARGGEQLAKLKRNYAKAVELLVELASLQTS
-     - MK EAAFSLAEAKFSGG-DFSHVV LQNVGKAQMKVR SKT DNVAGVKLPVFGLSRGGEQLSR LKKNY SKAVKLLVELASLQTS - - LQLASFSLAEVTYAAG -DIGYQVQESVRKANYTVQARQENVSGVVLPAFGLSRGGQQI QKSRDTYIKAVGTLVELASLQTA - - I K EA SFA LAKATWAAG -DFKDRII ESCKRPTVTMEVGTENI AGVRLPIFGVASGGQVI QSTREIYMKVLRDLVKLASLQTA - - IK EAS FA LAKATWAAG -DFKDRII ESCKRPTVTMEVGTENIAGVRLPIFGVASGGQVIQSTREIYMKVLRDLVKLASLQTA - - MR EAAFSLAEAK FAAG-DFSTTVI QNVNKAQVKVRAKKDNVAGVT LPVFGLARGGEQLSR LKRNYAKAVELLVELASLQTS - - MQTAAFS LA EVQYATGDNI SY LVQESVQNAR FQVKAKQENV SGVY LPTFGLGRGGQQVQKAKMVYTKAVETLVELASLQTA - -MRDASFSLAAAKYAAG -EFSNSVI ENV SNPTIAVKMTTENVAGVHLPTFGLSKGGQQINKSRESHIKAVEALIALASLQTA - -MK EAAFSLAEAKFTSG-DI NQVV LQNVTKAQIKIRTKKDNVAGVTLPVFGLARGGQQLAK LKKNYQSAVKLLVELASLQTS - -MK EAAFSLAEAKFTSG-DI NQVV LQNVTKAQIKIRTKKDNVAGVTLPVFGLARGGQQLAKLKKNYQSAVKLLVELASLQTS - - I R DAFFR LTEAEFLGA -N LKMF LY E-CQK QNVYVR SRV EQV SGVSLPFFFLDR SGQSLNECREKFLEVLEMLVDLCALKNS - - MKA SSFSLVSAKYTAG -EFSHVVVQNVKNSTYKVKLTQENI AGVRLPVFGLSKGGQSVANARQQY LKALDSLVK LASLQTA - -MR EAAFSLA EAK FTAG-DFSTTVI QNVNKAQVKIRAKKDNVAGVTLPVFGLARGGEQLAKLKRNYAKAVELLVELASLQTS - - MR EAAF SLA EAK FTAG -DFSTTVI QNVNKAQVKIRAKKDNVAGVTLPVFGLARGGEQLAKLKRNYAKAVELLVELASLQTS - - MQTAAFSLAEVTYATGENI GYQVQENVANARFKVRATQENVSGVY LPQFGLGRGGQQVQRAKEIYSRAVETLVELASLQTA
$--I K G S Y F T I T Q A Q F I A G-D I S L A V Q E S L K I P T Y R M E L Q V E N I A G V Q V P S F G L G K G G E Q I K E A Y S A F R H T L S L L V K I A S L Q T S$ - -MR EAAFS LA EAKFTAG-DFSTTVIQNVNKAQVKIRAKKDNVAGVTLPVFGLARGGEQLAKLKRNYAKAVELLVELASLQTS - -MQI AAFSLAEVTYAVGGDIGYTVQESAK SARFRIRAKQENVSGVLLPAFGLGKGGQQVQRCRETYARAVEALVELASLQTA - -Y EK STEKINLASAVNG-MVAVKSTAFTAK EYPEI QLSGHNI MGVVVPKIGIIGTNSYIDETADAYEELVEKIIAAAELETT --MR EAAFSLAEAKFTAG-DFSATVI QNVNKAQVKIRTKKDNVAGVTLPVFGLARGGEQVTKLKKNYGKAVELLVELASLQTS - - MR EAAFSLAEAKFTAG-DFSTTVI QNVNKAQVKIRAKKDNVAGVTLPVFGLARGGEQLAKLKRNYAKAVELLVELASLQTS - - MQI ASLSLAEVTYAVGGNIGYQI QESAK SARFRIRAKQENVSGVLLPAFGLGKGGQQVQRCRETYARAVEALVELASLQTA - -MRASSFSLAEAKYVAGDGVRHVVLQSVR SASLRVR SHQENVAGVKLPKFGLARGGQQVAACRAAHVKAI EVLVELASLQTS - -MR DAHFAWTRAKYAGGDAVKHAVLDGVDRANVRVMAHEDNVAGVKI PKFGLARGGARVREAKASYGEAIGLLSELASLQTA
- -MR EAAFSLAEAK FTAG-DFSTTVI QNVNKAQVKIRAKKDNVAGVT LPVFGLARGGEQLAK LKRNYAKAVELLVELASLQTS - - AQEALLLIAKAQYAAG-EFHQNVKDAVKRATIRLEI SSENIAGVMLPEVGLARGGQSI QRCRDKFKDLLMLLVKIASYQTS - -MQTAAFSLAEVQYATGDNIAYQVQESVQKAR FQVKAKQENVSGVY LPTFGLGRGGQQVQKAKLVYTRAVETLVELASLQTA
- -MRNASFSLAK SVWAAG-DFKGQI I EGI KR PVVT LSLSTNNVAGVKLPI FGVAAGGQVI NNTR ENY LQC LNMLVK LA SMQVA - -MGNASFSLAKAVWAAG-DFKGQI I EGIKRPVVT LSLSTNNVAGVKLPI FGIASGGQVINNTRENYLQCLNMLVKLASMQVA - - MRNASFA LAK SVWAAG -DFKGQII EGI KRPVVT LSLSTNNVAGVKLPIFGVAAGGQVINNTR ENY LQCLNMLVK LASMQVA - -MKASSFALTEAKYVAGENIKHTVLENVQTAT LKVRSRQENVAGVKLPKFGLARGGQQVQACRAAYVKAI EVLVELASLQTS --MR EAAFSLAEAKFTAG-DFSTTVI QNVNKAQVKIRAKKDNVAGVTLPVFGLARGGEQLAKLKRNYAKAVELLVELASLQTS - -MQTAAFSLAEVSYATGENI GYQVQESVSTAR FKVRARQENVSGVYLSQFGLGRGGQQVQRAKEIYSRAVETLVELASLQTA --MQI AAF SMA EVG FAMGNNI NFEI QQSVKQPR LRVR SKQENI SGVFLPTFGLGKGGQQI QKARQVYEKAVET LVQLASYQSA - -MK LASLSLA EAK FAMG -DI SHNVLQNVTKAQTKVR SKK ENVAGVNLPVFGLSRGGQQIDR LKKNYAKAI ELLVELASLQTS QFMR EAAFSLAEAKFTAG-DFSITVI QNVNKAQVKVRAKKDNVAGVTLPVFGLAKGGEQI SRLKRNYARAVELLVELASLQTS - -MQKAFI QLADAYWAAD-QFNTNVRESVKKALVRI EYSSENIAGVMLPNLGLDKGGFSI QKAKERFKEALYLLVKVASLQTS - - LKDATY S LANAVWSAE -DFK SLVI ESVGRPSVT LKLRGENIAGVLLPVFSLSSGGSAI QSVKTTHLAALDI LVELASLQIS - - LKDATY SLANAVWSAE-DFKSLVI ESVGRPSVT LKLRGENIAGVLLPVFSLSSGGSAIQSVKTTHLAALDI LVELASLQIS - - AKDA LFAYTEVKFVAS-DISPTVIQSVGNMPQLLLMTI DNIAGVRTPQFGLARGGQQI QKAREEFTKFLDSLVR LAELQTA $--I K G S Y F T I T Q A Q F I A G-D I S L A V Q E S L K L P T Y T L T L R V D N V A G V R V P A F G I G R G G E Q L R E A R D A F R E T L K L F V K I A S L Q V S$ - - I RGAYFTVSKAQFIAG-DIGLAVQESLKLPTYAMR LRVENI AGVRVPSFGIGRGGEQLREASEKFRETLRLLVKIASLQVS - -MQQA SF S LA EVQYATG-DIGYIVQESVK SASFRVRAKQENVSGVI LPAFGLSRGGQQVSKAR EVYTQALKVLVELASLQTA - - MQTAAFSLAEVTYATGDNINYQVQESVR SAR LRVRAKEENVSGVKLPSFGLGRGGQQVQKAKAVYSKAVETLVELASLQTA

Anopheles_gambiae/1-3220
Arabidopsis_thaliana/1-3286
Ashbya_gossypii/1-3281
A fumigatus/1-3241
Aspergillus_niger/1-3300
Bombyx_mori/1-2389
Bos_taurus/1-3273
C_briggsae/1-3231
Caenorhabditis_elegans/1-3289
Candida_albicans/1-3266
Candida_glabrata/1-3282
Canis_familiaris/1-3286
Ciona intestinalis/1-2712
Cryptococcus_neoformans/1-3301 Cryptosporidium_hominis/1-3020 C_parvum/1-3281
Danio_rerio/1-3286
Debaryomyces hansenii/1-3166
Dictyostelium_discoideum /1-3275 D_melanogaster/1-3287
Drosophila_pseudoobscura/1-3286
Encephalitozoon cuniculi/1-3199
Entamoeba_histolytica/1-3271
Gallus_gallus/1-3286
Homo_sapiens/1-3286
Kluyveromyces_lactis/1-3281
Leishmania_major/1-3286
Macaca_mulatta/1-3182
Magnaporthe_grisea/1-3286
Methanosarcina_acetivorans/1-3043
Monodelphis_domestica/1-3196
Mus_musculus/1-3286
Neurospora_crassa/1-3264
Oryza_sativa/1-3218
Ostreococcus_lucimarinus/1-3287
Pan_troglodyte/1-3286
Paramecium tetraurelia/1-3282
Pichia_stipitis/1-3290
_falciparum/1-3285
P_knowlesi/1-3284
Plasmodium_yoelii/1-3285
Populus_trichocarpa/1-3286
Rattus_norvegicus/1-3286
Saccharomyces_cerevisiae/1-3284
Schizosaccharomyces_pombe/1-3284
Strongylocentrotus_purpuratus/1-3252
Takifugu_rubripes/1-2930
Tetrahymena_thermophila/1-3270
T_annulata/1-3248
Theileria parva/1-3273
Trichomonas_vaginalis/1-3198
T_brucei/1-3264
rypanosoma_cruzi/1-3266
Ustilago_maydis/1-3291
Yarrowia_lipolytica/1-3278

FVTLDEVIKITNRRVNAI EH----VI IPRIDRTLAYI I SELDELEREEFYRLKKI QDKK FLTLDEAIKTTNRRVNALEN-----VVKPKLENTI SYIKGELDELEREDFFRLKKI QGYK FVI LDEVIKVTNRRVNAI EH-----VII PRTENTIAYINSELDELDREEFYRLKKVQEKK
 FVI LDEVIKVVNRRVNAI EH-----VII PRTENTIKYINSELDELDREEFYRLKKVSGKK FVTLDEVIKITNRRVNAI EH-----VII PRLERTLAYI I SELDELEREEFYRLKKI QDKK FVTLDEAIKITNRRVNAI EHGEFKLPFCPR LHPCLR PARTQA -
FITLDEAIKVTNRRVNAI EH-----VII PRI ENTLTYIVTELDEMEREEFFRMKKI QANK FITLDEAIKVTNRRVNAI EH-----VII PRI ENTLTYIVTELDEMEREEFFRMKKI QANK FI I LDEVIKITNRRVNAI EH-----VI I PRTENTIAYINGELDEMDREEFYRLKKVQEKK FII LDEVIKVTNRRVNAI EH-----VII PRTENTIAYINSELDELDREEFYRLKKVQEKK FVTLDEAIKITNRRVNAI EH-----VII PRI ERTLAYI ITELDEREREEFYRLKKIQEKK FVTLDESIKITNRRVNAI EH-----VII PKI ERTI SYIITELDEGEREEFFRLKKI QQKK FTI LDEVIRATNRRVNAI EH-----VVI PRLENTIKYINSELDEMDREEFFRLKKVQGKK FFSLDEEIKMTNRRVNALQN-----VVLPKLEDGMNYI LRELDEI EREEFFRLKKIQEKK FFSLDEEIKMTNRRVNALQN-----VVLPKLEDGMNYI LRELDEI EREEFFRLKKIQEKK FVTLDEAIKITNRRVNAI EH-----VII PRI ERTLTYIITELDEREREEFYRLKKIQEKK FII LDEVIKVTNRRVNAI EH-----VII PRTENTISYINSELDELDREEFYRLKKVQEKK FITLDEVIKITNRRVNAI EY-----VVKPKLENTISYI ITELDESEREEFYRLKKVQGKK FVTLDEVIKITNRRVNAI EH-----VII PRIDRTLAYI I SELDELEREEFYRLKKI QDKK FVTLDEVIKITNRRVNAI EH-----VII PRIDRTLAYII SELDELEREEFYRLKKI QDKK FRVLNSI LMSTNRRVNALEF-----NI I PRLENTVSYIVSELDEQDRGDFFRLKKVQNLK FLTLDTVIKITNRRVNALEH-----VVI PMTQATVKYI ETELDESEREEFFRLKLI QNKK FVTLDEAIKITNRRVNAI EH-----VII PRI ERTLSYIITELDEREREEFYRLKKIQEKK FVTLDEAIKITNRRVNAI EH-----VII PRI ERTLAYIITELDEREREEFYRLKKIQEKK FII LDEVIKVTNRRVNAI EH-----VII PRTENTIAYINSELDELDREEFYRLKKVQEKK WI TLDIAQKVTSRRVNALEK-----VVI PRVQNTLSYITSELDEQEREEFFRLKMVQKKK FVTLDEAIKITNRRVNAI EH-----VII PRI ERTLAYIITELDEREREEFYRLKKIQEKK FVI LDEVIKVVNRRVNAI EH-----VII PRTENTIKYINSELDELDR EEFYRLKKVAGKK MKRLLDEI EKTKRRVNALEF-----KVIPELIATMKYIRFMLEEMER ENTFRLKRVKARM FITLDEAIKITNRRVNAI EH-----VII PRI ERTLNYIVTELDEREREEFYRLKKI QEKK FVTLDEAIKITNRRVNAI EH-----VII PRI ERTLAYIITELDEREREEFYRLKKIQEKK FVI LDEVI KVVNRRVNAI EH-----VII PRTENTIKYINSELDELDREEFYR LKKVAAKK FLTLDEAIKTTNRRVNALEN----VVKPRLENTISYIKGELDELEREDFFRLKKI QGYK FVTLDEAIKTTNRRVNALEN-----YVTPRLQNTVKYI LSELDELEREEFFRLKKVQAKK FVTLDEAIKITNRRVNAI EH-----VII PRI ERTLAYIITELDEREREEFYRLKKIQEKK FVSLDQVI KVTNRRVNALEY-----VVI PRFTATMNYI DMELDEMSKEDFFR LKKV LDNK FI I LDEVI KVTNRRVNAI EH-----VII PRTENTI SYI NSELDELDREEFYRLKKVQEKK FFSLDEEIKMTNRRVNALNN-----IVLPRLDGGINYIIKELDEI EREEFYRLKKIKEKK FFSLDEEIKMTNRRVNALNN-----IVLPRLDGGINYIIKELDEI EREEFYRLKKIKEKK FFSLDEEIKMTNRRVNALNN-----IVLPRLEGGINYIIKELDEI EREEFYRLKKIKEKK FMT LDTAIKTTNRRVNALEN----VVKPRLENTITYIKGELDELEREDFFRLKKI QGFK FVTLDEAIKITNRRVNAI EH-----VII PRI ERTLAYIITELDEREREEFYRLKKIQEKK FII LDEVIKVTNRRVNAI EH-----VII PRTENTIAYINSELDELDREEFYRLKKVQEKK FVLLGDVLQMTNRRVNSI EH-----III PRLENTIKYI ESELEELEREDFTRLKKVQKTK FITLDEVIKITNRRVNAI EH-----VII PRI ENTISYITTELDEREREEFYRLKKIQEKK FVTLDEAIKITNRRVNAI EH-----VII PRIDRTLTYIVTELDEREREEFYRLKKI QEKK FITLDEVIKVTNRRVNALEH-----VVI PRFMEVQAY I NQELDEMSREDFFR LKKV LDFK FII LNEEIRMTNRRINALDN-----VLIPSIDRNLEYIRRELDEMEREEFYRLKMIKKHK FII LNEEIRMTNRRINALDN-----VLI PSI DRNLEYIRRELDEMEREEFYRLKMIKKHK FNVI DDVLRITNRRVNAMEC-----VLI PKYQAAI AFVDST LDENEREEFFRLKKVQET I WMT LDVAQKVTSRRVNALEK -----VVI PRMENT LNYI SSELDEQEREEFFRLKMI QKKK WVT LDLAQKVTNRRVNALEK-----VVI PRVQNTLSYITSELDEQEREEFFRLKMVQKKK FVI LDEVIRMTNRRVNAI EH-----VII PRLENTI SYIVSELDEADREEFFRLKKVQAKK FVI LDEVIKITNRRVNAI EH

