QUANTITATIVE AND EVOLUTIONARY GLOBAL ANALYSIS OF ENZYME REACTION MECHANISMS

Neetika Nath

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

2015

University of
St Andrews

600
YEARS

PhD Thesis

Quantitative and Evolutionary Global Analysis of Enzyme Reaction Mechanisms

Author:
Neetika Nath

Supervisor:
Dr. John BO Mitchell

This thesis presented for the degree of
Doctor of Philosophy

School of Chemistry
University of St Andrews
United Kingdom
Friday 17th March, 2015

Declaration

I, Neetika Nath, declare that this thesis titled, '*Quantitative and Evolutionary Global Analysis of Enzyme Reaction Mechanisms*' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. Except such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: ______________________________

Date: ______________________________

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.

Signed: ______________________________________

Date: ______________________________________

Supporting Statement

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 publication of this thesis:

PRINTED COPY

a No embargo on print copy

b Embargo on all or part of print copy for a period of ... years (maximum five) on the following ground(s):

  - Publication would be commercially damaging to the researcher, or to the supervisor, or the University
  - Publication would preclude future publication
  - Publication would be in breach of laws or ethics

c Permanent or longer term embargo on all or part of print copy for a period of 1 years (the request will be referred to the Pro-Provost and permission will be granted only in exceptional circumstances).

ELECTRONIC COPY

a No embargo on electronic copy

b Embargo on all or part of electronic copy for a period of ... years (maximum five) on the following ground(s):

  - Publication would be commercially damaging to the researcher, or to the supervisor, or the University
  - Publication would preclude future publication
  - Publication would be in breach of law or ethics

c Permanent or longer term embargo on all or part of electronic copy for a period of 1 years (the request will be referred to the Pro-Provost and permission will be granted only in exceptional circumstances).

Supporting statement for electronic embargo request:

Signature of Candidate: ____________________

Signature of Supervisor : ____________________

Date: ____________________

“Statistics is the grammar of science.”

Karl Pearson

Contents

Acknowledgements

First and foremost, I am deeply indebted to my promoter and supervisor Dr. John BO Mitchell, for his guidance and contribution to this thesis. His guidance, advice and flexibility, stimulating suggestions and encouragement helped me during all ups and downs I faced in my research.

I wish to express my deep sense of gratitude to Prof. Gustavo Caetano-Anollés, Department of Crop Science, University of Illinois at Urbana - Champaign, USA, for having fruitful discussions, sharing knowledge on enzyme function evolution, granting access to his lab for a scientific visit and in research collaboration with some parts of this thesis.

It is my pleasure to thank many collaborators and supporters from many parts of the world: Prof. Gustavo Caetano-Anollés (University of Illinois, United States), Minglei Wang (University of Illinois, United States), Syed Abbas Bukhari (University of Illinois, United States), Dr. Tanja van Mourik (University of St Andrews, United Kingdom), James L. McDonagh (University of St Andrews, United Kingdom), Dr. Lazaros Mavridis (Queen Mary, University of London, United Kingdom), Dr. Luna De Ferrari (Computational Systems Biology group at CISA, United Kingdom) and PD Dr. Reinhard Guthke (Hans Knöll Institute (HKI), Germany).

I would like to thank my colleagues Lazaros Mavridis, Luna De Ferrari, Rosanna Alderson, and James McDonagh for providing valuable information to build a strong background in enzyme function evolution and machine learning model. Additionally, I would like to thank other colleagues including Dr. Ludovic Castro, Luke Crawford, Leo Holroyd, Rachael Skyner, Ava Sih-Yu Chen, and Jose Garrido Torres for assisting in many ways during my research work.

I am very grateful for access to the EaStCHEM Research Computing Facility and to Dr Herbert Früchtl for its maintenance.

I am very grateful to the Scottish Universities Life Sciences Alliance (SULSA) and the Scottish Overseas Research Student Awards Scheme of the Scottish Funding Council (SFC) for providing enough financial grants to pursue this PhD.

I am happy to have wonderful friends from different parts of the world and would like to thank and a BIG hug to many including Himan (my partner), Piyush, Geetanshu, Iram, Ruchi, Rewis, Jessy, Rohit, Swati urf ChiChi, Iron Kirti, Satish, James, Laz, Rosie, Leo, Luna, Fabian, Nina, Judith, Priyanka, Kaveri, EHG (extremely hilarious group) - Shyam, Nouchali, Haifa, Ildiko, Vinodh, Shantanu, Sayantan, for supporting directly or indirectly during my PhD work. I would also like to thank to my Unsocial friends members of Santos' group - Himan, Jitu, Divu, Susu, Sonika, Neenu, Manu, Varu.

I received particular assistance in the writing of this work with respect of matters of grammar, style, vocabulary, spelling or punctuation. The assistance was provided by: Leo, Rosie, John and Naomi.

I would like to give my special thanks to my lovely parents and my brother whose love and blessings enabled me to successfully complete this work.

I offer my regards and benedictions to all of those who supported me in any respect during the completion of my thesis.

It takes a village to raise a PhD student.

Publications

- N Nath & JBO Mitchell, *Is EC class predictable from reaction mechanism?* **BMC Bioinformatics**, 13:60 (2012)
- RG Alderson, L De Ferrari, L Mavridis, JL McDonagh, JBO Mitchell & N Nath, *Enzyme Informatics*, **Current Topics in Medicinal Chemistry**, 12, 1911-1923 (2012)
- L Mavridis, N Nath & JBO Mitchell, *PFClust: a novel parameter free clustering algorithm*, **BMC Bioinformatics**, 14:213 (2013)
- JL McDonagh, N Nath, L De Ferrari, T van Mourik & JBO Mitchell, *Uniting Cheminformatics and Chemical Theory to Predict the Intrinsic Aqueous Solubility of Crystalline Druglike Molecules*, **Journal of Chemical Information and Modeling**, 54, 844-856 (2014)
- N Nath, JBO Mitchell & G Caetano-Anollés, *The Natural History of Biocatalytic Mechanisms*, **PLoS Computational Biology**, 10, e1003642 (2014)

Abstract

The most widely used classification system describing enzyme-catalysed reactions is the Enzyme Commission (EC) number. Understanding enzyme function is important for both fundamental scientific and pharmaceutical reasons. The EC classification is essentially unrelated to the reaction mechanism.

In this work we address two important questions related to enzyme function diversity. First, to investigate the relationship between the reaction mechanisms as described in the MACiE (Mechanism, Annotation, and Classification in Enzymes) database and the main top-level class of the EC classification. Second, how well these enzymes biocatalysis are adapted in nature.

In this thesis, we have retrieved 335 enzyme reactions from the MACiE database. We consider two ways of encoding the reaction mechanism in descriptors, and three approaches that encode only the overall chemical reaction.

To proceed through my work, we first develop a basic model to cluster the enzymatic reactions. Global study of enzyme reaction mechanism may provide important insights for better understanding of the diversity of chemical reactions of enzymes. Clustering analysis in such research is very common practice. Clustering algorithms suffer from various issues, such as requiring determination of the input parameters and stopping criteria, and very often a need to specify the number of clusters in advance.

Using several well known metrics, we tried to optimize the clustering outputs for each of the algorithms, with equivocal results that suggested the existence of between two and over a hundred clusters. This motivated us to design and implement our algorithm, PFClust (Parameter-Free Clustering), where no prior information is required to determine the number of clusters.

The analysis highlights the structure of the enzyme overall and mechanistic reaction. This suggests that mechanistic similarity can influence approaches for function prediction and automatic annotation of newly discovered protein and gene sequences.

We then develop and evaluate the method for enzyme function prediction using machine learning methods. Our results suggest that pairs of similar enzyme reactions tend to proceed by different mechanisms. The machine learning method needs only chemoinformatics descriptors as an input and is applicable for regression analysis.

The last phase of this work is to test the evolution of chemical mechanisms mapped onto ancestral enzymes. This domain occurrence and abundance in modern proteins has showed that the α/β architecture is probably the oldest fold design. These observations have important implications for the origins of biochemistry and for exploring structure-function relationships.

Over half of the known mechanisms are introduced before architectural diversification (nd <0.39) over evolutionary time. The other half of the mechanisms are invented gradually over the evolutionary timeline just after organismal diversification ($0.67 > nd > 1$). Moreover, many common mechanisms including fundamental building blocks of enzyme chemistry were found to be associated with the ancestral fold.

List of Figures

List of Tables

Lists of Abbreviations and Acronyms

ATP	Adenosine triphosphate
BLAST	Basic Local Alignment Search Tool
BRENDA	BRaunschweig ENzyme DAtabase
CATH	Class, Architecture, Topology and Homologous superfamily
CBC	Composite Bond Change descriptor
CSA	Catalytic Site Atlas
DAVID	Database for Annotation, Visualization and Integrated Discovery
DBSCAN	Density Based Spatial Clustering of Applications with Noise
EC	Enzyme Commission
EzCatDB	Enzyme Catalytic-mechanism Database
GO	Gene Ontology
HD	Human Designed descriptor
KEGG	Kyoto Encyclopedia of Genes and Genomes
kNN	K Nearest Neighbour
MACiE	Mechanism, Annotation and Classification in Enzymes
MOLMAP	MOLecular Mapping of Atom-level Properties
MS	Mechanistic Similarity descriptor

NADP	Nicotinamide Adenine Dinucleotide Phosphate
nd	Node Distance
OBC	Overall Bond Change descriptor
OS	Overall Similarity descriptor
PFClust	Parameter Free CLUSTering
PLS	Partial Least Squares
Poly	Polynomial kernel function
PSI-BLAST	Position-Specific Iterating BLAST
RBF	Radial Basis Function
RF	Random Forest
RMSE	Root Mean Square Error
SCOP	Structural Classification of Proteins
SVM	Support Vector Machine

1

Introduction

1.1 Motivation

With the growing amount of available genomic information, it is a challenge to make the process faster in executing the time required for the annotation of newly discovered proteins. Many informatics groups have tried to use protein sequence and structural information in order to understand and reproduce the classification system of enzymes with some success. The Enzyme Commission (EC) number system is designed to represent functional classification based on the overall chemical transformation in enzyme catalysis. The EC classification does not explore the detailed chemical mechanism of the enzyme reaction. Thus, it is essential to integrate such information in order to annotate function to newly discovered proteins.

The most essential property of enzymes is their ability to catalyze very specific chemical reactions. To make the functional annotation of proteins faster, it is essential to represent the chemical reaction of enzymes quantitatively. Numerous authors have attempted to represent enzymatic activity quantitatively based on quite different information such as substrate, reaction, cofactors. By definition, EC function describes the overall reaction of an enzyme. Thus, a full description of the overall chemical transformation from starting materials to products should, in principle, lead to the perfectly accurate assignment of the EC number.

Our motivation in this work is to understand enzyme function diversity.

For this we looked into the descriptors that encode the overall reaction as well as mechanism-rich information. Our intention with such information is to investigate the relationship between the reaction mechanism and EC classification using unsupervised and supervised classification methods.

Unsupervised global analysis of enzyme reactions is helpful to get better insight into the understanding of the diversity of chemical reactions of enzymes. We proposed a novel clustering algorithm, PFClust (Parameter Free Clustering), which is suitable for finding the structure of data when no prior information is available.

Moreover, supervised learning methods have also proven their worth in understanding the structure of the given data. Machine learning methods are suitable for molecular biology data to construct classifiers that can explain complicated relationships in the data. Machine learning algorithms, such as the support vector machine, can learn the patterns in the data with respect to given class and use that information to predict the function of newly discovered proteins.

Another important challenge in molecular biology is to understand how enzymes adapt their chemical mechanisms under evolutionary pressure. Such global study of enzyme reaction mechanisms may provide important insights for better understanding of the diversity of chemical reactions of enzymes.

Thesis Questions In this thesis, we address two important questions related to enzyme function diversity. First, to investigate the relationship between the reaction mechanisms as described in the MACiE (Mechanism, Annotation, and Classification in Enzymes) database and the main top-level class of the EC classification. Second, how well these enzymes', biocatalytic processes are adapted in nature.

To address these questions, we retrieve data from MACiE which is then used to quantitatively encode overall and reaction mechanisms as chemoinformatics descriptors. These chemoinformatics descriptors are used to investigate the relationship with the main top-level class of the EC classification. Enzyme function prediction is an important question in post-genomic bioinformatics to understand enzyme function diversity.

Traditionally, two strategies are used in order to address this question: first, transferring function annotation to similarly annotated proteins through homology, and second, machine learning algorithms to treat this as a classification problem against a fixed taxonomy, such as EC hierarchy.

This work examines the potential of a workflow designed using machine learning to automate the function prediction. In addition, this approach examines the potential to predict the solubility of drug-like molecules as a regression problem. It is noteworthy that the performances of the learning algorithms are highly dependent on the nature of the training data provided.

Moreover, we investigate the diversity of enzyme reactions using unsupervised clustering algorithms. By identifying critical features of enzyme functions, further use of such information could improve enzyme reaction classification, computational annotation, and function prediction for newly discovered proteins.

Furthermore, a challenge in molecular biology is to explore the chemical mechanisms used in biochemical reactions catalysed by ancestral enzymes. Such investigation has important implications for exploring structure-function relationships. The main challenge in molecular biology is to understand how new enzyme activities evolve in nature.

1.2 Contributions

The primary contribution of this thesis is the development of a sophisticated biostatistical method to examine the diversity of enzyme mechanisms and function prediction. Our work makes important contributions to the field of understanding function diversity, which leads to better function prediction. Another contribution of this thesis is a machine learning pipeline that can handle a number of problems including classification and regression. Also, it unifies inbuilt validation and evaluation of the machine learning algorithms to optimise the parameter and output.

We realise that there are missing pieces of the puzzle that need to be put together in order to meet the challenges of enzyme annotation, such as to blend various biological sources and to incorporate computational power to automate function prediction. In more detail, the contributions of this thesis are listed below:

- One of the main contributions of this thesis is PFClust (Parameter Free Clustering), a clustering algorithm that is suitable for use when no prior information is available.

- A quantitative analysis of enzyme function and exploring the biological attributes in each cluster.
- Another contribution is a workflow designed using various machine learning algorithms, with inbuilt internal validation and parameter optimisation.
- Also, we looked into function prediction of enzyme reaction mechanisms and evaluation of the predictions.
- We investigated how enzyme activities adapt in nature.

A summary of the claims and contributions is shown in Table 1.1, where the columns represent relevant reference, contribution in the paper and relevant chapters in this thesis.

Table 1.1: The following table lists my contributions in this work. Details are explained in respective chapters.

Reference	Contribution	chapters in thesis
PFClust [1]	My main contribution in this work is to design the experimental and validation studies and carry out the comparison of PFClust with the other methods.	Chapter 4
Enzyme function prediction [2]	I designed the workflow, carried out the majority of the computations and performed the statistical analysis.	Chapter 5
Solubility prediction of drug-like molecules [3]	I designed the workflow for this work and also produced machine learning R script in collaboration with Dr. Luna De Ferrari	Chapter 5
Natural history of enzyme biocatalysis [4]	I designed and executed the whole experiment for this work.	Chapter 6

1.3 Thesis Structure

Chapter 2: Enzymes, Function, and Bioinformatics In this chapter, we review the exquisite catalytic properties of enzymes and the active site residues that comprise their catalytic tool kits. Furthermore, we elaborate various definitions of functions and their usages for addressing the problem of protein function classification. In addition, we discuss the potential applications of this work in enzyme engineering and clinical implication.

Chapter 3: Data and Databases This chapter describes the descriptors and databases, from which the data is retrieved for use in this thesis. According to the definition of bioinformatics, well-defined quantitative representation of biological data is important for the interpretation. This chapter describes the limitations and challenges related to our work in this thesis.

Chapter 4: Quantitative Global Analysis of Enzyme Reaction Mechanisms In this chapter, we describe the unsupervised clustering analysis of enzyme reaction mechanisms. The motivation here is to understand the 'important' biological factors associated with enzyme reaction clusters. Based on the results from various clustering algorithms, we discuss the implications of a novel clustering algorithm: PFClust. Also, we discuss the issues suffered by various clustering algorithms such as optimising the input parameter and stopping criteria, and specifying the number of clusters in advance.

Chapter 5: Prediction of Enzymatic Function Here, we describe the bioinformatics workflow using machine learning methods that are applied for predicting enzymatic function with good accuracy. The motivation here is to test the enzyme mechanistic descriptors' performance for function annotation. Furthermore, this workflow is applied to a regression problem.

Chapter 6: Enzyme Function Evolution *Chemolution Study* Here we demonstrate the mapping of the enzymatic function onto enzyme domains. The motivation here is to test if the oldest folds have higher numbers of mechanistic step types compared to younger fold structures. We discuss the complete data culling process and its representation.

Chapter 7: Conclusion and Discussion This chapter outlines the conclusion of the thesis work and details its implications to other research areas, and beyond, while also identifying new research questions that could be addressed by future research.

2

Enzymes, Function, and Bioinformatics

Biochemical reactions, nearly all of which are mediated by a series of remarkable catalysts known as enzymes, shape all living systems. Enzymes exhibit remarkable selectivity and specificity for selecting a molecule and producing a single product. The list of functions done by enzymes ranges from alcoholic fermentation to ribozymes, including ribosomal RNA, which catalyze the formation of peptide bonds between amino acids. Enzymes are subject to the same laws of nature that govern the behavior of other substances, but enzymes differer from ordinary chemical catalysts in several important respects:

- High reaction rate: the rates of enzymatically catalyzed reactions are typically 10^6 to $10^1$2 times greater than those of the corresponding uncatalyzed reactions and are at least several order of magnitude greater than those of the corresponding chemically catalyzed reactions.

- Milder reaction conditions: enzymatically catalyzed reactions occur under relatively mild conditions: temperature below 100'C atmospheric pressure and nearly neutral pH. In contrast, efficient chemical catalysis often requires elevated temperatures and pressures as well as extremes of pH.

- Greater reaction specificity: enzymes have a vastly greater degree of specificity with respect to the identities of both their substrates and their products than do chemical catalysts: that is, enzymatic reactions rarely have side products.

- Capacity for regulation: the catalytic activities of many enzymes vary in response to the concentrations of substances other than the substrates. The mechanisms of these regulatory processes include allosteric control, covalent modification of enzymes, and variation of the amounts of enzymes synthesized.

However, listed properties of the enzymes depend on the external environment of the reaction. Temperature and pH both affect the rates at which a reaction takes place, and there exists an optimum value for each.

To eliminate the confusion in rationally naming the rapidly growing number of newly discovered enzymes, a scheme for the systematic functional classification and nomenclature of enzymes was adopted by the International Union of Biochemistry and Molecular Biology (IUBMB, we discuss this in detail in Chapter 3) [5]. Enzymes are classified and named according to the nature of the overall chemical reactions they catalyze. Enzymes are proteins made up of various combinations of 20 different amino acids, that catalyse biochemical reactions. In humans, enzymes represent drug targets [6] for clinical diagnostics due to their exquisite properties. Enzymes are a proficient and robust apparatus for executing functions within the body. Also, enzymes show remarkable evolutionary adaptation in nature.

Here, we discuss enzyme function and the features that are profoundly used in various function annotation schemes. The enzyme function prediction starts from sequence. Here, our work is not on technicality of sequences or structures of enzymes. Rather, our work depends on the definition of enzyme biocatalysis translated into chemoinformatics descriptors for function prediction and to understand the diversity of function. In addition, we have developed robust computational approaches to address such biological questions.

To describe the function of a protein is not an easy task as the function can be described at all levels of the enzyme function classification hierarchy, such as EC number. There are also examples of ‘moonlighting’ proteins, which play many roles in the cell (some nonenzymatic) by acquiring minimal changes either in sequence or structure. Thus, this suggests many ways one can define the classification and annotation of enzyme function by using information at different levels of data present, starting from sequence / structure or components of catalytic sites [7]. Our focus here is on the quantitative representation of function definition at mechanistic and overall levels of enzyme reaction to study the diversity of function.

Overall, the idea of such studies is to understand how well an enzyme adapts its function. Indeed, evolutionary relationships exist between proteins that show diverse folds or topologies, but share similar function. It is unlikely that these folds evolved independently. Many elegant studies [8, 9] suggest that enzymes evolved from pre-existing enzymes via gene duplication using common binding sites or mechanistic features to catalyze different reactions. The most probable scenario is that these folds evolved from a smaller, less diverse set of ancestral proteins. The earliest enzymes were probably weakly catalytic and multifunctional with broad specificities [10–12]. Gradually, evolutionary events (gene duplication, mutation and divergence) helped the evolution of more numerous, effective, and specific enzymes to evolve from the multifunctional enzymes [10]. Hence, they share a set of protein functions to effect the reaction [4, 13]. This in turn suggests that there are numerous functions which have evolved in unrelated structures.

Evolvability of promiscuous function provides immediate advantage to new enzymes to become positively selected [9]. Once the promiscuous function got selected, it can go through a series of reconstructions for improvement without abolishing the primary, native function of the enzyme [9]. For example enzymes belonging to *alpha/beta* hydrolase fold share conserved mechanistic features that are evolved to catalyze different reactions [13].

In this chapter, we review some aspects of enzyme function as well as traditional and modern ways of annotating newly discovered proteins and their limitations. We will also discuss evolutionary aspects of such studies, their application for clinical purposes and their limitations.

Definition 1.

Homologous - Proteins which have evolved from a common ancestral, and whose evolutionary relationship is evident from similarities in sequence, structure and/or function.

Analogous - Proteins where no evidence of a common ancestry is found, yet which are similar in some properties such as sharing the same fold. Analogous enzymes perform the same function.

Divergent - Proteins derived from a common ancestor to form different functionalities.

Convergent - Proteins without any trace of evolutionary relationship but which have evolved to possess a similar function.

2.1 Enzyme Function or Catalytic Tool Kit

The reaction centre of the enzyme structure is built of amino acid residues. The amino acid residues that form the binding site are arranged to specifically attract the substrate. Enzymes are very specific both in binding substrate and in catalysing their reaction. Molecules that differ in shape or functional group distribution from the substrate cannot productively bind to the enzymes.

For the proper functioning of the enzyme, there are many components that can be represented for function prediction models such as active sites, cofactors. Cactors often play a vital part for the catalytic entities. Functional components such as overall reactions, catalytic residue, cofactors or mechanistic step types can serve as chemoinformatics descriptors from enzyme reaction centre. These quantitative representations of an enzyme reaction centre are thus considered for further investigation. Such data, which is now available databases, can be used to investigate function annotation. Also, robust bioinformatics approaches can be designed which greatly accelerate the process of functional annotation of the genome sequence, as manual annotation of enzyme function would be a very labour-intensive and time-consuming process. For quick annotation of protein function, it is essential to represent data in a computationally accessible form, which may enhance accuracy as well as completeness of the protein function annotation [14].

In this section, we discuss the enzyme functional toolkits that are quantitatively represented and favoured for functional prediction and the contribution of such work towards our understanding of enzymatic function. In addition, we discuss the limitations of representing the data for examining the enzyme function diversity.

2.1.1 Enzyme Catalytic Residues

The enzyme catalytic residues are the central part of the enzyme structure where substrate molecules bind and undergo a chemical reaction. The catalytic residues are evolutionarily conserved, typically hydrophilic, and located inside the pockets of enzyme structure [15]. The catalytic residues are placed precisely in the pocket of the protein structure. Knowledge and improved understanding of the properties of enzyme active sites and their assorted catalytic mechanisms is vital for novel protein design and predicting protein function from structure [16].

Experimentally determining the catalytic residues requires extensive experimentation (mutagenesis experiments) followed by exhaustive testing of the enzyme's catalytic performance, including concentration assays. Thus, with the help of bioinformatics approaches one can boost the function annotation process.

Traditionally, catalytic residues are identified by multiple sequence alignment or structure template search with enzymes whose catalytic residues are already annotated [17]. The focus of multiple sequence alignment [18] is on the conserved sequence signature, which is evolutionary conserved patterns, as attributes for enzyme function prediction [16].

Using information on catalytic residues available in databases it was found that the most common catalytic residues in enzymes to effect the reaction are histidine, aspartate, glutamate, lysine, cysteine, arginine, serine, threonine, tyrosine and tryptophan [16, 19]. Most of the catalytic residues are recruited to contribute to the stability of the transition state. Generally these catalytic residues take part in general acid/base chemistry (proton acceptor and proton donor) or nucleophilic addition (nucleophile and nucleofuge) to complete the reaction [20]. Sometimes these common catalytic residues share the responsibilities by various means such as direct acid-base action or by increasing the effect of charge in the locality [16]. Some catalytic residues perform different functions as per the requirement.

A database that possesses the relevant information on the catalytic sites of enzymes is the Catalytic Site Atlas (CSA) [21]. Detailed information on catalytic residues of enzymes is very important to understand the structure-function relationship and for good annotation of novel proteins. Another such database is InterPro [22], a resource which provides sequence signatures for function analysis. Using sequence features, a study by Ferrari et al. [23], was able to successfully provide 99% accurate prediction of enzymatic activity. Sequence is less conserved than structure, making such methods vulnerable to false positive results. With available information on structure we need to add this information to make better predictions.

The main challenge in such studies is to reduce the time required for annotating the newly discovered protein by using fast, accurate methods. As sequence is less conserved than structure one needs to be careful of false positive results.

2.1.2 Enzyme Structure-Cofactor Relationships

A cofactor is a non-protein chemical compound that is required as an additional factor to be included in the enzymatic reactions. These cofactors help in catalysing chemical reactions either directly or indirectly. The cofactors are classified as either metal ions (such as Ca^{2+}, Mg^{2+}), also known as metalloproteins [24], or as small organic molecules (such as nicotinamide-adenine - dinucleotide phosphate (NADP)) [25].

A wide variety of enzymatic function is dependent on such cofactors: for example, photosystem oxygen-evolving complex (OEC) is a metallo-oxo cluster, containing Mg^{2+} and Ca^{2+} [26]. Another example is fructose - bisphosphate aldolase (EC: 4.1.2.13, MACiE: M0052), a zinc-dependent enzyme which catalyses the reversible aldol cleavage or condensation of fructose-1,6-bisphosphate into dihydroxyacetone - phosphate and glyceraldehyde 3-phosphate [27]. There are two distinct types of mechanistic reaction executed by this enzyme: class I (Schiff-base) and class II (metallo). The class II aldolases utilise zinc as an electrophile in the catalytic cycle; no obvious structural similarities between class I and II aldolases are found. Three forms of class I proteins are found in vertebrates which participate in glycolysis.

Some databases store information relevant on roles of cofactors as they participate in the chemical reaction, such as Metal-MACiE [24] and CoFactor [25, 28]. Metal-MACiE presents information on the metal cofactors involved in the catalytic mechanisms of the metalloenzymes present in MACiE, while the CoFactor database contains knowledge on each organic enzyme cofactor, integrated with the corresponding enzyme's sequence, structure and reaction.

This mainly contributes towards cofactor engineering, a subset of protein engineering [29]. This type of engineering is used to increase the efficacy of metabolic networks by optimising the production of metabolites. The most basic strategy is to change the concentration of the cofactors to either increase or decrease the efficiency of a metabolic network.

In addition, the understanding of how well cofactors affect the enzymes in nature is valuable information for evolutionary studies. The cofactors are evolutionarily conserved in nature. According to Ji et al. [30], cofactors played an important role in the early history of life, allowing primordial proteins to perform oxidative functions.

2.1.3 Catalytic Mechanisms

Enzymes achieve their enormous rate accelerations via the same catalytic mechanistic principles used by chemical catalysts, stabilising the transition state and thereby lowering the activation energy. Through evolution, some enzymes have simply become efficient in stabilising the transition state more and therefore providing a greater acceleration of the reaction.

Conventionally, the reactions are described with the use of the curly arrow convention to represent the electron rearrangements that occur in going from reactants to products. Such information is available mostly in papers and textbooks, but recent effort in collaborative project between the Thornton Group at the European Bioinformatics Institute and the Mitchell Group at the University of St Andrews [31] made this information easily available in computer-readable form in the MACiE database (this is discussed in detail in Chapter 3).

The EC classification system evidently shows its worth in cataloguing the overall reaction of enzymatic reactions. However, this system lacks the mechanistic information that is increasingly available now [31, 32]. In order to answer biological questions using computational methods, it is important to represent the enzyme reaction mechanism quantitatively. The information available in MACiE is very useful for representing enzyme reactions not only at the overall reaction but also at the mechanistic level of reactions. One way is by representing data in a fingerprint [20], another by using sophisticated bioinformatics or statistical methods to estimate the similarities between reactions [20, 33].

For investigating the enzyme reaction of hydrolase family EC 3.b.c.d Oliver Sacher [34] represented the data by combing active site and physiochemical effects on the chemical reaction. They showed that physiochemical property overall compares well with the EC system. The enzyme reactions can be represented by the elementary steps of the reaction such as bond breakage in addition to its physiochemical effects [35, 36].

Again, using MOLMAP descriptors defining the difference between the products and the reactants from the KEGG database [35], an investigation was carried out to automate an assignment of EC number using the sophisticated decision making algorithm, Random Forest.

Such information is essential for the annotation and association of the EC classification system to newly discovered proteins. However, one should consider another possibility where enzymes with very similar overall reactions (EC number) can have quite different mechanistic steps to effect the reaction [20]. For example, MACiE recorded six different mechanistic reactions for metallo-β-lactamase (M002 - Class A, M0015, M0016 & M0258 - Class B, M0210 - Class D & M0257 - Class C) where they possess hydrolase reaction (EC 3.5.2.6). Another well studied example is the mechanistically diverse enolase superfamily [32, 37].

2.2 Evolution of Enzyme Function

Enzymes are very adaptive by nature. Through the course of evolution, the scaffold of enzymes has improved their functional, and specifically catalytic, efficiency. The main driving force for these evolutionary advances in enzymes is the requirement of natural selection to improve molecular and cellular function, increasingly optimising both catalytic capability and regulation.

The ancestral proteins are thought to have had very broad specificity and performed multiple functions. With due course of time these enzymes evolved exquisite catalytic properties to effect the reaction [10]. Broadly speaking, during evolution some enzyme superfamilies exhibit divergent catalytic activity, whereas others possess common mechanistic features such as a cofactor, mechanistic step or strategy as a point to evolve into a new function. At some point, these features are shared during evolution to improve the quality of function.

The evolutionary strategies are concatenated to provide two scenarios which are broadly registered as using similar properties, either chemistry or substrate, to evolve enzyme function. The one where enzymes share similar catalytic residues but perform dissimilar catalysis is called the 'chemistry driven scenario'. Conversely, in the 'substrate driven scenario', different catalytic residues are recruited to yield the same required product [38,39]. For example, pairs of enzymes in tryptophan and histidine biosynthesis provide two examples of substrate-driven evolution. The increasing understanding of chemical mechanism and its role of active site features will continue to enrich our understanding of molecular evolution.

Such scenarios have suggested the possibility of proteins sharing common function with completely different structures. One striking example is that of the Ser-His-Asp catalytic triad [40], which is very commonly found in a number of folds that have no significant sequence or structural similarity. Another example is functional convergence found in antifreeze protein (AFP, also known as thermal hysteresis proteins) [41]: they have a dissimilar sequence in plants and fish, but perform the same function, producing a difference between the freezing and melting points by depressing the non-equilibrium freezing point.

This phenomenon is quite common and often occurs to preserve the overall function of protein [42]. Understanding how the enzyme function evolved is vital to get insight for annotation, function prediction, and protein engineering [38, 43]. Also, this is one of the most intriguing problems in molecular biology: to understand the vast diversity of protein function.

2.2.1 More Definitions Suggesting Evolution Strategy

Evolutionary evidence supports the idea that the computational representation of enzyme function should include structural elements which deliver catalytic ability. This is especially so in cases where enzymes perform different overall functions by utilising similar mechanistic steps. Such understanding will aid our ability to predict the function of newly sequenced enzymes and in efforts to engineer new functions into existing enzymes.

As we mentioned in the previous section, one of the reasons for the false assignment of function to a novel enzyme is due to mechanistically diverse enzymes. A mechanistically diverse enzyme [44–46] superfamily is a set of enzymes that utilize common mechanistic attributes, such as mechanistic steps, to catalyse different reactions. An example supporting this scenario is phosphoglucomutase (MACiE: M0194; EC 5.4.2.8) and phosphonoacetaldehyde hydrolases (MACiE: M0181; EC 3.11.1.1). Another well studied example is the pentein superfamily (CATH 2.60.40.1700) [47] which are functionally diverse proteins grouped together based on similarity at structural fold level β/α. That includes enzymes that modify guanidines. The enzymes in this superfamily participate in diverse biological roles including gene regulation, translation and signalling. Assigning structure and function to penteins is difficult due to low sequence similarity between members of this superfamily.

Another definition that is preferred for classification of enzyme function is ‘functionally distinct enzymes’. ‘Functionally distinct enzymes’ are groups of divergently evolved enzymes which perform different overall reactions and for which no common mechanistic steps are found to complete the reaction [45, 48].

2.2.2 Biostatistics to Study Evolution

Understanding how well enzyme function adapts its nature is still a challenging task in molecular biology. To understand evolutionary trends of proteins, the biological data can be represented quantitatively to study the overall trend. The quantification of the relationships between various genomic and molecular variables are termed as ‘quantitative evolutionary genomics’ [49].

Quantitative evolutionary genomics has helped to understand dependency between structure and function [50]. For example, Log-normal distribution shows the global trend of evolution rates between orthologous genes [49] and Power-law like distribution [49, 51] represents a membership in paralogous gene families. A power-law-like distribution shows that a few parts occur many times and most occur infrequently.

2.3 Bioinformatics

The functional annotation of newly discovered proteins is a fruitful area for the application of the phenomenally large quantity of protein data now available. Broadly speaking, there are two major concerns related to the annotation of enzyme function: first, the clear definition of an enzyme function, which is important in order to get correct prediction, and second, a fast and efficient bioinformatics approach to manage the huge datasets as well as save time. Developing robust bioinformatics tools for managing huge data sets or mapping onto genome is ongoing research with many benefits for reducing the execution time of function annotation tasks.

Furthermore, the development of bioinformatics tools will improve the quality of addressing biological questions, which in turn improve the explanation of the enzyme chemical mechanism, molecular evolution, and structure-function relationship [52].

The challenge is found when the definition of protein's function is not clear. Protein function annotation is a multi-step process, proceeding from sequencing the corresponding nucleotides to building a predictive model for annotating function. In practice, it is done by classifying newly discovered proteins based on functional domains, folds or motifs followed by assigning homologous annotation.

In this section, we discuss the challenges faced during the performance of a wet experiment to annotate function and how designing bioinformatics models, tools and approaches can be beneficial. Further, we also discuss some biostatistical methods that are currently very famous in exploring problems for clear interpretation of enzyme function prediction and understanding the diversity of enzyme function.

2.3.1 *In lieu* of Wet Experiments

Even before high throughput techniques, biologists had collected large volumes of data that end up in repositories where most of the proteins are annotated as 'hypothetical protein' [53]. The final aim after the data is produced through sequencing experiment or determining structure of the protein is to annotate the produced sequence and make that available publicly such that this information can be used for further studies (see Figure 2.1).

One of the major goals of developing a robust bioinformatics approach is to provide a cheaper and faster computational alternative to wet lab experiments. For better annotation of protein function, the vast accumulation of data that is currently available may hold enough information to be used. Such experiments would be dependent on previously determined targets to predict outcomes for newly discovered proteins. For example, given previous sequence signatures, it is possible to characterise protein function [23] with good accuracy.

Another goal of this work is to understand the underlying mechanism of how enzymes adapt to execute sophisticated reactions. To achieve this goal, one can accumulate selective information from various sources to guide an experimentalist in designing and coordinating an experiment. One of the important data sources is the Enzyme Portal [54] which integrates publicly available information about enzymes, such as small-molecule chemistry, biochemical pathways and drug compounds. This portal is designed to display available enzyme-related information publicly. Bringing together chemoinformatics and biological data, so that information can be explored in one place, is useful for further studies and a very important step where ample amount of data are available.

2.3.2 Functional Prediction by Database Search

Various bioinformatics methods are designed to identify the function of a novel protein sequence or structure through a homologous search in a database. This is a very straightforward way to identify the function of an unknown protein by searching for a strong similarity from the protein sequence against available databases such as RefSeq or UniProtKB/Swiss-Prot. With the help of a search algorithm, such as BLASTp, PSI-BLAST (Position-Specific Iterating BLAST) [55], one can achieve this task. Information available in databases is very valuable for understanding enzyme function and for annotating function to novel enzymes. However, such information may not be enough for final prediction due to misannotations of definitions present in databases [56].

The annotation of a newly discovered protein can be done by using either the sequence or the structure of the protein.

The annotation of a sequence or a structure is important as it bridges the gap between the sequence or structure to the biological process of the

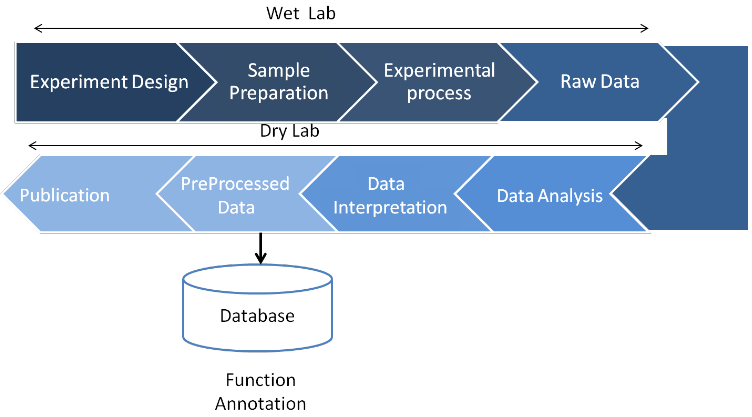


Figure 2.1: This figure illustrates workflow to generate raw data through 'wet' bench experiments which is further considered for bioinformatics analysis. It is becoming very popular to combine the experimental and computational approaches. The main aim of this is to produce publicly available data and annotate newly discovered proteins.

organism. However, similar sequence does not necessarily mean similar function [57]. This suggests more detailed analysis is required so that we can collect attributes that could help for annotation.

To overcome this difficulty one can use a structure homology search instead [58]. Structures are 3-10 times more conserved than sequences [59]. Nonetheless, in biology we can find many striking examples suggesting similar protein structures perform different functions. Gerlt and Babbitt [60] have shown that the functionally diverse enolase superfamily has conserved structures that include the $(\beta/\alpha)_8$ TIM barrel domains, sharing a common catalytic site.

The existence of such examples in nature adds to the difficulty of annotating function to enzyme structure or sequence. In summary, there are many enzymes whose sequence / structures are found to be very similar but which perform different functions, and the reverse is also true. We have discussed some examples and definitions in the previous section.

It is found that 'mechanistic' and 'transformational analogues' are not a rare phenomenon [61] in nature. Here, 'mechanistic analogues' are defined as those enzymes which use the same mechanism to perform related reactions, in other words they share a similar EC number until the third level. The 'transformational analogues' are enzymes which share identical EC numbers but use different mechanistic steps to effect the reaction, for example the metallo-beta-lactamase. To annotate such examples is difficult using homologous search. Thus, we need more sophisticated and robust methods such as machine learning algorithms that have already begun to successfully address this question [2, 62, 63].

It is noteworthy that assignment of function prediction depends on the accuracy of the database entries which in turn depends on any divergence of sequence and function in the course of evolution [56]. Moreover, the subjectivity of the definition of function and various aspects of biology could lead to false prediction of the function annotation [64]. The assignment of function to domains also depends on multi-domain organization of proteins, and low sequence complexity [65].

Another difficulty is where proteins can have a molecular function, a cellular role and be part of a complex pathway. Simplistically, one can use the EC numbering scheme to annotate the function.

There are various databases available for annotating function through homology search, such as RefSeq, PDB, InterPro etc. In addition to these databases, there are many databases available where information is integrated in one place such as FunTree, UniProt, SFLD. Also, for further investigation one can examine proteins of analogous function with the information available in MACiE. In order to perform a database search one needs to keep the database up-to-date, which means assigning meaningful biological information to newly discovered proteins. Collecting, organizing, and interpreting such data often requires the input of experts in the biological field of study [66].

2.3.3 Functional Prediction by Biostatistical Methods

With recent developments in biostatistics and advanced statistical methods such as machine learning algorithms, it is possible to design an efficient predictive model to extract patterns from a given dataset [2, 62, 67–69]. One strategy for predicting function is to reduce the problem to a classification exercise using data whose ontology allows each item to be identified with a specific category, for example top level EC class.

In statistics, classification is defined as a problem that identifies the patterns in datasets and further arranges together those which are alike and separates those which are unlike. The idea is that the new individual items should be placed into groups based on one or more criteria. The application of classification analysis can be seen in various research fields such as bioinformatics, cheminformatics [68], drug discovery, toxicogenomics and many more. Classification methods could be further divided into two approaches, as supervised and unsupervised algorithms.

Supervised algorithm A supervised classification algorithm uses defined examples to learn patterns and, based on this learning, it then classifies the data points. It has been shown that many machine learning methods have profound improvement on the prediction of function when chemoinformatics descriptors were designed using sequence or structure information [23,70,71]. In bioinformatics and system biology, machine learning methods are widely applicable [72].

The machine learning method focuses on prediction, based on learning from known properties from the training data. Machine learning algorithms are also beneficial in investigation into the structure of the data and handling a huge dataset. In other words, a machine learning algorithm constructs a model in order to predict the outcome of an experiment. However, the disadvantage of machine learning algorithms is that they hugely depend on the nature, source and quality of the data.

We preferred supervised classification methods because we wanted to see if our descriptors could show patterns which could be used further to annotate enzyme functions. This is a machine learning task, to infer function from given examples. Although algorithms of machine learning methods are often very complex, they nonetheless work on the very basic philosophy of learning from examples.

Machine learning algorithms are applicable to problems such as predicting the structure of a protein [72]. Prediction of protein structure has been a challenge for decades and with the advent in technology we are able to get some successful output. Machine learning methods have many advantages to map the input sequence of amino acid to the features of output sequences.

Where an inadequate amount of information is available for two proteins sharing the same function annotation, machine learning algorithms can be very profitable through extracting more information from multiple proteins [2, 23, 73, 74].

Unsupervised Clustering Analysis As its name suggests, this type of classification is not supervised by any examples. In this method, the data points are grouped together based on some criteria such that they can distinguish between self and non self. There are two ways a distinct cluster is defined, either by finding greater similarities within the members of the group or finding clear separation between the clusters. Sometimes clustering analysis is only used to summarise the data which could further be used for analysis purposes. In biology, bioinformatics, pattern recognition and social science, cluster analysis is commonly used for understanding the data or to annotate function.

Indeed, human eyes are skilled in grouping objects based on certain criteria, for example, a child can label a photograph as a building, vehicle, people etc. Biologists have used clustering analysis to create a taxonomy of living things: kingdom, phylum, class, order, family, genus and species. Among many unsupervised classification algorithms, hierarchical clustering is the simplest and very popularly used by biologists. An example of hierarchical clustering in biology is Gene Ontology (GO) which classifies genes into hierarchies of biological processes and molecular functions. Moreover, three structural classification databases which define sequence-structure-function relationship are SCOP, CATH and DALI. The EC nomenclature and classical taxonomy are both hierarchical methods used to classify enzymes based on biochemical classes and organism-level morphological features, respectively.

Another database using microarray data to study a large variety of biological mechanisms, including association with diseases, is the Database for Annotation, Visualization and Integrated Discovery (DAVID) [75, 76]. This database is popularly used to understand biological meaning of gene list using various sophisticated statistical methods.

As indicated schematically in Figure 2.2, biophysical information, with bioinformatics analyses of an entire set of related or non-related proteins, can be used to identify novel function by one of the two strategies using either supervised or unsupervised classification method. A new protein can be classified either as a member or non-member depending on its feature vector by using these machine learning methods.

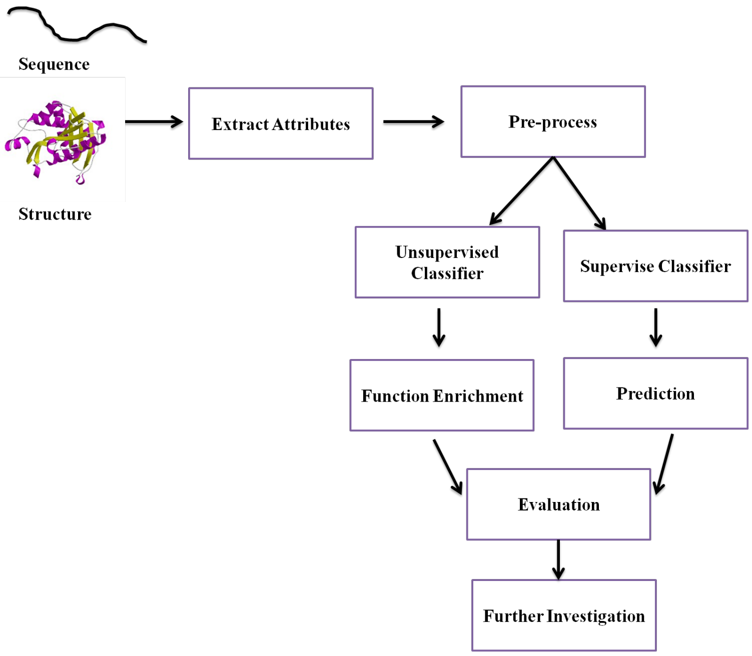


Figure 2.2: Schematic representation of an interactive approach to function annotation and prediction. It starts with extracting useful information from sequence or structure such as catalytic residues, reaction entities, to build a predictive model. Once the data is pre-processed, it can be further used for annotation using either supervised learning or unsupervised method. To get fewer false positives, the practice of evaluation is highly recommended.

2.3.4 Enzyme Catalytic Tool Kit

Our intentions with information related to catalytic residues are to investigate sequence property to understand the mechanism of the enzyme reaction. Based on input types, prediction methods are divided into sequence-only, structure-only, or a combination.

Multiple sequence alignment and database search are the two easiest ways to annotate newly discovered proteins, where two proteins are evolutionarily related. However, examples exist when the related sequences are diverged so much that they lose any evidence of homology [77]. In either of the cases, catalytic residues proved to be a good predictive attribute.

Using sequence information, such as sequence profile or location of the active site one can design a computational approach to annotate the novel function [78]. Considering only functionally annotated sequence, catalytic residues and position of catalytic residues within the sequence for glycohydrolase family of enzymes Sterner et. al. [78] could obtain $\approx$ 80% of prediction using kNN classification method.

Chou et al. [79] extracted the catalytic site from the non-catalytic sites of serine hydrolase with $\approx$99% predictive accuracy using a covariance matrix. Many times the enzyme catalytic sites are mistakenly predicted as non-catalytic antibodies [80]. It is noteworthy, while developing the prediction model, that the catalytic pocket of the enzymes is buried much deeper inside than non-catalytic antibodies.

The gap between the sequence profile and function annotation is usually bridged based on the sequence similarity with enzymes whose functions are experimentally known [81]. Even with the excellent databases detailing biological properties embodied in protein sequence patterns and motifs, this method shares many of the limitations of sequence alignment approaches. One basic problem here is that sequence is not as conserved as structure.

Structures of proteins are much more conserved in nature than their sequences [82]. Distant evolutionary relationships, undetectable by sequence comparisons, can be revealed by similarities in structures. Common ancestry can be indicative of a functional relationship, although the correlation between fold and function is strong for only some folds [83], others are highly functionally promiscuous. Structural genomics seeks to determine the structures of all protein folds, which would ultimately be highly valuable for data annotation and have other applications such as drug design. Luiz C. Borro [84] predicted enzyme function classification at the top EC level with an accuracy of 45%. In his study, he combined the strengths of statistical and data-mining methods using structural parameters. On the other hand, Dobson and Doig's [85] approach combined many support vector machine models with structural parameters and could predict EC class to an accuracy of only 35% for the top ranked prediction.

The idea of such methods is to give clues for functional annotation of proteins which could be used for further analysis. From the above mentioned studies it can be concluded that predictive models (either supervised or unsupervised method) for prediction of enzyme class from protein sequence or from its structure is tricky as the prediction accuracy is pretty low. Thus, we need to develop more computational approaches that integrate various sources for function prediction. The interpretation of the results for the biologist will be improved when computational power is improved.

2.4 Applications

2.4.1 Enzyme Engineering

Enzyme design and engineering is the ability to incorporate rational change to the protein structure to enhance the activity of an enzyme. Enzyme design has fruitfully shown its worth in producing new metabolites, to allow new pathways for reactions to occur and to convert certain compounds into others [86, 87]. To achieve the goal of enzyme engineering one can alter the substrate specificity (e.g., $NADP^+$ versus NAD^+) and catalytic efficiencies without altering the overall reaction of enzymes [86, 88, 89].

The ultimate goal of redesigning enzymes is to enhance their catalytic efficiency, bearing in mind the properties of enzyme structure and chemical reactions which closely resemble the enzyme reaction in nature [90].

Two complementary approaches have been developed over the past decades for enzyme engineering: directed evolution and rational design [91]. Directed evolution mimics the natural evolution process in the laboratory which could be achieved via two different pathways, one by randomly recombining a set of related sequences(e.g. gene shuffling) and the other by random changes in one protein sequence (e.g. error prone PCR). Whereas rational design involves alteration of knowledge-based specific and selected residues in a protein to cause predicted changes in function, which are introduced by site-specific mutagenesis.

Understanding catalytic structure has potential application in improving the enzymatic activity by selectively introducing single point mutations in the proteins [87]. For example, mutation of His94 to Asn uncovers an aldolase activity from L-ribulose-5-phosphate epimerase, which is a key enzyme in the bacterial arabinose metabolic pathway. This mutant is able to perform aldol condensation, which was not present in the wild type [8].

2.4.2 Drug Discovery

The use of drugs to treat various maladies has a long history, but the modern pharmaceutical industry, which is based on science (rather than tradition or superstition), is a product of the twentieth century. Identifying the target of disease is not sufficient in order to achieve successful treatment, a lead drug needs to be developed. There are many aspects that should be considered while engineering a drug. For example, a drug must influence the target protein in such a way that it does not interfere with normal metabolism [92]. Bioinformatics methods have been developed to virtually screen the target for compounds that bind and inhibit the protein. Bioinformatics methods and tools play a vital role in every aspect of drug discovery, drug assessment and drug development [93, 94].

Typically, drug discovery starts by identifying the target protein, choosing a biochemical mechanism involved in a disease condition, and the process is completed by approval from FDA, Food and Drug Administration.

In recent years, many major pharmaceutical companies have invested heavily in high throughput screening (HTS) [93, 94]. Not only conventional multivariate statistical methods, i.e. principal components' analysis and partial least squares, but also sophisticated machine learning methods are of great utility and continue to improve commercial tools [92].

Encouragingly, the machine learning methods have shown their potential to use chemoinformatics descriptors for developing methods for better drug targets [92]. Using chemoinformatics descriptors, which combine chemical properties and high throughput screening measurements, a classifier can be trained for 'virtual screening' for discovering molecules with specific therapeutic target affinities from potentially millions of representations [3, 92, 94].

In biochemistry and pharmacology, an active area of study is to target enzyme inhibitors as drug molecules [53]. During the lead compound screening, protein function prediction can impact on target selection. In cases where experimental information is not available, *in silico* function prediction methods can prove rewarding to provide functional insight for the target protein [53].

2.4.3 Clinical Implications

Enzymes play a major role in analytical diagnosis [6]. Due to their specific and selective nature, they are preferred as diagnostic analytes. The measurement of the enzymes in the serum level indicates damage in the tissue. For example, when a physician performs an assay for liver enzymes, the purpose is to indicate the potential damage to liver cells.

Alkaline phosphatase (ALP, MACiE: M0044, EC 3.1.3.1) is a hydrolase enzyme which is responsible for removal of phosphate group in many types of molecules. In children, these enzymes are important for the growth of bone. Increase in the concentration of these enzymes could lead to rheumatoid arthritis. Low concentration of such enzymes, which is rare, could cause hypophosphatasia [6].

Another such example is D-alanine transaminase (ALAT, MACiE: M0066, EC 2.6.1.2), commonly found in plasma. Clinically, the concentration of this enzyme has been used to diagnose hepatocellular injury in order to determine liver health [6].

An increase in our understanding of how enzymes work will surely have implications for the clinical treatment of disease.

2.5 Summary

The existence of a complex relationship between structure and function implies that there is a richness in the diversity of function that is still uncovered. This kind of work has the potential to be applied in enzyme technology, which is one of the corner stones of Industrial Biotechnology [95]. The benefit of knowing the structure-function relationship is to guide the experimental work such as site directed mutagenesis, protein-protein interaction studies and identification of ligands (e.g. inhibitors). The research in this area involves both fundamental and applied enzymology, biocatalysis, molecular modelling, structural biology and diagnostics. The overall goal is to develop new and more sustainable products, processes and services to meet human needs or to improve processes to produce existing products from new raw materials and biomass.

In summary, we can see that the structure-function relationship is complex, and annotating function of newly discovered proteins is not as straightforward as searching through a database. Here, we have explored the complex relationship between sequence, structure and function. In our opinion, the development of computational technologies will lead to more productive exploitation of the information contained in such data.

The next chapter is about data and databases, where we discuss in detail the databases from which we have extracted information for our work in this thesis.

3

Data and Databases

Enzymes are proteins which catalyse the chemical reactions necessary to support life and significantly enhance the rate of biological processes. Traditionally, enzyme functions are classified based on biochemical reactions using a system by the Enzyme Commission (EC) of the International Union of Biochemistry and Molecular Biology (NC-IUBMB)[1] [5]. The EC system represents the hierarchy based on the overall catalytic transformation, which fails to represent the correlation between structure and function [83]. To get a better understanding of the structure-function relationship and for predictability of function we require a good quantitative representation of enzymatic function. For this the data should be quantitative, and computationally accessible, informed by sequence and structure to enable use of genomic information for functional inference, and other applications. In this chapter, we discuss some databases and attributes that are derived to represent enzyme function at overall and mechanistic levels to get some insight into enzymatic functions.

A great deal of information [90] exists for enzymes in well known databases, including their 3D structure (CATH [96] and SCOP [97]), sequence (UniProt Knowledgebase[2]), catalytic reaction (Mechanism, Annotation and Classification in Enzymes[3] [98]), for metabolic pathways (Kyoto Encyclopedia of

[1] NC-IUBMB: http://www.chem.qmul.ac.uk/iubmb/enzyme/

[2] UniProtKB: http://www.uniprot.org/

[3] MACiE: http://www.ebi.ac.uk/thornton-srv/databases/MACiE/

Genes and Genomes: KEGG[4]) and kinetic data (BRENDA[5] [99]).

Some databases combine various information on protein structure, function, and catalysis to explore structure-function relationships, such as Structure - Function Linkage Database (SFLD[6] [100]) and Enzyme Catalytic - mechanism Database(EzCatDB[7] [101]) . Also, databases such as FunTree [102] map available information of enzyme function onto the phylogeny build based on protein structure to explore evolution of enzyme function. Moreover, the Caetano-Anollés group has developed Molecular Ancestry Network (MANET[8]) to facilitate evolutionary studies [103]. Such applications represent interesting and exciting ways to explore evolution.

In this chapter, we discuss resources that we use in our work to explore and understand enzyme functions. The data for overall study in this thesis is retrieved from the MACiE database. We also discuss the importance of data management in our work. By the definition of Bioinformatics (see Definition 2), it is clear that organising and arranging the data is important to get a clearer and better understanding of the relevant scientific questions.

Definition 2 (Bioinformatics)**.**

bio - *informatics*: bioinformatics is a tool to help understand the conceptual biology in terms of molecules and applying '*informatics techniques*' and to organise the informatics associated with these molecules. In other words, bioinformatics is an information management system to address scientific questions for molecular biology and its applications. This definition is adapted from [104].

3.1 Enzyme Mechanism Reactions Database

Our understanding of enzyme structure and function applies to the problems in enzyme engineering and drug design [83, 105]. In this section, we discuss the databases dedicated to enzyme function at overall and mechanistic levels. Also, we discuss various quantitative representations of enzyme catalysis.

[4] KEGG: http://www.genome.jp/kegg/

[5] BRENDA: http://www.brenda-enzymes.info/

[6] SFLD: http://sfld.rbvi.ucsf.edu/django/

[7] EzCatDB: http://mbs.cbrc.jp/EzCatDB/

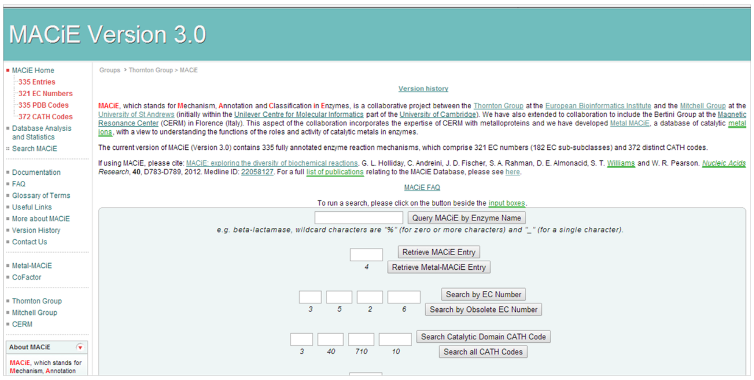
[8] MANET: http://manet.illinois.edu/

3.1.1 MACiE Database

MACiE integrates a wide range of enzyme reactions with details of stepwise chemical transformation in the reaction [98,106]. MACiE is a publicly available resource for enzyme reactions. The home page of MACiE (as shown in Figure 3.1) illustrates various search possibilities from the database such as by enzyme name, MACiE entry, by EC code, by CATH code as well, for information regarding enzyme mechanisms if known. MACiE holds entries for each EC sub-subclass representative, where the crystal structure of an enzyme is available in the PDB, and sufficient evidence of its mechanistic steps exists in the literature. The reaction steps of enzymes are thought to be an evolutionary unit of enzyme function evolution [107], which is the minimal unit present in the database to study. Such a database is valuable to advance our knowledge of the chemistry of enzyme reaction mechanisms and also for better understanding how enzymes adapt their chemical mechanisms under evolutionary pressure.

The current version of MACiE(V 3.0) possesses 335 enzyme entries [108]. An example of MACiE's representation of mechanism of catalysis, M0033 (EC Number: 5.1.99.1 , corresponds to methylmalonyl-CoA epimerase), is shown in Figure 3.2. The distribution of enzyme folds in the database is as follows: out of 335 entries there are 321 entries which are unique functions (EC level 1) at the 4th level of the EC number. These 335 enzyme entries are associated with 308 enzyme entries which are assigned to a CATH H homologous superfamily; the rest of the enzyme entries are not assigned to the CATH H-level superfamily. These 308 enzymes are associated with 236 unique CATH-H superfamilies.

A survey of the MACiE database [98] suggests that there are important catalytic residues such as histidine, cysteine and aspartate: these residues are more likely to act as reactants rather than spectators. The most commonly occurring mechanistic step types are proton transfer and nucleophilic reactions, which are supported by the generally nucleophilic nature of the naturally occurring amino acid residues [106]. Such studies suggest that MACiE possesses data that reveal the general trends in enzyme catalysis over a broad base of different enzyme reaction mechanisms and can be used for further such studies.



MACiE, which stands for Mechanism, Annotation and Classification in Enzymes, is a collaborative project between the Thornton Group at the European Bioinformatics Institute and the Mitchell Group at the University of St Andrews (initially within the Unilever Centre for Molecular Informatics part of the University of Cambridge). We have also extended to collaboration to include the Bertini Group at the Magnetic Resonance Center (CERM) in Florence (Italy). This aspect of the collaboration incorporates the expertise of CERM with metalloproteins and we have developed Metal MACiE, a database of catalytic metal ions, with a view to understanding the functions of the roles and activity of catalytic metals in enzymes.

The current version of MACiE (Version 3.0) contains 335 fully annotated enzyme reaction mechanisms, which comprise 321 EC numbers (182 EC sub-subclasses) and 372 distinct CATH codes.

If using MACiE, please cite: MACiE: exploring the diversity of biochemical reactions. G. L. Holliday, C. Andreini, J. D. Fischer, S. A. Rahman, D. E. Almonacid, S. T. Williams and W. R. Pearson. *Nucleic Acids Research*, **40**, D783-D789, 2012. Medline ID: 22058127. For a full list of publications relating to the MACiE Database, please see here.

MACiE FAQ

To run a search, please click on the button beside the input boxes.

e.g. beta-lactamase, wildcard characters are "%" (for zero or more characters) and "_" (for a single character).

Query MACiE by Enzyme Name
4
Retrieve MACiE Entry
Retrieve Metal-MACiE Entry
3 5 2 6
Search by EC Number
Search by Obsolete EC Number
3 40 710 10
Search Catalytic Domain CATH Code
Search all CATH Codes
MACiE, which stands for Mechanism, Annotation

Figure 3.1: This figure illustrates the homepage of the MACiE database that provides various gateways for searching information on enzymatic reactions.

3.1.2 Components of MACiE

Information provided in MACiE can be divided into catalysis specific information and non-catalysis specific information (see Table 3.1). In the following section, we discuss the catalysis specific information and the quantitative representation of the data fetched from MACiE.

Table 3.1: Annotation Components in MACiE

Catalysis specific information	Non-Catalysis specific information
Enzyme name	PDB code
EC code	non-catalytic UniProt Code
Catalytic CATH domain	non-Catalytic Domain CATH code
Reactant and products	Species name
Bond involved and formed, cleaved and changed in order	Other databases such as KEGG, GO, SFLD
Reactive centres	Literature
Overall Reaction comment	
Mechanistic annotations	

3.1.3 Catalytic Information Available in MACiE

Here, we consider only catalytic information from the MACiE database that is appropriate to design scaffolds for engineering new functions, such as catalytic residue and mechanistic step types. To understand enzymatic function we use bond involved, formed, broken and changed in order, and quantify these features into chemoinformatic descriptors [20]. We choose five descriptor sets to represent enzymatic reactions by using fingerprints from [20] and similarity scores using [33]. The *human designed* descriptor is designed in-

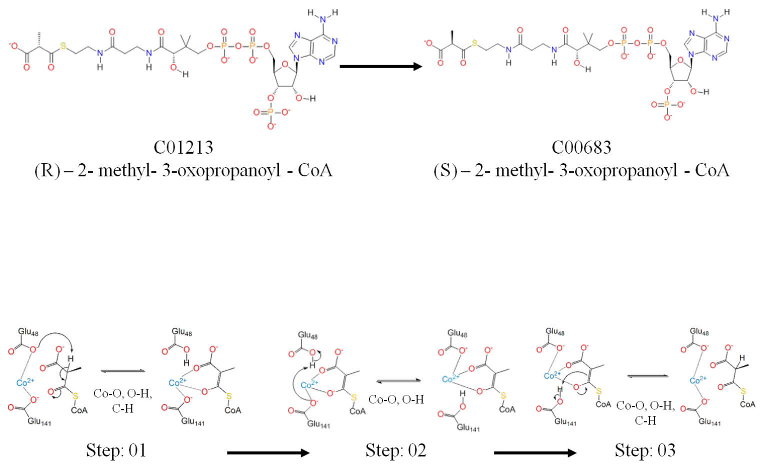


Figure 3.2: This figure shows an example of the overall reaction and catalytic steps of M0033 represented in MACiE. M0033 is a member of isomerase (enzyme name: methylmalonyl-CoA epimerase (EC 5.1.99.1)).

tentionally to express the important and correlated features of enzyme activity. One would expect to get accurate function prediction if the quantitative representation of enzyme function follows the definition of EC classification. Here, we discuss chemoinformatics descriptors to understand ‘best’ features from enzyme function clusters (detail in Chapter 4), whether representation in terms of mechanistic steps is better for enzymatic function prediction (detail in Chapter 5) and the evolutionary role of these enzymatic functions (detail Chapter 6). Also, we discuss the mechanistic annotations retrieved from MACiE for Chapter 4 and 6.

***Human Designed (HD)*:** This descriptor is designed in order to represent the specific features involved in enzymatic activity. These features are based on the overall reaction, not mechanism. For example; f:X-H, the total number of bonds to hydrogen formed. Some features quantified in this descriptor are deliberate attempts to correlate with specific EC classes, such as water.OH-.su is set to 1 whenever water (or OH-) is a substrate on the left hand side of the overall reaction equation, it should take the value 1 for all EC 3.__.__.__ . Moreover, one feature also calculates the difference between the molecular weight (Mod_Diff) of the largest substrate and largest product, this is specifically beneficial to detect EC 5.__.__.__.

***Overall Bond Change (OBC)*:** This descriptor is designed based on the overall reactions of enzymes. Holliday et al. [31] list the number of covalent bonds between a given pair of elements i.e. count both the number and the bond change. For instance, the descriptor C.C_0.1 gives the number of carbon-carbon single bonds formed in the reaction, O.O_2.1 is the number of oxygen-oxygen double bonds in the starting materials that become single bonds in the products. These descriptors depend only on the overall reaction, not on mechanism.

***Overall Reaction Similarity (OS)*:** The algorithm computes the overall reaction similarity of any two of the 260 mechanisms present in MACiE V 2.4. The similarity score is the Tanimoto coefficient of the bond changes in the reaction. Due to the nature of similarity matrix, also to prevent overfitting, similar columns and rows are deleted for training the model. These numbers range between 0 and 1 where 1 means the identical reaction.

***Composite Bond Change (CBC)*:** MACiE also represents stepwise reactions of the reaction pathways. This information is represented [106] in this descriptor. The *Composite Bond Change* descriptor is derived by summing all bond changes in a stepwise reaction mechanism; hence they depend on mechanism, not just on overall reaction. This means that, for example, a C-O single bond formed in one step and broken in a subsequent one will appear as both C.O_0.1 and C.O_1.0 in the *Composite Bond Change* description of the mechanism.

***Mechanistic Similarity (MS)*:** The mechanistic similarity is computed by first aligning the steps of two reactions using a Needleman-Wunsch algorithm, as explained in [33]. The similarity is calculated in both directions, first canonical in MACiE and then reverse, the best similarity out of two is selected for further analysis. Thus, the range of similarity score is between 0 and 1 where 1 means identical mechanism.

***Definition of molecular mechanism (mechanistic step types)*:** For enzyme mechanistic step type definitions, the data are retrieved from the MACiE database, specifically the functional annotations describing the chemical nature of individual reaction steps; frequently observed examples are ‘proton transfer’ and ‘bimolecular nucleophilic substitution’. These MACiE annotations relate specifically to the steps of the mechanisms by which the reactions occur, rather than to the overall chemical transformation; the EC number covers the latter. This is one of the important and interesting features in MACiE; that it provides mechanistic annotations[9] [106] or mechanistic definitions for each step of the catalytic reaction. For example, in MACiE, M0033 (methylmalonyl-CoA epimerase, EC 5.1.99.1) uses three mechanistic steps to convert substrate ((R)-2-methyl-3-oxopropanoyl-CoA) to product ((S)-2-methyl-3-oxopropanoyl-CoA) by using two mechanistic annotations’: ‘proton transfer’ and ‘assisted keto-enol tautomerisation’. There are 51 mechanistic definitions available in MACiE[10], among which ‘proton transfer’, ‘bimolecular nucleophilic addition’ and ‘unimolecular elimination by the conjugate base’ are the most commonly preferred annotations in order to complete the reaction (see Figure 3.3).

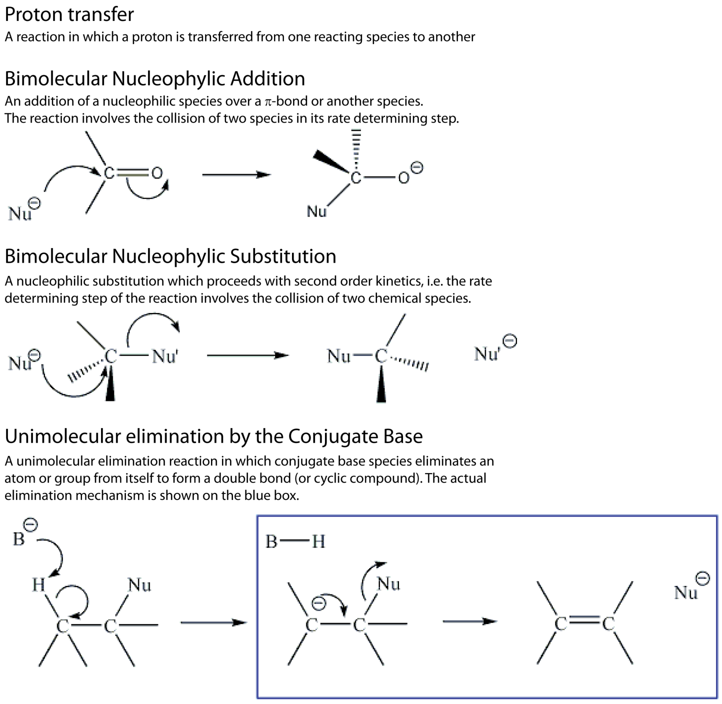
3.1.4 Definitions of Enzyme Function

Enzyme Commission number All characterised enzymes have an Enzyme Commission number (EC number), also known as IUBMB’s Enzyme nomenclature system [5], that represents the overall chemical transformation of a substrate to a product. This is a very popular system to classify enzyme activity. The classification system is a hierarchy where the reaction is divided into 6 classes, (oxidoreductases: EC 1, transferases: EC 2, hydrolases: EC 3, lyases: EC 4, isomerases: EC 5 and ligases: EC 6), which are then split at a further three hierarchical levels. The second level subclass and third sub-subclass usually describe the bonds or functional groups of the enzymes. The fourth level defines the actual substrate in the reaction. Identifying and annotating enzyme function is important for biological, medical, environmental and industrial problems.

Although the EC number system is very useful for many tasks, this system fails to reflect any mechanistic, sequence, protein structure or evolutionary information. Its design represents only the overall reaction of an enzyme, which makes it difficult to use for global classification or comparisons of biochemical activities. Indeed, the classification was designed before

[9] “mechanistic annotations" is used interchangeably with “mechanistic step types"

[10] The reaction definitions are available in following URL: http://www.ebi.ac.uk/thornton-srv/databases/MACiE/glossary.html



Proton transfer

A reaction in which a proton is transferred from one reacting species to another

Bimolecular Nucleophylic Addition

An addition of a nucleophilic species over a π-bond or another species.
The reaction involves the collision of two species in its rate determining step.

Bimolecular Nucleophylic Substitution

A nucleophilic substitution which proceeds with second order kinetics, i.e. the rate determining step of the reaction involves the collision of two chemical species.

Unimolecular elimination by the Conjugate Base

A unimolecular elimination reaction in which conjugate base species eliminates an atom or group from itself to form a double bond (or cyclic compound). The actual elimination mechanism is shown on the blue box.

Figure 3.3: This figure illustrates the definitions of the mechanistic step types, which include fundamental building blocks of enzyme chemistry: 'proton transfer', 'bimolecular nucleophilic addition', 'bimolecular nucleophilic substitution', and 'unimolecular elimination by the conjugate base'. We follow MACiE's terminology, though the latter could perhaps be better described as 'unimolecular elimination from the conjugate base', being the second and last step of the E1cB 'unimolecular elimination via the conjugate base' mechanism. These definitions are adapted from MACiE glossary www.ebi.ac.uk/thornton-srv/databases/MACiE/glossary.html

much information concerning enzyme structures and mechanisms was available.

MACiE is designed to be as complete as possible at the 1st, 2nd and 3rd levels of EC, but only representative at the 4th level. Its coverage, relative to the numbers of nodes for which the PDB structures exist, is 6/6 (1st level); 54/57 (2nd level); 165/194 (3rd level); 249/1547 (4th level) according to figures collated in 2010 [31, 98, 106].

QuickGO Gene Ontology (GO) is an impressive consortium for the development of control vocabularies for shared use in different domains [109].

The GO database represents the relationship between gene and gene product information across all species. The GO terms describe the functions at three levels: molecular functions (the activity described at the molecular level that the protein performs), biological process (assemblies of proteins and their functions such as in a metabolic pathway) and cellular components (the location where a protein performs its function). A review [110] discussed shortcomings of the GO database. One of the major shortcomings of using annotations from GO [111] is how to deal with a term that is represented in several, possibly overlapping, ontologies. For example, in MetaCyc [112] contains a term for which the ID is 'NAD BIOSYNTHESIS III'. This term is synonymous to GO:0019360 in GO, which corresponds to nicotinamide nucleotide biosynthesis from niacinamide.

One can also find GO information in MACiE entries. For example, M0033 (enzyme name:methylmalonyl-CoA epimerase, EC: 5.1.99.1) suggests two molecular functions: isomerase activity (GO:0016853) and methylmalonyl-CoA epimerase activity (GO:0004493). This information is directly annotated using AmiGO[11].

3.2 Metal-MACiE

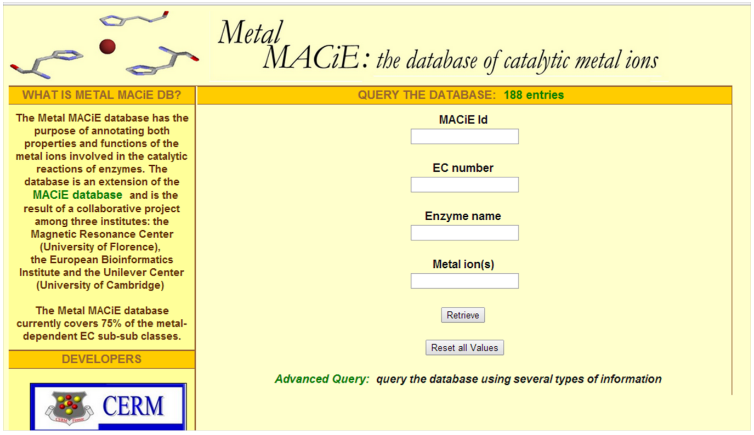
The Metal-MACiE database is a sister database of MACiE, possessing complementary information on enzyme metal cofactors[12] [24, 113]. The role of metal cofactors for enzymatic activity is well established. This information is collected and gathered in Metal-MACiE, providing properties and roles of metal ions involved in the reaction [113]. Currently, there are only 188 enzyme entries from MACiE that are metalloenzymes. These enzyme entries can be examined using various search terms such as using MACiE id, EC number, enzyme name or according to metal ions (Figure 3.4). The entries are selected based on the metal-dependent enzymes by annotating the enzyme structure using the PDBSProtEC database[13] [114]. Where enough evidence was not present to suggest the use of a particular metal ion in enzyme activity, a literature search for such evidence was conducted.

It appears that the metal ion which is most often involved in the functioning of enzymes is magnesium, followed by iron and zinc (see Figure 3.5). We

[11] AmiGO: http://amigo1.geneontology.org/cgi-bin/amigo/go.cgi

[12] Metal-MACiE: http://www.ebi.ac.uk/thornton-srv/databases/Metal_MACiE/home.html

[13] PDBSPortEC http://www.bioinf.org.uk/pdbsprotec/



Metal MACiE: the database of catalytic metal ions

WHAT IS METAL MACiE DB?

The Metal MACiE database has the purpose of annotating both properties and functions of the metal ions involved in the catalytic reactions of enzymes. The database is an extension of the MACiE database and is the result of a collaborative project among three institutes: the Magnetic Resonance Center (University of Florence), the European Bioinformatics Institute and the Unilever Center (University of Cambridge)

The Metal MACiE database currently covers 75% of the metal-dependent EC sub-sub classes.

DEVELOPERS

QUERY THE DATABASE: 188 entries

MACiE Id

EC number

Enzyme name

Metal ion(s)

Advanced Query: query the database using several types of information

Figure 3.4: This figure shows the homepage for Metal-MACiE where one can retrieve properties and roles of metals in the catalytic mechanisms present in MACiE.

found that in many Metal MACiE entries, the enzyme uses a two-metal ion catalysis that presumably involves the same metal ion, for example M0058 (adenylate cyclase, EC: 4.6.1.1) uses two metal ions and both are Mg^{2+} in the cell. For Figure 3.5, we consider the occurrence of one metal ion only. Figure 3.5 shows that magnesium is the most abundantly used metal cofactor, while second place goes to iron which almost every time is used in oxidoreductase reaction (EC 1._._._). Whereas, we found that zinc as a metal cofactor is not specific to any one type of biocatalytic reaction.

3.3 Molecular Ancestry Network (MANET)

The MANET database [103] contains information on enzyme structure traced onto an evolutionary time line. MANET explores the evolution of modern metabolism by mapping enzyme domain structural data from SCOP [97,115] onto the KEGG, and is based on phylogenetic reconstructions depicting the evolution of protein fold architecture (shown in Figure 3.6). The protein fold architecture age definition is designed and evaluated by Gustavo Caetano-Anollés et. al. [116] to explore the usage of enzymatic function in the dataset of folds.

Biocatalytic mechanisms operating in metabolic enzymes were traced along an evolutionary timeline of appearance of domain structures defined

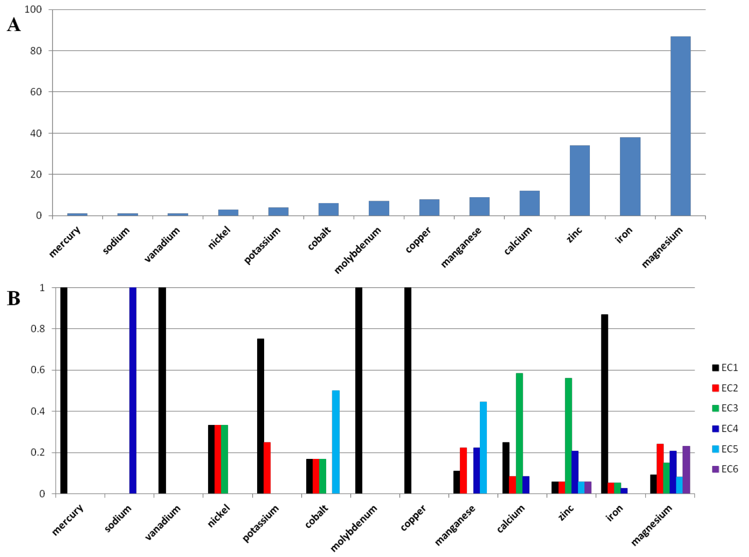


Figure 3.5: This figure shows the distribution of metal cofactors in the catalytic mechanisms of metalloproteins in the MACiE database, Part A. Part B represents the metal distribution specific to EC top class.

at the homologous superfamily (H) level of structural abstraction of CATH. Hereafter, we refer to these fold superfamilies as H-level structures. CATH unifies domain structures hierarchically from bottom to top into sequence families (SF), homologous superfamilies (H), topologies (T), architectures (A) and classes (C). H-level structures are considered evolutionary units. The timeline (*nd*) was built directly from a phylogenomic tree describing the evolution of 2,221 H-level structures [117].

3.4 Challenges and Limitations in Bioinformatics

3.4.1 Challenges Faced in Bioinformatics

Research in bioinformatics presents a number of exciting challenges and opportunities for biologists, computer scientists, information scientists and in particular bioinformaticists. These challenges are related to the current status of the flood of raw data, and evolving knowledge arising from the study of the genome and its manifestation. The challenges faced in bioinformatics and its applications have been reviewed multiple times by various

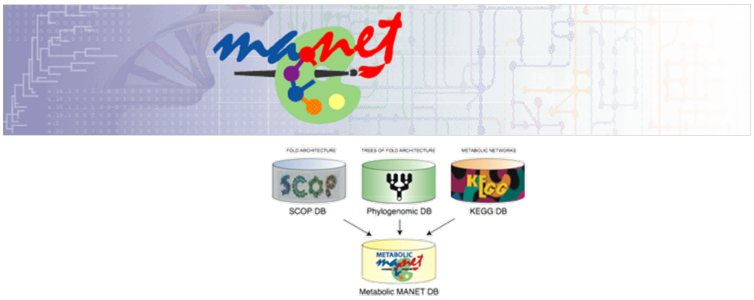


Figure 3.6: Principles of database design: The metabolic MANET database links the KEGG database, the SCOP database and a database of phylogenomic trees of protein fold architecture.

authors [104, 118–121]. By the definition of bioinformatics (see Definition 2), as mentioned in the beginning of this chapter, data mining, curation and pre-processing are among the exciting challenges [119] we have addressed in our work. These are important processes to deal with at the beginning of any study. Data curation and pre-processing are equally important whether treated together or separately. These issues together will condition the quality of pattern discovery as well as reduce the danger of over-fitting to the statistical models which further improves the interpretations.

It is important to gather relevant data to get better analysis, for example adding sequence or catalytic sites serves to facilitate better predictions of unannotated enzymatic function [23]. In Chapter 5, we have used mechanistic enzymatic activity to predict enzymatic function. Here, we use various pre-processing stages for the data in order to improve the prediction. There are various web sources that integrate relevant biological information together to give a clearer picture, such as SFLD, MACiE, FunTree and MANET. With these resources available it is important to perform rigorous data culling. This information is used for our work that is discussed in Chapters 4, 5 and 6.

3.4.2 Caveats

For the purpose of our work in this thesis we have retrieved data mainly from the MACiE database. In addition to MACiE we data mined mechanistic annotations, enzyme metal (MACiE-Metal), GO annotation (GO), enzyme pathways (KEGG) and evolutionary time line (MANET). Thus, it is important to mention that the result and interpretation mostly depends on the information in the database, which in turns depends on the manual extraction of information from the primary literature. Thus, the annotation used in our study is as good as that in the literature. There are some caveats or limitations that could lead to difficulties in interpretation; for example, in serine protease like mechanisms some literature studies note oxyanion hole stabilising residues and some do not.

Based on this potentially inconsistent information, each entry in MACiE is evaluated manually [106]. It should be noted that the information in the literature in turn depends on the accuracy and reliability of the structural data in the PDB, and other chemical and biochemical studies from which the catalytic residues and mechanisms have been proposed. In cases, where more than two possible mechanistic step pathways were noted, the most likely pathway was reviewed and selected by an expert.

3.5 Summary

In this chapter, we discussed the resources used in this work. We delve into the quantitative representation of the data as chemoinformatics descriptors coupled with the importance of data culling and pre-processing, which is at the heart of all interpretations. We also discussed the advantages and disadvantages of the resources we are using.

4

Quantitative Global Analysis of Enzyme Reaction Mechanisms

In this chapter we will discuss the clustering analysis of enzyme reaction mechanisms. Here, we analyse clusters of chemical mechanisms defined by chemoinformatics descriptors, using unsupervised global analysis. These descriptors map the chemical changes which are represented with curly arrows to track the electron movements. These descriptors have been explained in the previous chapter (Chapter 3). In this study, we use two descriptors: *OBC* and *CBC*. Formally, the enzyme reaction is defined by EC annotation, which lacks information for mechanistic steps. There is supporting evidence in the literature [33, 122] to suggest that in nature there are some examples of different overall reactions following similar reaction steps to effect the reaction. Studies of this kind give better insight at the finer level of enzyme function classification, further improving our understanding of structure-function relationships.

Fuelled by the availability of chemoinformatics resources, several methods have emerged to quantify overall [36, 69, 123, 124] and mechanistic [33] enzyme reactions. In this chapter, we investigate the significant patterns of reaction entities defined from MACiE [31, 106]. For this, we first clustered enzyme reactions using PFClust [1], seeking biologically rich information provided in the clusters of enzyme reactions.

Here, we will first discuss some similar studies where clustering analysis is performed using enzyme reaction, followed by brief descriptions of various clustering algorithms. Next, we describe the results produced by PFClust for enzymatic reaction descriptors and further its biological enrichment analysis.

4.1 Previous Studies of Cluster Analysis

Traditionally, enzyme overall chemical functions are manually classified by experts using the Enzyme Commission (EC) number system, lacking any evolutionary and mechanistic information. This EC system is a popular label to annotate enzymatic function and has many benefits. However, with the increase of data and computational technology it is of importance to design an automated system that can annotate newly discovered enzyme functions. A study of a similar nature was conducted by [69] where they represented overall enzyme chemical reactions in a vector fashion to optimize these enzymes' reactions into 21 groups, which is far more clusters than the six overall reactions in the EC classification. This suggests that we need robust representation of the data and method to determine the enzymatic reaction groups.

4.2 Data

The data was retrieved from the MACiE database [98]. MACiE is a repository of enzymatic reactions at overall and stepwise reactions. To find the similar patterns of enzymatic reaction, we use the *OBC* and *CBC* definitions by Holliday et al. [31]. The data matrix of the counts of bond formation, bond breakage or bond order change between each pair of elements are reported in the columns of the data matrix, with each enzyme reaction corresponding to one particular row. Next, the Tanimoto similarity of the enzyme reactions are calculated using "proxy" [125] packages in R [126]. For each of *OBC* and *CBC*, we calculate the similarities of all pairs of reactions, and each reaction is represented by a vector comprising 320 similarities.

For further investigation, data from different resources were mapped onto the clusters formed by PFClust. Resources such as mechanistic step types (or mechanistic annotation) are available in MACiE, for example 'proton transfer', 'electron transfer', 'bimolecular nucleophilic addition' and 'unimolecular elimination by the conjugate base'. Functional annotations were taken from QuickGO database [127, 128], for metal-cofactor we have used the Metal-MACiE database [24, 113]. For pathway analysis we gather information from the KEGG database [129].

4.3 Motivation

Global study of enzyme reaction mechanisms may provide important insights for better understanding of the diversity of chemical reactions of enzymes. Our motivation is to endeavour to address the challenge of how the chemical mechanisms of enzyme reactions cluster in a space defined by chemoinformatics descriptors, using unsupervised global analysis. Moreover, we designed a clustering algorithm, PFClust, which is parameter free.

For a global analysis of enzyme mechanisms, we have performed an unsupervised clustering analysis, the idea being to find the closest reaction neighbours. Before performing clustering analysis, we asked two questions: first, how to determine the number of clusters, and second, how to validate the results? Validation plays an important role in the analysis. Using an internal (Silhouette width) validation measure, one can decide the optimal number of clusters by selecting the best score from a range of possible numbers of clusters, and an external (Rand Index) validation measure can be used to compare the results with gold standards. Initially, we used well-established clustering algorithms such as hierarchical clustering, k-means, and density based clustering. Finally, we analysed results of PFClust.

4.4 Brief Introduction of Various Clustering Algorithms

Clustering is a very useful approach for discovering groups, identifying internal distribution and patterns. The main idea behind clustering algorithms is to group the entities together based on similar properties. For example, imagine a basketful of different colour (yellow, red, green) balls: a clustering algorithm will group these balls into homogeneous groups based on one property, in this case it is colour [130]. Broadly, the clustering algorithm can be grouped into three categories, first, hierarchical clustering [131], second, partition clustering [132, 133] and third, density based clustering [134]. These algorithms are well documented and have been applied to various research fields.

It is not uncommon practice, in biology, to use cluster analysis for determining natural structure in data. Clustering analyses have predominated not only in microarray data analysis [124, 135–138] for the interpretation of the results but also in neuroscience [139] and also for bioinformatics analysis [135, 136], image analysis [140], pattern recognition [138] and in pharmaceutical industry [141]. The simple task of clustering, by definition, has many limitations attached to it. In the following section, we will briefly describe various clustering algorithms and their limitations.

4.4.1 Hierarchical Clustering Method

For a given dataset D with n objects, in this study $n = 320$ *enzymes*. A hierarchical algorithm clusters a given set of n objects in stepwise fashion. First, the two closest objects are linked together, the closeness of the object is evaluated based on a (dis)similarity measure. There are various ways to calculate (dis)similarity, for example Euclidean distance or Tanimoto coefficient. In this study, we have calculated Tanimoto coefficients [125] for counts of enzyme reaction entities as a vector. The next step is to find the next connection between the first group that will lead to second level and so on. These steps are iterated until no data points are left to be connected as clusters [131].

The clustering algorithms preferring hierarchical strategy such as agnes seek the hierarchical structure in the data, where the objects are linked together. There are two approaches in hierarchical clustering algorithms: the top-down (also known as Divisive) and bottom-up (also known as Agglomerative) approach (for visual illustration see Figure 4.1). The difference between these two algorithms is that in one the clustering starts by considering every data point to be included in one cluster, and in the other, each data point is considered as a singleton before starting the clustering. The disadvantage of using this clustering algorithm is the vagueness of termination criteria. Without prior knowledge of data, it is difficult to determine the cut-off value for determining the number of clusters.

4.4.2 Partitional Clustering Method

In partitional clustering algorithms [132, 133], such as in the k-means clustering algorithm [133], one tries to find the centre of highly dense group

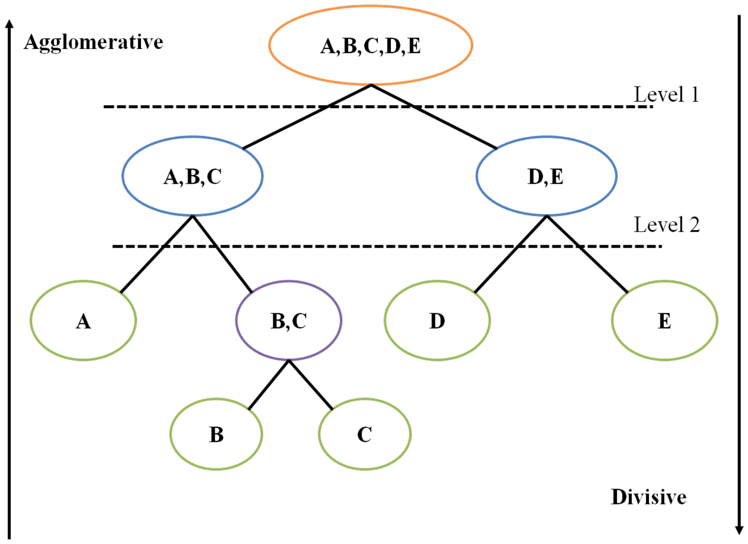


Figure 4.1: This figure illustrates the strategy of hierarchical clustering where top-down arrow suggests divisive clustering and bottom up suggests agglomerative clustering. Number of clusters can be decided at level 1 or level 2. In some cases, it is possible to get A, B, C, D and E (in this example) all in different clusters when using agglomerative clustering algorithm. makind A, B, C, D and E all singletons.

iteratively till a homogeneous group is formed. The partition clustering algorithm finds clusters simultaneously.

In k-means clustering, in the first step k clusters are either selected randomly or input by assigning as cluster centres, then each object is assigned to the nearest centre. Next, recalculation is done by calculating the mean of the elements in the cluster provided. Again the average of the cluster is treated as the next centre point and recruits new members iteratively. These steps are iterated until no reassignment of object occurs (see Figure 4.2). The obvious limitation found in this algorithm is to decide the number of clusters (k), as an input parameter. Deciding this parameter is not easy if one is working with a new dataset. In addition, this algorithm is sensitive to outliers and noisy data. Nevertheless, this algorithm is popular because of its easy interpretation and simplicity of implementation [142].

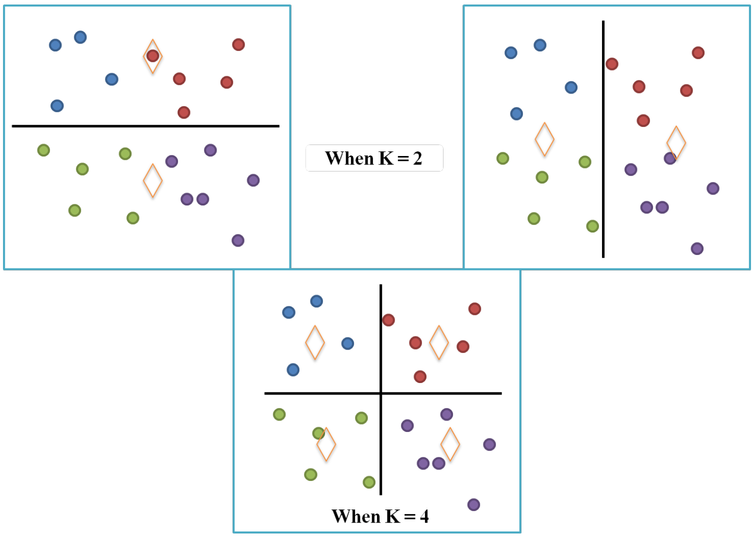


Figure 4.2: This figure illustrates k-means clustering algorithm. In this given example, when k is selected to be two, there are two possible ways this algorithm will optimise the groups, while when k is selected to be four it will find four groups in the data provided. In this algorithm parameter k plays an important role for determining the structure of given dataset.

4.4.3 Density-Based Clustering

Density-based clustering [134] clusters a highly dense group of objects together, separating them from the other density-based clusters. The first step is to calculate densities of the given dataset and then to determine the number of clusters of maximum density. In this algorithm, clusters are regarded as regions in the data space in which the objects are dense, and which are separated by regions of low object density (noise). Such algorithms are very efficient to handle large and noisy datasets. An example of density-based clustering is DBSCAN [143].

The DBSCAN algorithm recognises clusters of local density using two global parameters: ϵ - (Reachability distance) and $MinPts$ (minimum number). A cluster in this algorithm is defined if at least two objects lie within defined radius ϵ, also known as ϵ - neighbourhood, and there is a minimum number of points, $MinPts$, within that distance. For example, in Figure 4.3, instance p is a part of a cluster C and instance q is density reachable from point p with minimum number of points if q is also a part of cluster C. Here, q is a core object which obeys the reachability distance containing at least minimum number of objects. As a consequence, it can find arbitrary shapes of the cluster in the given dataset. The results are highly dependent on these two parameters. The popular application of this algorithm is in image analysis as well as in zoology, specifically for species divergence [134, 143].

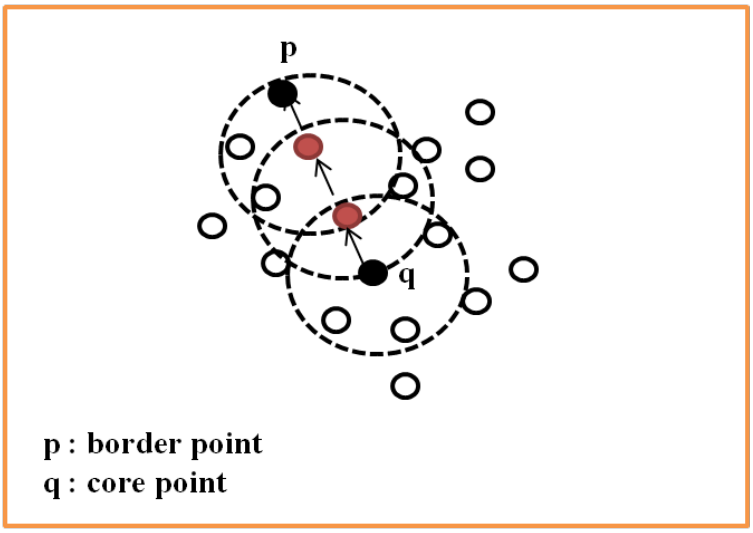


Figure 4.3: This figure illustrates a part of the algorithm to find the density population in the data based on two parameters explained in main text. This figure shows the connection between the core data point q and border point p. Definition explained in main text.

4.4.4 PFClust: Parameter Free Clustering

Our motive in this work is to overcome the limitations of deciding the number of clusters, hence we aim to design a novel clustering algorithm PFClust[1] is to overcome the pitfalls of current methods. The idea behind this algorithm is to find "meaningful" groups, where enzyme reactions can find themselves clustered with similar reactions as neighbours. This clustering algorithm is suitable when no prior information, except similarity matrix, is available. This algorithm is discussed in four sections; first randomization, from which 20 candidate thresholds of intra-cluster similarity (T_i) are retrieved, second the clustering for each threshold, third, cluster optimization, and fourth, once the best threshold (T) has been selected, convergence. The PFClust algorithm is also illustrated in Figure 4.4.

Step 1 Randomization: The first step in this clustering analysis is to estimate the thresholds. For a given data set, a random number of clusters (k; $1 \leq k \leq n$, where n is the number of data points) are chosen and each data point is randomly assigned to a cluster. For each cluster i, the mean intra-cluster similarity was computed using Equation 4.1: which is defined as the expected value of the distribution,

$$E[X_i] = \frac{1}{\binom{n_i}{2}} \sum_{j=2}^{n_i} \sum_{q=1}^{j-1} S(\alpha_j \alpha_q) \tag{4.1}$$

where n_i is the number of members of the cluster i and $S\alpha_j\alpha$ is the similarity between elements α_j, α_q. This step is repeated $N = 10^r$ times, where r is the iteration of the algorithm. We retrieved 20 T_i from the intra-cluster mean similarity, $E[X_i]$, distribution at 95% - 99.75% significance levels, that is the 5% of the clusters with the highest mean intra-cluster similarities. Using this number of thresholds provides a way of reducing the random element of our sampling. This step is illustrated in Figure 4.4 in Step B.

Step 2 Clustering: Now, for each T, a similarity–based clustering is performed for the given data set. First, the two most similar elements are placed together in a cluster. In each iteration, a new element is added to

[1] PFClust is designed on the JAVA platform

an existing cluster, provided that two criteria are fulfilled. First, the average similarity between the new element and the existing members of a cluster must be greater than $0.85 * T$, and second, the resulting intra-cluster mean similarity must be greater than T. The P% of T cut-off was selected as a way to restrict the intra-cluster variation of the similarities since, in a very tight cluster, outlier members could be included because, even if they are distant from the other cluster members, the total $E[X]$ could still be above T. A value of P = 85% of T gives the optimal results with respect to the Silhouette width as well as the number of clusters (see Table 4.1), with multi-member clusters and singletons being shown separately.In this way, clusterings are obtained for each of the 20 values of T_i. This step is illustrated in Figure 4.4 in Step C.

Table 4.1: The table summarizes the performance of the different P values for the threshold inclusion rule. The numbers of multi-member clusters and singletons are given separately, so that the total numbers of clusters at each P values are 36, 14, 11, 10, 10, and 10 respectively. The Silhouette width and the average of the standard deviations of the distributions of intra-cluster similarities in each cluster are also shown.

P Value	Clusters	Singletons	Silhouette width	Avg Std
T	30	6	0.1719	1.1740
0.95*T	10	4	0.5240	4.2715
0.90*T	10	1	0.5650	4.3176
0.85*T	10	0	0.5961	4.6604
0.80*T	10	0	0.5955	4.8175
0.75*T	10	0	0.5955	4.8175

Step 3 Cluster optimization: Here, we optimize the cluster memberships for the clustering for each value of T_i. For each point in each cluster, we calculate its average similarity with all members of its current cluster. If this average similarity $<T$, and if its average similarity with the elements of any alternative cluster is greater than with the parent cluster, then we move the point to the alternative cluster. Based on the highest Silhouette width (averaged over all points, singletons counting -1) of the outputs for each T_i, the best T is selected. This step is illustrated in Figure 4.4 in Step D.

Step 4 Convergence: The whole process (Steps A, B and C) is repeated 4 times and outputs (clusters) are compared. Thus, we have four clusterings output. The Rand Index is calculated for each pair of outputs (6 pairs) to estimate the concordance between different clusterings. If the average Rand Index is >0.99, this means that there are no significant differences in the outputs of 4 runs, the algorithm is said to be converged, and it stops. Otherwise, the output with the lowest average Silhouette width is discarded, and steps A to D are repeated.

When PFClust was compared with other well-known clustering algorithms (including hierarchical, partition, and density based clustering algorithms) using a synthetic data set, for more detail see [1], it outperformed other algorithms. We have explained various validation measures used in this algorithm in the following section.

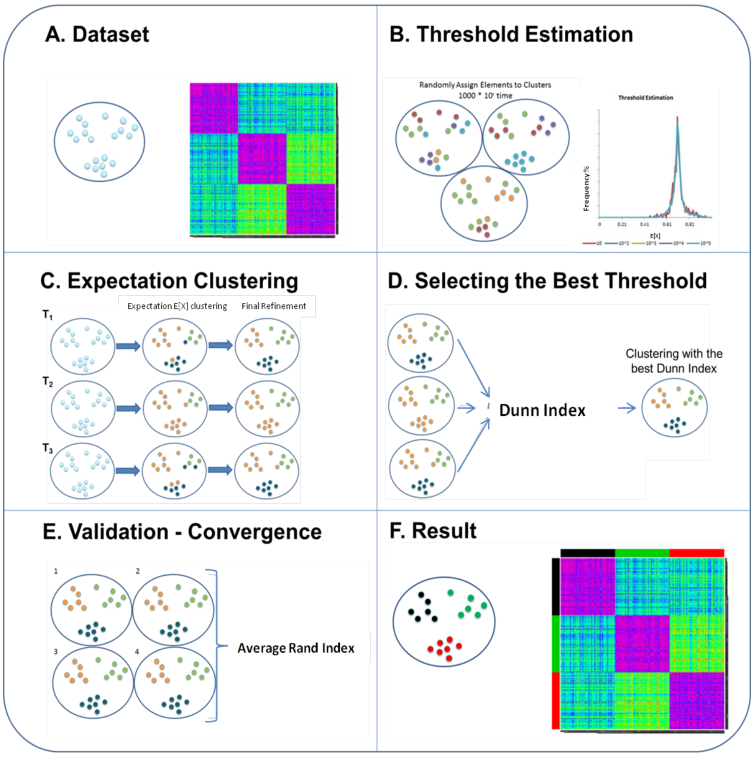


Figure 4.4: This figure illustrates the work-flow of PFClust. Each step is explained in the main text.

Performance of PFClust Performance evaluation: The performance of PFClust was done on the following configuration:

- Hardware: 2.2 GHz Intel(R) Core(TM) i5-3470S CPU @ 2.90 GHz, 8.00 GB RAM
- Operating system: Scientific Linux release 6.3 (Carbon)
- JVM: 1.6.0_45-b06

The time difference was significantly lower for larger dataset (size $>$ 1000 data point) from 35000 seconds to 10 seconds [144].

Limitations of clustering methods Overall, three major limitations in the above mentioned clustering algorithms are: first defining the number of clusters k, second input parameters, and last validating the results. There is no straightforward 'best' way to evaluate clustering methods, as the results are dependent on the dataset provided. Different techniques often highlight different patterns in the data, so complementary methods may be helpful in analysing a single data set. This also makes interpretation of the results harder. To evaluate the results, many authors combine different evaluation measures, discussed later in this chapter, to get a clearer interpretation of the results [135, 136].

In the next section, we will discuss the work-flow of evaluating results from different clustering algorithms using 'clValid' [145] and 'fpc' [146] packages in R. Moreover, we also discuss results from PFClust.

4.5 Evaluating Clustering Solutions

There is more than one definition of clusters depending on the dataset used for that particular study [147]. In fact, most authors define different grouping criteria to cluster an item, for example, a cluster or group is formed based on the principle of minimum distance between two items or by maximum separation of clusters. Today, with well known clustering criteria, we need a measure that validates the output.

Validity is a certain amount of confidence that is added to the patterns recognised by the cluster algorithms [135, 147]. Validation also serves an important implication on the problems or limitations discussed in the previous section, by defining the number of clusters or to optimise the parameters [148]. Broadly, the validation methods are grouped into external and internal validation. The fundamental difference between the two types of validation method is that the external validation method uses some reference classification method whereas there is no external label required in internal validation methods. Both of these methods are equally important [149]. In PFClust, we have used both validation criteria.

External Validation: Standard external validation measures take gold - standard class labels and compare with the labels provided by the cluster algorithm *via* contingency table of the pairwise assignment of data items. Probably the best known index is the Rand Index (Rand, 1971) (Equation 4.2), following the simple criteria of comparing gold-standard class labels with labels provided. The Rand Index is defined as:

$$RAND = \frac{a+b}{a+b+c+d} \tag{4.2}$$

Where a is the number of pairs of instances that are assigned to the same cluster in clustering (C_1) and to the same cluster in clustering (C_2); b is the number of pairs of instances that are in the same cluster in C_1, but not in the same cluster in C_2; c is the number of pairs of instances that are in the same cluster in C_2, but not in the same cluster in C_1; and d is the number of pairs of instances that are assigned to different clusters in C_1 and C_2.

Internal Validation: In contrast to external validation, internal validation evaluates the intrinsic quality of the cluster. The qualities we are interested in here are *compactness*, *connectedness* and *separation* of the cluster. Here, *compactness* suggests finding homogeneity of intra-cluster variance, *connectedness* provides the degree of partitioning observed local densities and groups data items together with their nearest neighbours in the data, and *separation* includes the measure to quantify the degree of separation between the individual clusters. All these aspects hold important places in internal validation separately as well as with some combinations. The most popular combination is between *compactness* and *separation.* Several techniques therefore assess both intra-cluster homogeneity and inter-cluster separation such as Silhouette width [150] and Dunn Index [151].

The Silhouette width is a useful measure when one is seeking compact and clear separation between clusters. Once the data is clustered the distance within and between clusters is quantified with respect to each object i. Suppose object i belongs to cluster A then average dissimilarity of object i is computed to other members of the same cluster, this is assigned to a_i. Next, the average dissimilarity of i to all objects that are different from cluster A is computed. Then, the minimum distance from i to objects not belonging to cluster A is recorded in b_i, which is also known as neighbour of object i. Note that the construction of b_i depends on the other clusters, so it is an underlying assumption that there are more than one clusters. The number of s_i is obtained as following Equation 4.3.

$$s_i = \frac{b_i - a_i}{max(a_i, b_i)} \tag{4.3}$$

Another method to assess the *compactness* and *separation* of the cluster output is Dunn Index. Dunn Index is defined in Equation 4.4.

$$D_c = \min_{C_k \in C} \left(\min_{C_l \in C} \frac{dist(C_k, C_l)}{max_{(C_m \in C)} diam(C_m)} \right) \tag{4.4}$$

where $diam(C_m)$ is the maximum intra-cluster distance within cluster C_m and $dist(C_k, c_l)$ is the minimal distance between pairs of data items i and j with $i \in C_k$ and $j \in C_l$. Higher Dunn Index is better for a given assignment of clusters. One of the limitations is that this method is computationally costly. The Dunn Index measures the ratio between the smallest cluster distance and the largest intra-cluster distance in a partitioning.

Statistical analysis: In order to find association between the clusters and biologically relevant information we use residual test statistics, which performs cell-to-cell comparisons within the cluster. For this test, the first step is to summarise the data into a contingency table, to get better insight into the clustering results. In this matrix, column is represented by the label for each cluster (arbitrary number was assigned to cluster in order to keep track) and row represents the counts of mechanistic annotations defined in MACiE. The χ^2 test [148, 152] of the mechanism annotations within each cluster suggests a significant association between the annotations and the clusters at $P < 0.001$.

Further, we examine the patterns of association within observed frequency of mechanistic annotation using residual test statistics to determine strong or weak association. For this, we arranged data into contingency table, mentioned earlier, and calculated how much deviation of observed frequency from the expected by using residual analysis [153, 154]; Equation 4.5.

$$r = \frac{(Observed - Expected)}{\sqrt{Expected}} \tag{4.5}$$

where $Expected = {}^{(n_i - n_j)}/_{n}$. $Expected$ is the expected frequency count between mechanistic step type i and clusters j. n_i is the total number of sample observations at cluster i of Variable A: mechanistic step type; n_j is the total number of sample observations at cluster j of Variable B: Clusters labels, and n is the total sample size. The results of this test are represented as a heatmap in the next section.

4.6 Result

There is no straightforward way to determine number of clusters. However, a common way to deal with such a situation is to compare different algorithms using various validation methods hoping to get common results. Using such a strategy and results from PFClust, we discuss the result in three different sections: first in Part A, the comparative analysis of various clustering algorithms. Second in Part B, we discuss the results from PFClust algorithm. And third in Part C, we describe extending the results of PFClust to EC top class.

In order to test PFClust, we used a number of synthetic 2D datasets and the 224 protein domains in 11 CATH superfamilies. To show how well the PFClust algorithm applies to biological questions, we discuss results of 224 protein CATH domains only. In order to validate PFClust results, we compared results from PFClust with six other current state-of-the-art algorithms. These are (i) the hierarchical clustering algorithm Hierarchy [155]; (ii) the hierarchical AGglomerative NESting (Agnes), (iii) the partitional k-means clustering algorithm [133], (iv) Clustering Large

Applications (Clara) [156], which is based on repeated k-means clustering of samples, (v) Density-Based Algorithm for Discovering Clusters in Large Spatial Databases (DBSCAN) [157] and (vi) Model-Based Clustering (Mix Model) [158].

Direct comparison of results (k), where k is the number of clusters, for all these clustering algorithms is not possible. Only PFClust and DBSCAN amongst the methods considered here can determine the clusters, and in fact the latter algorithm requires two parameters to be optimised before it decides the number of clusters. Hence, for the five other methods, we use the externally defined 'correct' number of clusters (this definition includes singletons in the count of clusters) as a given parameter and compare how well each algorithm clusters the data compared to the original classification. In order to compare the different clustering approaches, we select the Rand Index as a measure of agreement between the externally known 'correct' clustering and that produced by the clustering algorithms.

The agreement between the PFClust and CATH classification is nearly perfect with a Rand Index of 0.996. There is only a minor difference between the original classification and the classification of PFClust, where the 1q27A00 protein domain is classified as a singleton by PFClust, whereas CATH has it assigned to the 3.90.79.10 (Nucleoside Triphosphate Pyrophosphohydrolase) superfamily. We also test the other clustering algorithms against this dataset and set the number of clusters to 11 for the five algorithms requiring this parameter. Figure 4.5 visually illustrates the agreements and disagreements between the different clustering algorithms. We see that Mix Model and Clara are the top performing clustering algorithms, reproducing the exact CATH classification.

For each of the aforementioned algorithms, the Silhouette width was used as the criterion for identifying the number of clusters similarly to the way we ran DBSCAN. Since all the algorithms depended on a single parameter k (number of clusters, inclusive of singletons), we varied this number from 2 to 50 and the results are shown in Figure 4.6. Note that these data are considered separately and do not contribute to the main results described previously, for which purpose the 'correct' number of clusters was instead passed to Hierarchy, Agnes, Clara, k-means and Mix Model as a parameter.

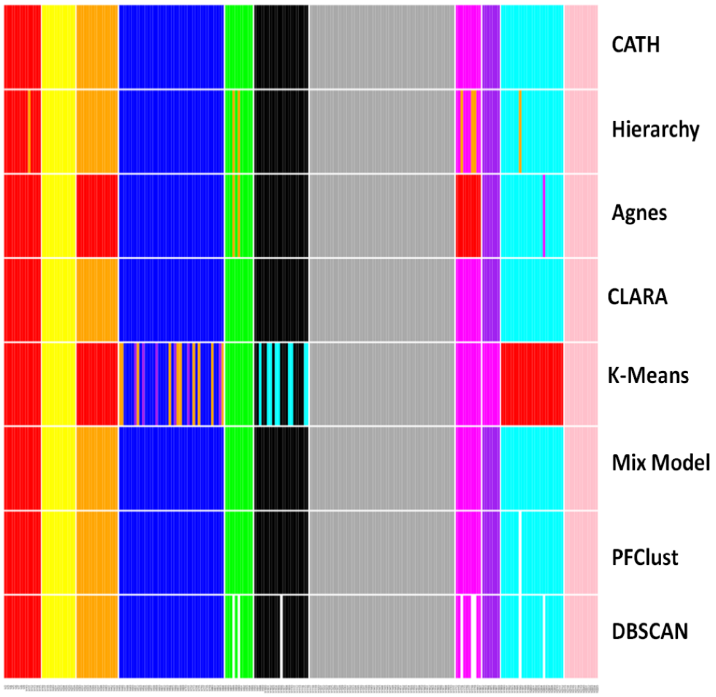


Figure 4.5: The results of each of the different clustering algorithms as coloured lines. Heat map of protein domain to protein domain density similarities. On the row side, the protein domains are coloured according to the CATH classification; on the column side, the protein domains are coloured according to PFClust.

4.6.1 Part A

To address the difficulty of determining the number of clusters, we designed a workflow which includes results from various clustering algorithms, illustrated in Figure 4.7, for better interpretation of the clustering of enzyme reactions. In this workflow, first the enzymatic reaction (*OBC* and *CBC*) [106] is calculated into similarity matrix using the Tanimoto coefficient. Next, using several well known validation measures we solved two criteria: first, to find the number of clusters and second, to optimise the required parameters. Ideally, if all the cluster algorithms agreed on single results, that could be the best possible answer [135]. Using several well known evaluation indices, we optimized the clustering outputs for each of the algorithms with equivocal results (Figure 4.8) that suggested the existence of between two and

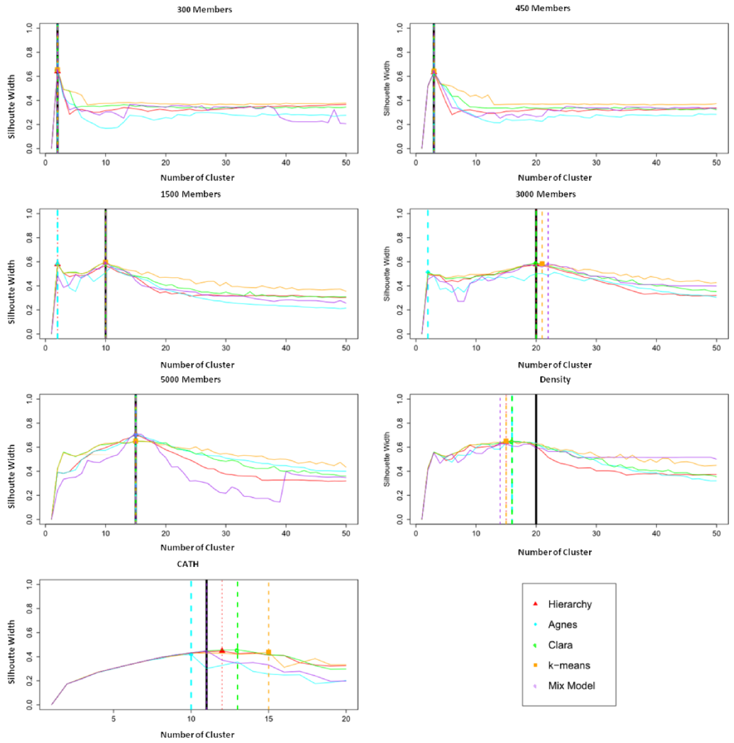


Figure 4.6: As an addendum to the main work, we test the use of the Silhouette width as a characteristic measure from which to decide the correct number of clusters. We ran the deterministic methods once each. We also ran the stochastic Clara and k-means algorithms 100 times each for every number of clusters, *k*, between 2 and 50. The run with the best Silhouette width for a given algorithm was selected, thus deciding the number of clusters to report.

over a hundred clusters.

4.6.2 Part B

Here, in part B we have analysed results generated from PFClust. We feel that this algorithm provides reasonable results without any additional information being required. We understand that our results are limited by the annotations available in the databases. For *OBC* descriptor, PFClust produced 39 clusters and 57 singletons. We found that there are 8 MACiE enzymes that were found to be singletons in both of the datasets; those are: M0128 (photinus-luciferin 4-monooxygenase, EC: 1.13.12.7), M0140

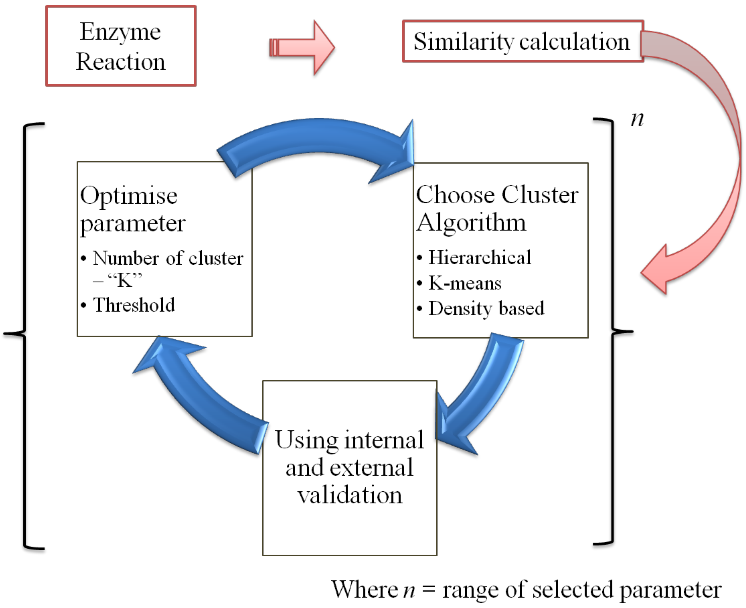


Figure 4.7: This figure illustrates the workflow designed to determine the number of clusters using various state-of-the-art algorithms.

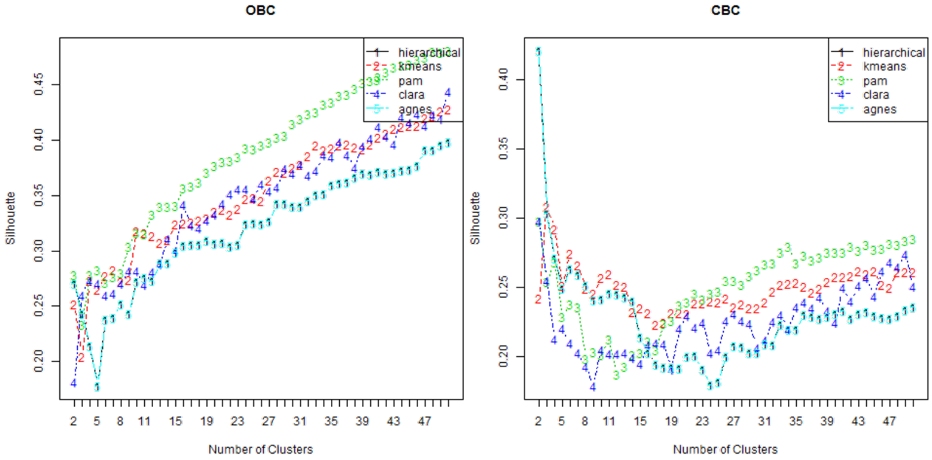


Figure 4.8: This figure illustrates the performance of different clustering algorithms using Silhouette width. Here, x-axis represents Silhouette width and y-axis is the range of number of clusters.

Table 4.2: The table summarizes the comparison between PFClust and the other six clustering algorithms based on the Rand Index between the clustering predicted by the method in question and the original gold standard clusters.

Data Set	300	450	1500	3000	5000	CATH
Hierarchy	0.88	0.94	0.89	0.920	0.981	0.964
Agnes	0.94	0.97	0.83	0.820	0.976	0.906
Clara	0.96	0.98	0.952	0.948	0.987	1.000
K-means	0.96	0.98	0.958	0.966	0.986	0.738
Mix Model	0.960	0.98	0.959	0.911	0.990	1.000
PFClust	0.96	1.00	0.958	0.949	0.986	0.996
DBSCAN	0.97	0.97	0.930	0.92	0.978	0.977

(ribonucleoside-triphosphate reductase, EC: 1.17.4.2), M0145 (isopenicillin-N synthase, EC: 1.21.3.1), M0207 (pyruvate, phosphate dikinase, EC: 2.7.9.1), M0212 (nitrogenase, EC: 1.18.6.1), M0286 (O - phospho - L - seryl - tRNASec : L - selenocysteinyl - tRNA synthase, EC: 2.9.1.2), M0294 (succinate dehydrogenase (ubiquinone), EC: 1.3.5.1) and M0297 (alkylmercury lyase, EC: 4.99.1.2). There were 13 clusters (Figure 4.11) produced when the data were clustered using *CBC* descriptors. For further investigation, we created a contingency table with enzyme function annotation "mechanism annotation" in a given cluster. Applying χ^2 test, we concluded that there is an association between the mechanistic annotations and clusters of enzymes at $p < 0.001$. Further "mechanism profile" is created using standardized residual analysis for further analysis. These results show fewer clusters when clustered with *CBC* than *OBC* descriptors, suggesting that enzymes often use different mechanistic steps to perform similar functions. These two distinct patterns observed in this analysis suggest that using the EC classification system is not sufficient to annotate enzyme function. We aim to find conserved patterns of mechanistic steps that lead the enzymes to perform different functions.

Further, we looked for the association of 'mechanistic annotation' and various other information related with enzyme function such as KEGG [129], QuickGO [128], Metal MACiE [24] within a cluster group of enzymes (Figure 4.9), in order to get better insight of the cluster patter we found. The "mechanism profile" retrieves strong signals as shown in heat map (Figure 4.10 and 4.11) where colour yellow represents the high (>2) association of the “mechanisms” within the clusters.

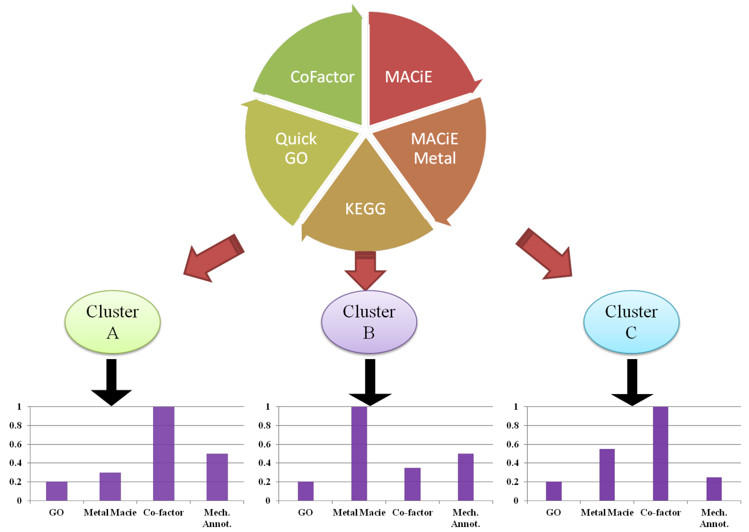


Figure 4.9: This figure illustrates the data mining for analysing the output of PFClust. For this work, we retrieved data from various resources such as KEGG, Metal MACiE, Quick GO, CoFactor to associate, and determined the important features of enzyme reaction in each clusters.

4.6.3 Propensities of EC Classes to Cluster Together

Here, we show the pairs of reactions grouped together in the cluster with respect to the EC top class. For this, first we calculated the pairs of reactions to be grouped together out of $\frac{1}{2} \times (320) \times (320 - 1) = 51,040$ pairs of reactions in the dataset (Figure 4.12). Figure 4.12 illustrates the propensities for the different EC classes, with the *OBC* and *CBC* in different shades of red. To calculate propensities for different EC classes we divide each of these class proportions by the overall value. Where the values are above 1, this indicate that pairs of top level EC class members are more likely to cluster together than are randomly chosen mechanisms. We can see from Table 4.3 and Table 4.4 that 29% of all possible pairs of hydrolases are clustered together when clustered using *OBC* dataset and 60% when clustered by *CBC* descriptors. This example shows that there are more enzymes that use similar enzymatic reaction steps to complete the reaction as a hydrolase (EC 3). Moreover, for ligases (EC 6) it was suggested that 30% of possible pairs were clustered together when clustered according to *OBC* descriptors,

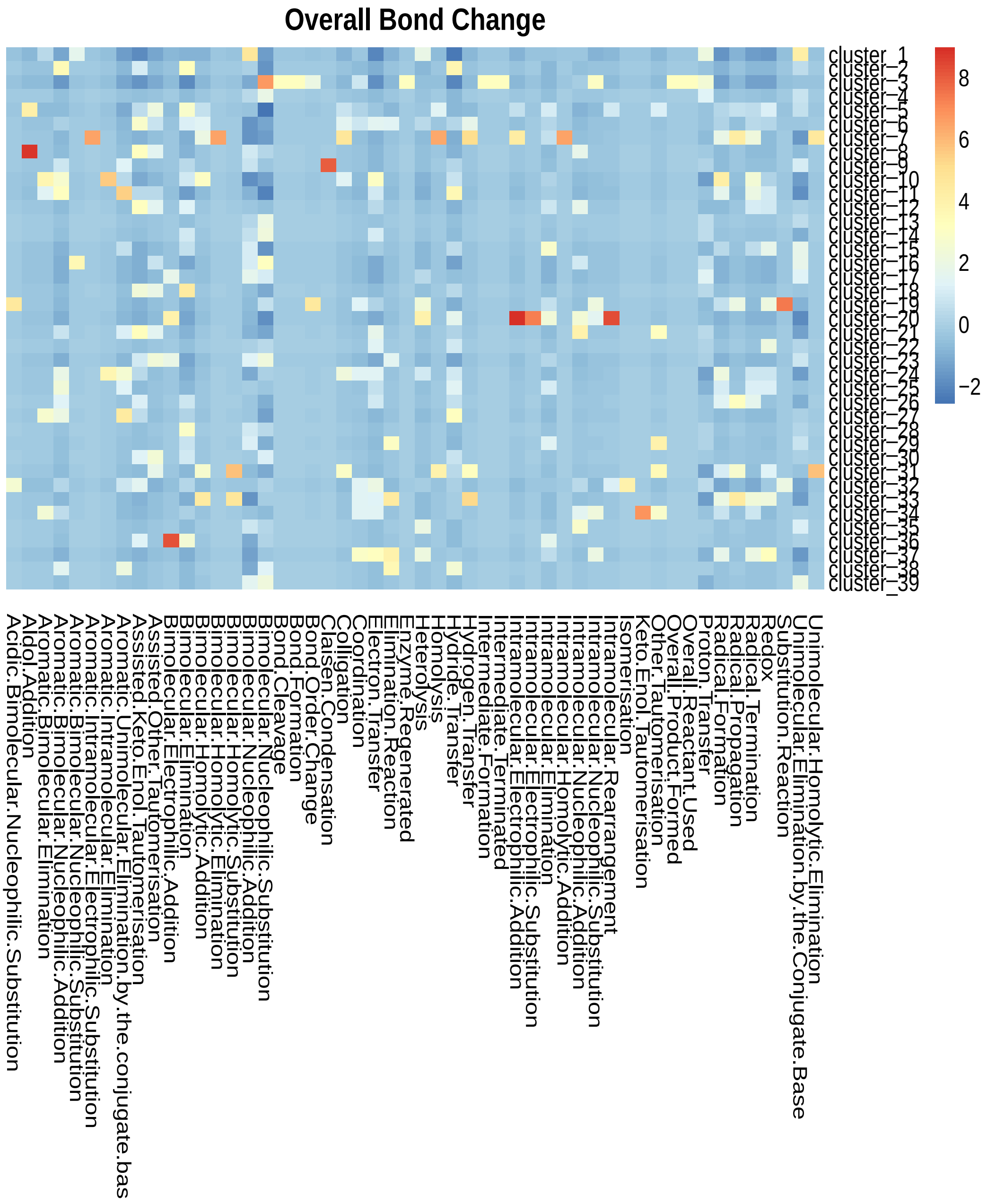


Figure 4.10: Heat map representing the clusters (y-axis) of *OBC* descriptors and their association with mechanistic annotations (x-axis). The colour key represents the strength of the signals found in each cluster where red means stronger association and blue means weak.

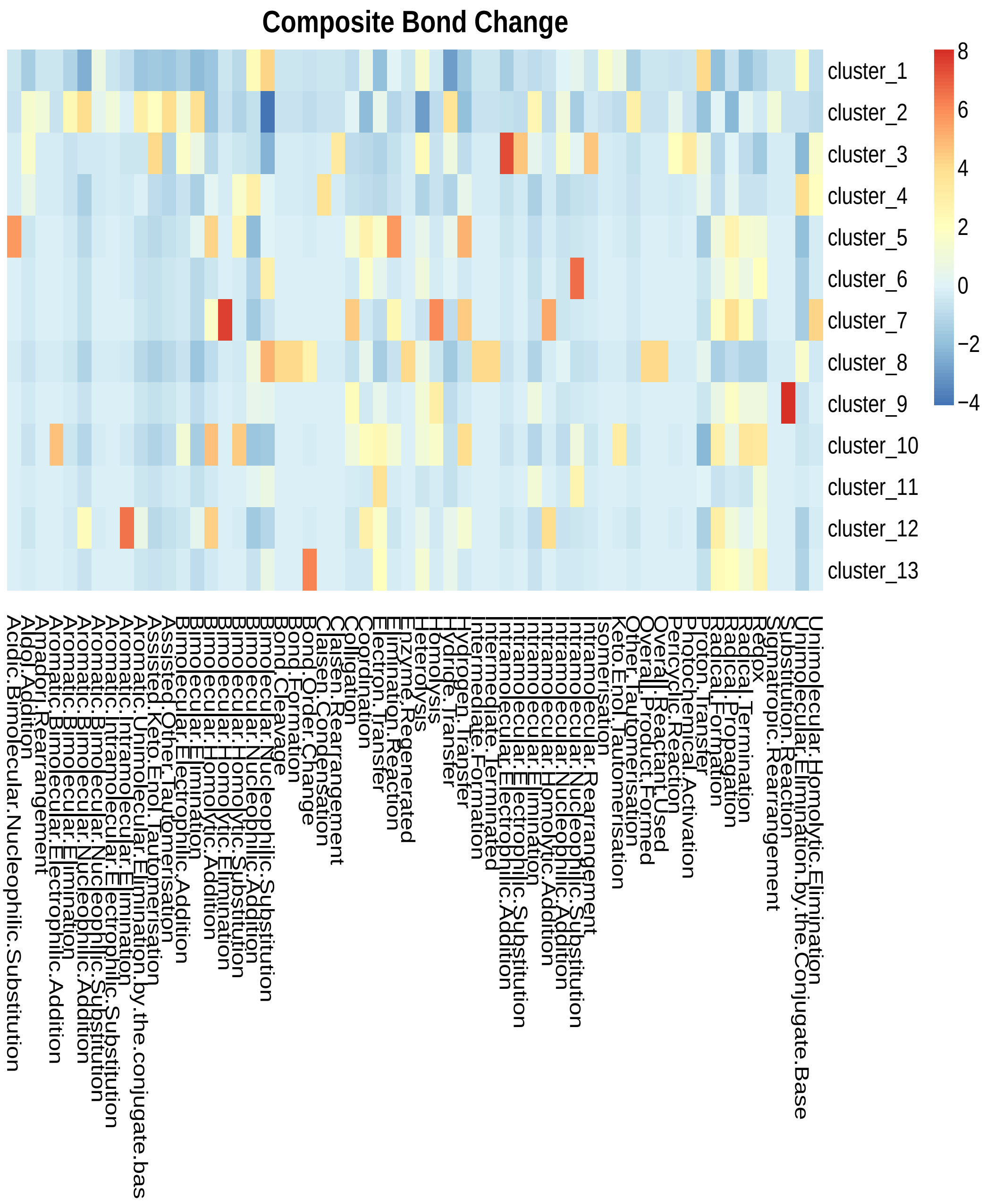


Figure 4.11: Heat map representing the clusters (y-axis) of *CBC* descriptors and their association with mechanistic annotations. The colour key represents the strength of the signals found in each cluster where red means stronger association and blue means weak.

whereas when clustered according to *CBC* we noticed that 40% of all possible pairs were grouped together. Overall we noticed *CBC* has a greater proportion of enzymes grouped together than *OBC*.

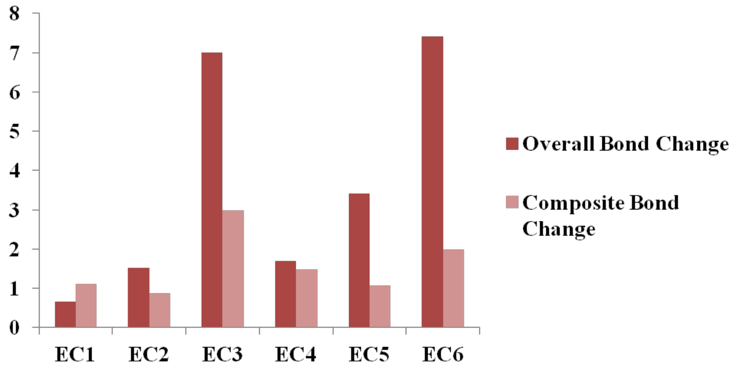


Figure 4.12: Bar plot representing the proportion of enzyme *OBC* and enzyme mechanisms pairs *CBC* grouped together in the same cluster for each EC top class, ECx (x = 1 to 6).

Table 4.3: When clustered with *OBC*: proportion and propensity of the pairs of enzymes, according to the top EC class, grouped in PFClust

OBC	EC1	EC2	EC3	EC4	EC5	EC6	All same EC
Proportion	0.027	0.063	0.292	0.070	0.142	0.309	0.121
Propensity	0.667	1.522	7.006	1.692	3.416	7.420	2.090

Table 4.4: When clustered using *CBC* descriptors this table shows the proportion and a propensity for each EC top class.

CBC	EC1	EC2	EC3	EC4	EC5	EC6	All same EC
Proportion	0.224	0.278	0.603	0.301	0.218	0.404	0.328
Propensity	1.104	0.88	2.974	1.483	1.076	1.076	1.620

4.7 Two Case Studies

It is difficult to provide a discussion of the complete results therefore here, we discuss two case studies, and further information for complete discussion of the results is available in the appendix A1.

Case 1: In *OBC* clustering these two enzymes, M0033 (methylmalonyl - CoA epimerase EC 5.1.99.1) and M0070 (methylmalonyl-CoA decarboxylase, EC 4.1.1.41), are assigned to different clusters whereas according to *CBC* clustering they are grouped together. Another MACiE enzyme M0182 (methylisocitrate lyase, EC: 4.1.3.30) is grouped together with M0070 when clustered with *OBC* and these three enzymes are grouped together when clustered with *CBC*. These enzymes share common reaction elements i.e. order change from single bond to double bond between carbon and oxygen and a bond formation between carbon and hydrogen. All these three enzymes participate in a similar pathway (see Figure 4.13): ec00640 Propanoate metabolism[2] also known as propinoate metabolism. The metabolism of propionic acid (propanoate) begins with its conversion to propinoyl coenzyme A [159]. Studies concerning the role of enzymes involved in pathways have gathered special focus for *Mycobacterium tuberculosis* [160]. It is important to understand propinoate metabolism as any change in propinoate pathway could lead to accumulation of toxic metabolites [160].

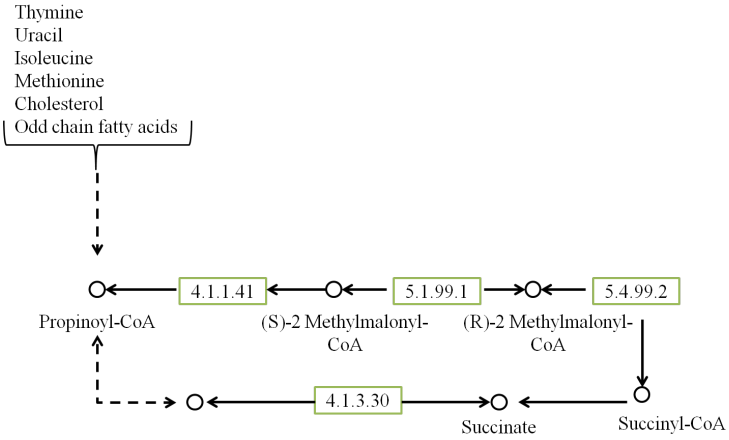


Figure 4.13: This figure illustrates the precursors of pathways for propionate. A number of clinical disorders arise from errors at various steps in these pathways. Here, broken arrows indicate the presence of several reactions to complete the pathways.

Case 2: Another interesting *OBC* cluster of enzymes: M0123 (adenylylsulfate reductase; EC Number: 1.8.99.2) and M0279 (phosphoadenylyl-

[2] KEGG: http://www.genome.jp/kegg-bin/show$_{p}athway?ec$00640

sulfate reductase (thioredoxin) EC Number: 1.8.4.8) participating in sulfur metabolism ec00920 [3]. When clustered by *CBC*, we found that M0279 is clustered with another group of enzymes which included M0153 that also participates in similar pathways. Whereas, M0123 was treated as a singleton for *CBC*. Sulfur is an essential element for life and the metabolism of organic sulfur compounds plays an important role in the global sulfur cycle. Sulfate reduction can both occur as an energy consuming assimilatory pathway, participated by M0279 and an energy producing dissimilatory pathway, participated by M0123 (see Figure 4.14). Assimilatory pathway is more commonly found in different organisms, producing reduced sulfur compounds for the biosynthesis of S-containing amino acid, while in dissimilatory pathway, which is restricted to anaerobic bacteria, sulfate is the terminal electron acceptor of the respiratory chain producing large quantities of inorganic sulfide. It was interesting but not surprising to find that the MACiE entries M0123 and M0279 share similar overall reaction entities: those are oxygen and hydrogen bond formation, sulfur and oxygen bond formation and sulfur-oxygen bond order change.

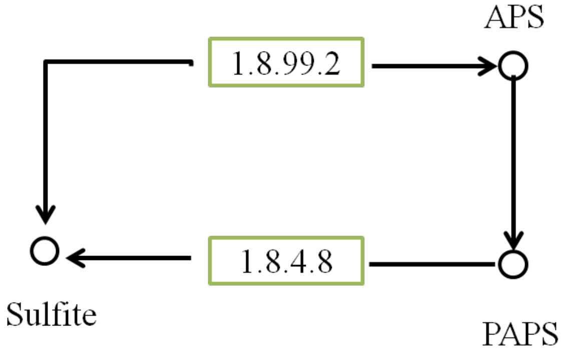


Figure 4.14: This figure illustrates precursors of pathways for sulfur metabolism. This illustrates that the enzymes participating in assimilatory pathway and dissimilatory pathways are grouped together as overall they perform similar reactions using different step types.

[3] KEGG: http://www.genome.jp/kegg-bin/show$_{p}athway?ec$00920

4.8 Summary

In this chapter, we looked into the diversity of enzyme chemical reactions where chemical reactions were defined for overall as well as for mechanistic reactions. Ambiguous results from various clustering algorithms led to the design of an in-house clustering algorithm, PFClust, that outperformed other clustering algorithms in recognising the data structure.

We have performed a clustering analysis of enzyme reactions described first by *OBC* descriptors, and second by *CBC* descriptors. We find that the *CBC* descriptors cluster the data into significantly fewer clusters than *OBC* descriptors, suggesting that different functions tend to share similar mechanisms. Moreover, at a finer level of enzyme classification, we have also observed that enzymes often use different mechanistic steps to perform similar functions.

Our results suggest that, in spite of the simplicity of PFClust, our method was able to capture the important features of an enzyme reaction. Our study shows an interesting diversity of reaction clusterings, where every clustering suggested different factors that are similar in the clusters, such as metal cofactor, mechanism annotation etc. The distinct patterns observed in this analysis suggest that using the EC classification system is not sufficient to annotate enzyme function. In future, we aim to find conserved patterns of mechanistic steps that lead the enzymes to perform different functions and can create a library. By screening a library of enzyme variants one can discover variants with greatly improved activities for various cyclization reactions [161].

5

Prediction of Enzymatic Function

5.1 Introduction

One of the many challenges in post-structural genomic functional elucidation is to design computational methods to annotate function of uncharacterised proteins (or unannotated proteins), especially enzymes. Numerous informatics groups have addressed this issue using protein sequence [23, 85, 162] and structural information to understand and to predict EC number[1], successfully providing up-to 97% correct prediction [23,163]. Some have also tried predicting enzyme function by defining overall transformation and not considering mechanism of enzymatic activity, also ignoring protein structure and function [35, 36, 164].

Our motivation is to investigate the relationship between the reaction mechanism as described in the MACiE database and the main top-level class of the EC classification. In order to do this, we generate supervised machine learning models to predict EC class from data on the chemical reaction or its mechanism.We consider two ways of encoding the mechanistic information in descriptors, and also three approaches that encode only the overall chemical reaction.

We compare enzyme mechanistic descriptors derived from the MACiE

[1] Note that EC number will be used interchangeably with EC class

database and use multivariate statistical analysis for assessment of enzyme classification. We investigate the relationship between reaction mechanism as described in the MACiE [98, 106] database and top class of EC number.

We evaluate 260 well annotated chemical reaction mechanisms of enzymes using machine learning methods, placing them into the six top level EC classes retrieved from MACiE V 2.4. Moreover, we compare the classification performances of three supervised learning techniques, Support Vector Machine (SVM) [Vapnik, 1998], Random Forest (RF) [Breiman, 2001] and K Nearest Neighbour (kNN), for the reaction mechanism classification task using five different descriptor sets from MACiE data. In this chapter, we discuss the machine learning algorithms that have been used to address problems like this.

5.2 Introduction: Enzyme Function Prediction

Many informatics groups have attempted to predict EC number by using different biological information such as sequence, structure, or catalytic residue. Sometimes these features are used independently and sometimes a combination of these featured are used in order to get a better understanding. In this section, we discuss features that are used for predicting function.

5.2.1 Using Sequence

A well designed kNN model (with $k = 1$) [23] was used for predicting enzyme function via InterPro [22] signatures. This is very local prediction which basically means annotating function to the closest neighbour, but the coverage of this model is global. This model successfully provided 97% accuracy and this result suggests that this method can predict all EC classes using sequence features. The success of this method is broadly based on the homologous enzymes. However, there are many examples in biology suggesting that homologous enzymes do not necessarily catalyse the same reaction [165]. Hence, using sequence information alone to assign function should be viewed with some scepticism [166].

5.2.2 Using Structure

A study by Dobson & Doig [85] used various structural properties, such as secondary structure properties, amino acid propensity, surface properties and ligands, to predict enzymatic function using SVM model with 35% accuracy. Even though the structures are much more conserved than sequence, these figures suggest the difficulty of predicting function using structure properties. Another study [162], using a combination of sequence and structure, was able to predict 33% correct EC class prediction. Functional sequence properties are very valuable, but, when the 3D information is added it makes the picture much clearer.

5.2.3 Using Overall Chemical Transformation

Using biochemical transformation patterns of a given reactant pair [164], we can access features of enzymatic reaction such as substrate and product, creating an 'RDM pattern', in order to annotate enzymatic function. Here, 'RDM pattern' represents the chemical transformation of enzyme catalysis, more detail is present in [164]. In another study by [34], the bonds reacting in the substrates of enzymatic reactions catalysed by the EC 3 hydrolase enzyme family were used to create physicochemical descriptors, which were found to correlate well with EC number. Representation of enzyme catalysis can also be done using the MOLMAP descriptors [35], which defines types of covalent bonds based on physicochemical and topological properties, and correctly assigns with 95% accuracy at the EC class level. Also, the study in [36] used reaction difference fingerprints (RDF) to map reactions to EC class, giving 83% correct predictions. These figures support the use of reaction descriptors for mapping activity onto enzymes. These studies suggest that a good design of descriptor could lead to an accurate prediction of function.

5.3 Method

To investigate the relationship between an enzymatic reaction, as defined in MACiE [98, 106], and the EC top class we used state-of-the-art machine learning algorithms such as Support Vector Machine (SVM), Random Forest (RF) and kNN (k Nearest Neighbour). Studies by [33] and [20] inspired our motive to investigate the extent to which prediction of EC number is possible. By definition, quantitatively representing overall reaction from starting material to the product would lead to an accurate assignment of EC number. Here, we have encoded the enzymatic activity into descriptors at two levels, first at overall reaction and second at granular steps of reaction.

Figure 5.1 illustrates the complete work-flow starting from data preprocessing till the evaluation of the final model. We have extended this work-flow for regression analysis as well [3][2]. Cheminformatics models have been used for compound property prediction, particularly in pharmaceutical industries. In work by [3], apart from other machine learning methods we incorporated Partial Least Square(PLS). Here, we are particularly interested in predicting solubility of drug-like molecules. Solubility of the drug candidate is important as it reflects the bioavailability, especially with oral drugs.

In the following section, we will describe each step in the work-flow (Figure 5.1) starting from data culling till validating the data. We will also explain briefly regarding classification and regression algorithms, and respective results are in the last section.

5.3.1 Data Culling

For Enzyme Function Prediction: For this, we retrieved data from MACiE, where enzymes are limited to 260 entries from the MACiE database V 2.4 adding further 60 entries in V 3.0 [2]. Basically, the enzymatic activity is encoded numerically into five sets of descriptors that illustrate at two levels the enzymatic reaction; first the overall reaction and second, mechanistic step types. These descriptors are discussed in Chapter 3.

For Predicting Solubility of Drug-Like molecules: For this analysis [3], we retrieved data from work by Llinas et al., [167] where solubilities of 122 compounds are reported from the CheqSol method. The Solubility Challenge dataset [168] was used as a benchmark dataset. We trained our model with canonical training:test split of 94 being in training and 28 molecules in test set. The SMILES were culled from various resources such as ChemSpider [169] and some data directly from database [168]. As a benchmark, we present our method's predictions of the solubility challenge set based solely on chemoinformatics descriptors. The datasets are available in the supporting information: see Appendix B for more detail.

[2] NOTE: my contribution to this paper is to execute work-flow (Figure 5.1) for the descriptors quantified by James McDonagh

5.3.2 Step 1: Data Preparation

It is important to pre-process the data to improve the performance of the model. This step conditions the quality of the patterns discovered and assists in clear interpretation. Also, to avoid over-fitting of the data one would like to pre-process data before analysing it [170]. There are various ways to pre-process the data and in our work we used two normalising methods. First, commonly used standardised method of variable scaling or z-score normalization Equation 5.1. This will equalise the priority of all the attributes. Second, PCA scaling method [171], specifically for regression analysis in [3], which is widely used in many studies to scale the data. This method transforms the data into smaller subsets, which reduces the correlation of the data. To compare the difference between two scaling methods we have also performed our analysis on the raw dataset.

$$z = \frac{x - \mu}{\sigma} \tag{5.1}$$

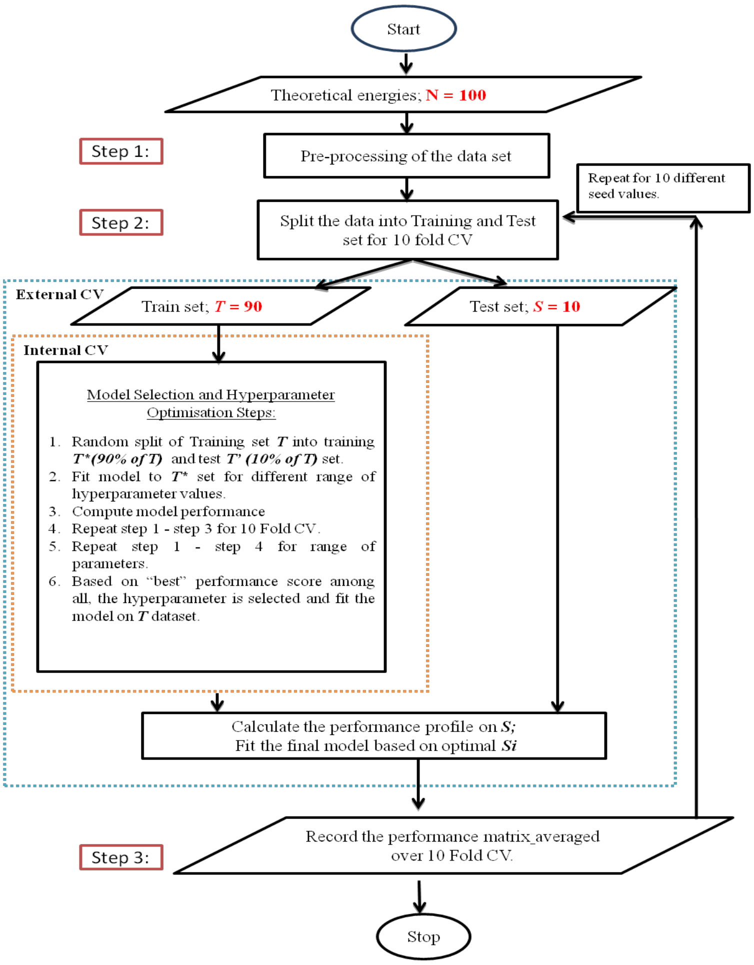
where: x is the dataset mentioned earlier [2] and [3]; μ is the original mean of the population; the mean is set to zero for normalization, σ is the standard deviation of the population.

In study [2], only three out of five datasets were scaled; those are *CBC*, *OBC* and *HD* descriptors. The other datasets, *MS* and *OS* similarity, are not scaled as these descriptors are already scaled in range of 0-1.

5.3.3 Step 2: Internal and External N-Fold Cross-Validation

It is a common practice to evaluate the performance of different machine learning methods by cross-validation methods such as bootstrapping, Leave-One-Out-CV (LOOCV) and N-fold cross-validation [172]. Cross-validation methods are validation techniques to generalise the model for various independent datasets. For this, the dataset is split into a training and test set, where the training set (seen dataset) is used to build model or tune parameter (internal cross validation measure) and the test set (unseen dataset) is fitted to the final model with constant parameters (external cross-validation).

The 10-fold cross-validation has been widely accepted as a reliable method for calculating generalization accuracy, and experiments have shown that cross-validation is relatively unbiased [172,173]. The design of 10-fold cross-


Start
Theoretical energies; N = 100
Step 1:
Pre-processing of the data set
Repeat for 10 different seed values.
Step 2:
Split the data into Training and Test set for 10 fold CV
External CV
Train set; T = 90
Test set; S = 10
Internal CV
Calculate the performance profile on S; Fit the final model based on optimal Si
Step 3:
Record the performance matrix averaged over 10 Fold CV.
Stop

Model Selection and Hyperparameter Optimisation Steps:

1. Random split of Training set *T* into training ***T*(90% of T)*** and test ***T' (10% of T)*** set.
2. Fit model to ***T**** set for different range of hyperparameter values.
3. Compute model performance
4. Repeat step 1 - step 3 for 10 Fold CV.
5. Repeat step 1 - step 4 for range of parameters.
6. Based on "best" performance score among all, the hyperparameter is selected and fit the model on *T* dataset.

Figure 5.1: This illustrates the work-flow for both classification and regression problem. The complete work-flow is divided into 3 major steps, which is explained in the main text.

validation is illustrated in Figure 5.2, where 10% of the dataset is considered to be a test set which is kept aside to fit with the final model. The cross-validation method runs 10 times where each test set is used exactly once to fit the model. The final estimated result is averaged over 10 folds. At Step 2, in Figure 5.1, the 10-Fold Cross-validation is executed as an internal (red dotted box) and external (blue dotted box) cross-validation.

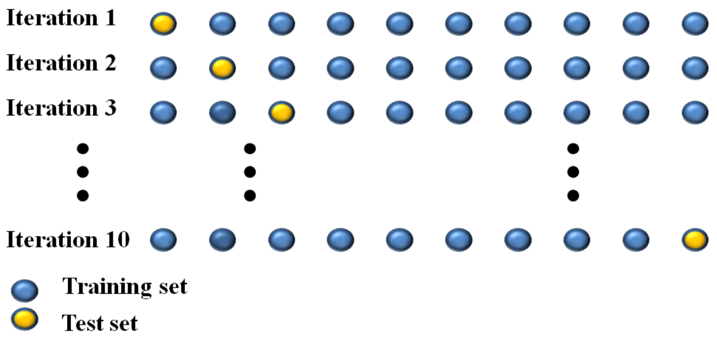


Figure 5.2: Cross validation: this figure illustrates an example of 10-fold cross-validation where blue circle represents the training set and yellow test set. In each iteration one 10% of the dataset is considered to be a test set which is later used to validate the model fitting. In our analysis we have used this validation measure twice an in internal validation and external validation measure.

Internal 10-Fold Cross-Validation for parameter selection: In step one, the data is split into test and training set (90% of the original dataset). The training portion of 90% of the original data is further split into 10 new folds of 9%, with nine (81% of the original data) being used to build each model and one (9%) as an internal validation; this process of model building and internal validation is repeated to predict each of the 10 possible internal validation folds. This internal cross-validation step is repeated 20 times, once for each possible value of the parameter being assessed. Next, based on the 'best' scores (accuracy Equation 5.2 for [2] or RMSE Equation 5.3 for [3]) in the internal validation folds, the optimum parameter is selected. Finally, the model is fitted on the complete training set of 90% of the original data using the selected parameter values.

External 10-Fold Cross-Validation: The given 90%:10% split of the data into training and test sets was used to fit the final model for each fold of the main 10-fold cross-validation, once the optimum parameter values have been selected. The average accuracy or RMSE values over the 10 folds were considered in order to compare the usefulness of different descriptor sets and to evaluate the performance of the fitted models.

The internal and external validation step was repeated over 10 times to get coverage over standard deviation.

5.3.4 Machine Learning Methods

Enzyme Function Classification

Machine learning is the study of algorithms that enable computers to learn and evolve the behaviours that allow them to interpret data. Machine learning is categorised into two main groups, based on the principle of the algorithm to perform classification or regression problem. The two groups are: first, supervised learning, where classification or regression is done based on prior information, and second, unsupervised learning, where no prior information is available to perform classification or regression, such as clustering analysis which we already discussed in Chapter 4. Many machine learning methods are available. We focus on the analysis and comparison of performance of three commonly used supervised approaches, Support Vector Machine (SVM), Random Forest (RF) and K Nearest Neighbour (kNN) for classification problem and PLS model for regression problem.

The following section describes the basic concepts of various machine learning algorithm used in this study [2, 3]. We will introduce the basic concepts with relevant detail reference.

Support Vector Machine (SVM): Support vector machines, is very popularly used for classification and regression problems, which was developed by Vapnik [174]. This model has been successfully applied to various fields of study such as chemoinformatics, structure-function prediction [2, 3, 68, 175] or to classify protein function [163, 176]. Due to very sophisticated mathematical equations, SVM has mostly been described as a black box. However, this algorithm can be explained basically by four main concepts: 1. the kernel function, 2. the separating hyperplane, 3. the maximum margin hyperplane, and 4. the soft margin [177, 178].

First, the data $D = (x_1, y_1), ..., (x_n, y_n) \subset \mathbb{R}^3$, where $x_i(i = 1, ..., n)$ is a vector representing chemoinformatics descriptors. Whereas y_i is prediction labels (EC class representation and LogS for respective studies). The D is mapped onto a higher dimension feature space defined by the kernel trick function $[k(x_i, x_j) =< \phi(x_i).\phi(x_j) >]$ (as illustrated in Figure 5.3), to allow SVM to perform a two dimensional classification of a set of originally one dimensional data [179, 180]. Next, a clear separation between two classes is evaluated by creating a hyperplane. To get a clear separation the hyperplane is between the two classes by maximal distances from the given training set. This is the idea behind maximum margin in SVM. In cases where the data are not well separated or data points are on the 'wrong' side of the margin, soft margins are allowed to consider those points. That means this provides flexibility to push some data points through the margin of the separating hyperplane without affecting the final results.

For support vector regression, the goal is to find a function $f(x) = \omega^t x_i + b$, where ω is a vector of weight and b is the coefficient, that captures the deviation of $f(x)$ from the actual target y_i for all training sets (see Figure 5.4) (for more details [181]). This deviation should be at most ϵ in magnitude, ϵ is the loss function. This is the tolerance level for making prediction.

Here, we have used two kernel functions: Polynomial Function $k(x, x')$ $= (scale < x, x' > +offset)^{degree}$ and Gaussian Radial Basis Function $k(x, x') = exp(-\sigma|x - x'|^2)$. These kernel functions are suited to tackle (non-)learner classification or regression problems. For more details, there are excellent review papers on SVM [68, 181]. SVM has many applications for classification and regression problems in bioinformatics and chemoinformatics [2, 3, 175, 182–184].

Random Forest (RF): A Random Forest is an ensemble of decision trees $T_1(x), ..., T_i(x)$, each tree is generated by stochastic recursive partitioning of a bootstrap sample of the training set. Trees are constructed by the Classification And Regression Trees (CART) algorithm [185] without pruning. As the instances progress through the tree, they are partitioned into increasingly homogeneous groups, so that each terminal node of the decision tree is associated with a group of instances with similar properties. Each split

[3] We will use this definition of data throughout this chapter

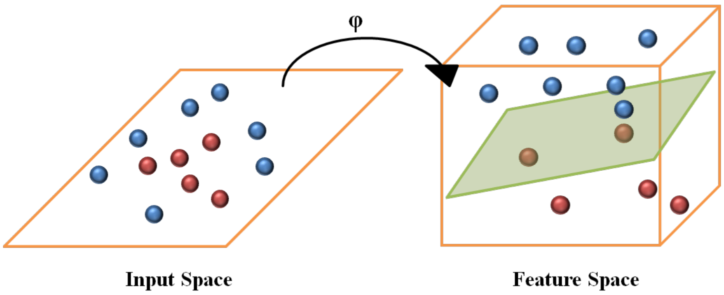


Figure 5.3: SVM maps the data point to higher dimensions: In this figure we show a simple example of two classes where the linear classification is difficult. Using kernel trick the data is transferred to a higher dimension that makes the hyperplane.

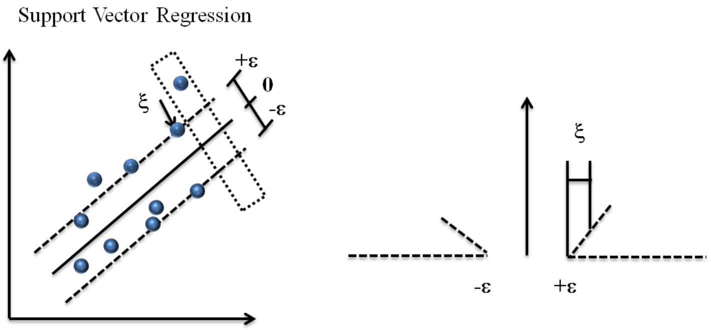


Figure 5.4: This figure illustrates support vector regression. Here ϵ works as a loss function to give flexibility for prediction.

within a tree is created based on the best partitioning of the bootstrap sample, according to the GINI impurity ((GI) criterion see Figure 5.5), that is possible using any of a randomly chosen subset of *mtry* descriptors. This random subset is freshly chosen for each node. If *ntree*, the number of trees in the forest, is held constant then *mtry* is the only parameter that needs to be optimised. For each tree, approximately one third of the training set molecules do not appear in that tree's bootstrap sample, and constitute the so called out-of-bag data; conversely, every molecule is out-of-bag for about a third of the trees.

Trees grown in this way can then be used to predict unseen data. By the concept of consensus decision making, RF produces the results. Classification of an unseen data point is done by putting the input data down each tree in the forest. Based on the plurality of the votes by each tree, class assignment is predicted. Whereas, in regression the output of each given molecule is averaged to produce the final prediction for y value.

Advantages of RF include not having to split the data into separate training and test sets (if out-of-bag validation is used), and especially RF's tolerance of unimportant descriptors. This means that it is not usually necessary to carry out descriptor selection with RF [175, 186–189].

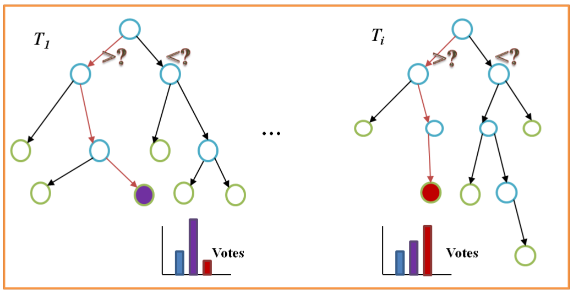


Figure 5.5: In Random Forest, each tree (T_i) is built and trained independently. During testing, each test point v is simultaneously pushed through all trees (starting at the root) until it reaches the corresponding leaves. This figure illustrates an example of decision tree, where each split is based on the question leading to a leaf node (green circle).

K Nearest Neighbour (kNN) (for classification only): kNN is one of the simplest classifier models to understand. Underlying function of this algorithm is very straightforward, queries are classified based on the nearest data point class [190]. The basic idea of this algorithm is illustrated in Figure 5.6. The algorithm can be defined in three simple steps: first, computing distance, second assign k and third, votes.

In classification step, first, a given instance q (the query) whose attribute is referred to $q.A_i$ is classified into one of the classes. In kNN, the class of q is found as follows:

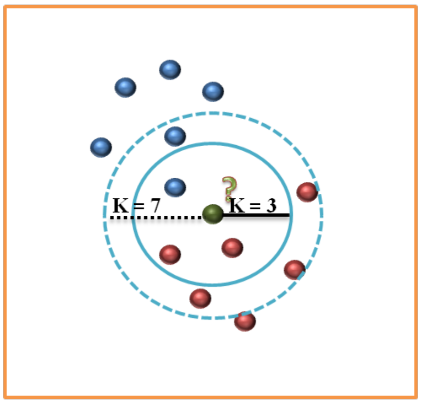


Figure 5.6: If $k = 3$, then in this case query '?' instance will be classified as red since two of the nearest neighbours are red.

First: The distance is calculated between query q and rest of the data points. Various distance functions can be used to compute the distance [190], such as Euclidean distance, which is a very popular distance metric.

Second: Find k instances in the data set that are closest to q. Example in Figure 5.6, k is assigned to either $k = 3$ or $k = 7$.

Third: Based on the maximum votes the q is assigned to the class. In example Figure 5.6, the maximum votes go to red class, hence, the instance is assigned as red.

In a case where $k = 1$, the class label for q is therefore copied directly from the nearest neighbour, as this strategy was used in [23] to annotate protein function. The principle behind distance is that instances with the same class label (in [2], EC class) are expected to have smaller separating distance compared to instances that fall under different classes.

Partial Least Squares (PLS for Regression only): PLS regression is very commonly used to reveal the structure of chemoinformatics descriptors [175, 191–193]. The strength of PLS is to solve the multicollinearity problem (a statistical phenomenon which suggests that two or more predictor variables are highly correlated leading to dubious conclusions) and provide an additional benefit of noise filtering. In PLS, the input data set D is decomposed into T, which is referred to as a latent variable, factor or principal component, then factor T is regressed with the property of interest Y (in study [3] LogS). The factor T explains more information on the predictor and is also correlated as much as possible with Y. This solves multicollinearity problem. The matrix representation of the PLS is illustrated in Figure 5.7.

PLS can be explained in matrix form, where first scores for X matrix are calculated by considering any one vector of Y; Let u_1 = any y. Next a weight matrix W is calculated for X-blocks; $w_1^T = {}^{u_1^T}/_{u_1^T u_1}$, which is then scaled to be in range of 0-1 using; $w_1^T = {}^{w_1^T}/_{w_1^T w_1}{}^{1/2}$. Further, for calculating the weight of X-loadings; (L); $t_1 = Xw_1$. Then, X-scores are used to calculate Y-loadings (Q); $q_1^T = {}^{t_1^T}/_{t_1^T t_1}$. This is further used to calculate new Y-scores (U); $u_{1,new} = {}^{Tq_1}/_{q_1^T q_1}$. These steps will continue to iterate until converged to a stable solution. Further, calculation will deal with finding 'inner' relationship between loadings U and T that is simply calculated using Y and X blocks.

Finally, latent variable for X is calculated as $p_1^T = {}^{t_1^T X}/_{t_1^T t_1}$. For the iteration to calculate next latent variables, the information linked to the first variables are subtracted from the original data $E = X - t_1 p_1^T$ and $F = Y - bt_1 q_1^T$. Now, E and F are used as new X and Y at the start of the computation. The overall process is repeated until the desired number of latent variables are extracted. Hence, the desired number of latent variables is the user-defined parameter.

Range of parameters to optimise:

- SVM: We have used two kernel functions in studies [2, 3]. In Gaussian RBF kernel, the number of free hyper-parameters are C (cost parameter) and kernel width l. For Polynomial kernel, there are three parameters to be regularized: C, *scale* and *degree*. The cost value C is to control the complexity of the decision boundary.

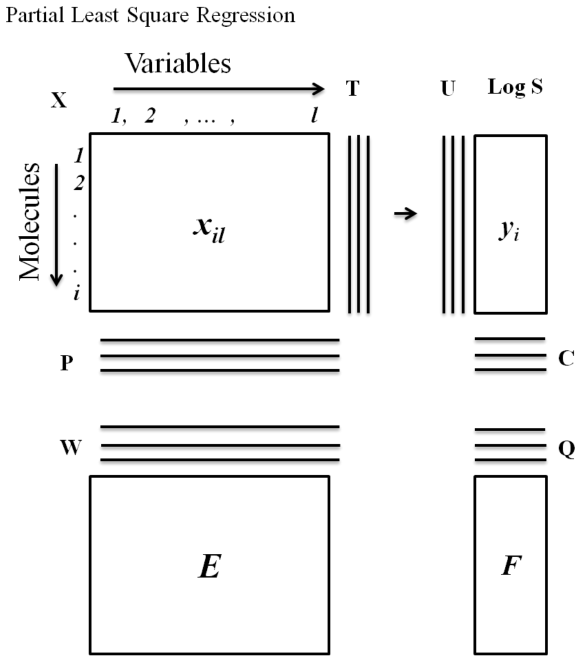


Figure 5.7: This figure illustrates the matrix calculation of PLS. Where first scores (T and U) are computed in such a way that they possess maximum information of respective matrix (X and Y) and are related in some respect. Further used for calculating loadings (W and Q).

- RF: we treat the subset size *mtry* as a hyper-parameter of the method, where M is the number of descriptors.
- kNN: K (number of the nearest neighbour) is the parameter that needs to be selected.
- PLS: the number of components (or latent variables) are optimised.

R packages used in this study [126]: We have used CARET [194,195] package to design the overall work-flow (for regression and classification, Figure 5.1). CARET package calls other packages to build model such as, for RF, CARET calls package 'randomForest' [196]. For SVM, we used two functions 'svmpoly' and 'svmradial' for polynomial kernel and radial kernel respectively, the models were implemented using 'kernlab' [197] package. The PLS, pls package [198] was called. Whereas, kNN was implemented in caret package itself.

5.3.5 Step 3: Validation

Statistical Significance Test

Statistical Test Formulas

For classification, using accuracy measure Equation 5.2 to evaluate the performance of the model is very common practice [199] and easy to interpret. Once, the final model is fitted with an unseen dataset, accuracy is calculated [200]. This measure estimates the performance of the classifier, and is easily explainable and widely used to quantify the proportion of correct predictions made. Nonetheless, this measure is not good when the distribution of the data class is imbalanced. Here, we have used accuracy as the criteria for selecting the parameter in internal validation phase, hence our training model prediction might suggest small bias towards the large class.

	Positive	Negative
Test positive	TP	FP
Test Negative	FN	TN

- TP : true positives, the number of objects that are correctly classified to the decision class,

- TN : true negatives, the number of objects that do not belong to the decision class and were not classified to it
- FP : false positives, the number of objects that are incorrectly classified to the decision class,
- FN : false negatives, the number of objects that do belong to the decision class but were incorrectly classified to the other class.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.2}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i^{obs} - y^{pred})^2} \tag{5.3}$$

$$R^2 = \left(\frac{\sum(x^{pred} - \bar{x})(y^{pred} - \bar{y})}{\sqrt{\sum(x^{pred} - \bar{x})^2 \sum(y^{pred} - \bar{y})^2}}\right)^2 \tag{5.4}$$

Moreover, for regression analysis, we used Root Mean Square Error (RMSE) (Equation 5.3) and R^2 (Equations 5.4 (R^2 here is the square of the Pearson correlation coefficient, not the coefficient of determination)) for evaluating the performance of our method. And n is the number of molecules; y^{obs} is the observed output and y^{pred} is the predicted output.

Statistical Significance Test: The permutation test is a widely used technique in various research areas such as in bioinformatics and chemoinformatics where the question is how well algorithm A performed compared with algorithm B on a particular problem characterised by a data set D. For this, we used the permutation test to calculate exact P-values (Equation 5.5) suggested in [201] for the commonly used 10-fold cross-validation methods.

$$P = \frac{n}{N} \tag{5.5}$$

where n is the number of permutations of the mean difference in the performance of two regression models, which can be more extreme than the observed mean difference, and N is the total number of possible reassignments of the paired differences given the results. In more detail, the procedure consists of the following steps:

1. A given paired-difference (B_0) of accuracy/RMSE scores obtained by different classification/regression models is given by $B_0 = (R_A^1 - R_B^1) + (R_A^2 - R_B^2) + \cdots + (R_A^{10} - R_B^{10})$ where R_A^1 is the accuracy/RMSE scores for test set predictions made by model A for each fold (1 ... 10) in the 10-fold cross-validation.

2. For this statistical test, 1024 permutations are created via all 2^{10} combinations: $B_p = \pm(R_A^1 - R_B^1) \pm (R_A^2 - R_B^2) \pm \cdots \pm (R_A^{10} - R_B^{10})$.

3. The rank of true difference in the performance (B_0) is used as an indicator of the p-values among the 1024 permutations. The p-value is computed as: $P = n/1024$ where n is the number of permutations which have $|B_p| \geq |B_0|$.

5.4 Results

In this section, we describe the results from two studies [3] and [2].

5.4.1 Classification: Enzyme Function Prediction

Our results for enzyme function prediction suggest descriptors representing overall reactions gave more accurate prediction as compared with mechanistic information. Thus out of 4 descriptors (as mentioned earlier in Chapter 3, section 3.1.3), *OS* and *OBC* have consistently given better predictive accuracy than descriptors annotating mechanistic information. This implies that EC function annotation can be predicted by overall reactions better than mechanistic information can, and suggests that reaction mechanisms cannot be used to assign EC class. EC classification system classifies enzymes based on overall transformation without incorporating mechanistic information into consideration. In Figure 5.8, we graphically represent the performance of various machine learning methods of different descriptors. We represent detailed results in the following paragraphs:

Results of MACiE enzyme descriptors Out of all five descriptors used for classification prediction the best prediction was given by *Human designed* descriptors. In summary, for all four machine learning methods the relative performances of the descriptors are the same: *Human Designed* >*Overall Reaction Similarity* >*Overall Bond Change* >*Composite Bond Change* >*Mechanistic Similarity.*

Overall, we found that RF outperformed other classifier models used here, when predicting function from all the descriptors. The two kernel functions for support vector machine, Poly and RBF, performed similarly, coming in second and third place respectively, leaving k Nearest Neighbours ranked last. In detail, the classifiers', performance for individual descriptors are as follows:

- For us it is unsurprising that predicting enzyme function using *Human designed* descriptor was best among all as they are designed to best represent the features specific to enzyme reactions. The training set performance showed 90% accuracy (results are shown in Figure 5.8), while when tested with unseen dataset the performance was reduced to 83% (results shown in Figure 5.8). The classifiers', performance was as; RF >SVM (poly) >SVM (RBF) >kNN.
- *OBC* showed ≈ 68% EC class prediction, as compared to *CBC*, which showed 62% correct EC class prediction. For both descriptors the classifiers performed similarly: RF ∼ SVM (RBF) >SVM (poly) >kNN.
- Similarly, *OS* descriptor performed better by correctly predicting 70%, while *MS* gave 60% accuracy. The performance of the classifier for both *OS* and *MS* was similar: SVM (RBF) ∼ RF ∼ SVM (poly) >kNN.

Results for the EC top class: Among the EC class, *Human designed* descriptors predict EC 1, oxidoreductase, and EC 5, isomerase, much better than the rest of the descriptors, as these descriptors were engineered to capture specific features from the respective EC classes. For example, in EC 5, out of 30 MACiE enzymes 27 enzymes employ the simple stoichiometry of one starting material being transformed into one isomeric product. Interestingly the ranking of the descriptor performances matches with the

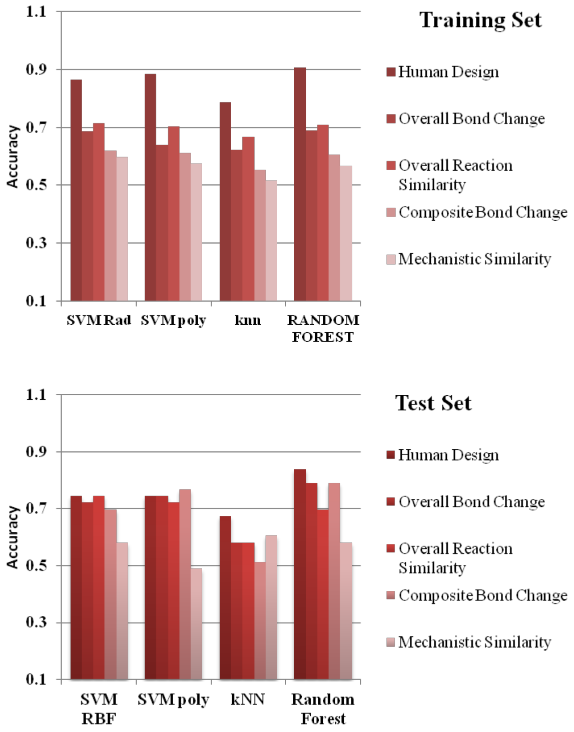


Figure 5.8: Performance of different classifiers in cross-validation. The Figure shows the accuracy achieved by each of the four classifiers for each of the five descriptor sets in the cross-validation.

overall prediction performance (see Table 5.1). It was difficult to predict EC 2 when descriptors were encoded based on overall reaction as compared to when mechanistic information was added. This is in contrast to EC 5, where prediction performance was worst among all when descriptors represented mechanistic information. This suggests that mechanistic information based descriptors did not have enough information to be able to detect enzymes catalysing isomerase activity.

Table 5.1: Performance of different classifiers in cross-validation for individual EC class.

Descriptors	EC 1	EC 2	EC 3	EC 4	EC 5	EC 6
HD	0.961	0.816	0.948	0.835	0.952	0.877
OBC	0.823	0.394	0.849	0.605	0.500	0.731
OS	0.865	0.500	0.828	0.568	0.628	0.600
CBC	0.870	0.406	0.680	0.495	0.276	0.654
MS	0.817	0.363	0.722	0.334	0.333	0.315

5.4.2 Regression Analysis

From regression analysis in [3], we show the results (see Table 5.2) for predicting LogS on a benchmark dataset, solubility challenge dataset [167]. Our results showed that machine learning workflow prediction for regression analysis was much better than various commercially available learning methods [202]. However, these results are not directly comparable because of the nature of cross validation used in this study.

Table 5.2: Solubility Challenge dataset: average over ten repetitions of 10-fold cross-validation of RMSE(standard deviation) for the log S calculation.

Machine learning	Raw data	variance scaling	PCA Scaling
RF	0.9 ± 0.01	0.93 ± 0.01	1.12 ± 0.01
SVM	1.17 ± 0.04	0.93 ± 0.02	0.95 ± 0.02
PLS	1.08 ± 0.04	1.03 ± 0.02	1.15 ± 0.01

The machine learning method was further used to test the predictions of solution free energy using physics-based theory alone and quantitative structure − property relationship (QSPR) models, designed by James McDonagh [3]. While direct theoretical calculation does not give accurate results in this approach, machine learning is able to give predictions with a root mean squared error (RMSE) of ∼1.1 log S units in a 10-fold cross- validation for our Drug-Like-Solubility-100 (DLS-100) dataset of 100 drug-like molecules. We find that a model built using energy terms from our theoretical methodology as descriptors is marginally less predictive than one built on Chemistry Development Kit (CDK) descriptors. Combining both sets of descriptors allows a further but very modest improvement in the predictions. However, in some cases, this is a statistically significant enhancement. These results suggest that there is little complementarity between the chemical information provided by these two sets of descriptors, despite their different sources and methods of calculation. Our machine learning models are also able to predict the well-known Solubility Challenge dataset with an RMSE value of 0.9−1.0 log S units.

Here, we used cheminformatics descriptors to predict the solubility of drug-like molecules. As a benchmark, we also present our method's predictions of the solubility challenge set based solely on cheminformatics descriptors (see Table 5.3). As suitable crystal structures are not available for all molecules in the solubility challenge, we could not calculate the theoretical energies.

Tables 5.2 and Table 5.3 demonstrate that our method can make predictions for the solubility challenge dataset within the coveted 1 log S unit RMSE error and, in fact, makes predictions that are consistent with some commercially available methods and deep-learning methods. A recent publication [202] reported RMSE scores of 0.95 log S units [202] for the commercially available package MLR-SC62 and 0.90 log S units for a deep-learning method [202]. However, these results are not directly comparable with ours, for two reasons. First, our results have been calculated for a 10-fold cross-validation and for the canonical training:test split (see Table 5.3). Second, the deep-learning result (RMSE = 0.90) given by Lusci et al. [202] is contingent on correcting eight putative errors in the CheqSol solubility data.

5.5 Summary

This chapter summarises two studies [3] and [2], where we used similar work-flow for two different problems, regression and classification respectively. Our results from enzyme classification strongly suggest that the use

Table 5.3: RMSE for the Log S Calculation using the solubility challenge dataset with its original training:test split

Machine learning	Raw data	variance scaling	PCA Scaling
PLS	0.86	0.91	0.91
RF	0.93	1.03	1.02
SVR	1.08	1.07	1.08

of mechanistic information has a diminished EC prediction performance relative to overall transformation. For regression, chemoinformatics descriptors suggest enough potential to predict solubility of drug-like molecules. Moreover, it is evident in both the studies that RF outperformed other machine learning methods, with close competition from SVM.

6

Enzyme Function Evolution: *Chemolution Study*

6.1 Background

THe three-dimensional (3D) atomic structures of contemporary proteins provide clues about how both structure and function unfolded in the course of billions of years of evolution [203]. The phylogenomic analysis of protein domain occurrence and abundance in modern proteomes [13, 204] enables retrodictive views of protein evolution that are unanticipated [117, 205] and can be used to study structural change and the relationship between protein structure and function [116]. Two recent studies of this kind showed congruently that the α/β architecture is probably the oldest type of fold design [13, 204].

An interesting observation [13, 206], regarding the Enzyme Commission (EC) [5] definition of the overall function of enzymes, is that the oldest fold structures were associated with the largest number of enzyme functions [13, 60, 206, 207].

Understanding how enzymes adapt their chemical mechanisms under evolutionary pressure is still a challenging task in molecular biology. In this chapter, we explore the chemical mechanisms used in biochemical reactions catalysed by ancestral enzymes. We ask questions about the ways in which enzyme structure and chemical mechanism have evolved together, and about the evolutionary origination of new enzyme structures and new catalytic mechanisms. As mentioned in Chapter 3, the EC classification does not explore the detailed chemical mechanism of the enzyme reaction [19,98]. MACiE [19,98] definitions of enzyme mechanisms and ages of domain structures (MANET) [103] derived from phylogenomic analyses of protein structure [13,117,208] dissected the evolutionary appearance of novel structures and functions. It has been suggested that the difficulty of evolving novel stepwise chemical reaction mechanisms could be the dominant factor limiting the divergent evolution of new catalytic functions in related enzymes [208]. We put this concept to the test with phylogenomic analysis of protein domain structure and careful annotations of reaction mechanisms. Our observations have important implications for the origins of modern biochemistry and for exploring structure-function relationships.

6.2 Method

As we mentioned in Chapter 3, the data is retrieved from MACiE V3.0. The mechanistic step types were mapped onto the phylogenetic timeline from the MANET database. These databases are discussed in detail in Chapter 3. The calculation of the *nd* value is also mentioned in Chapter 3 with the definition of components of MACiE used in this chapter.

6.2.1 Data Culling

Here, we emphasise on the mapping of mechanistic step types onto the catalytic domain structure in MACiE. In many enzymes, not all domains were actually involved in catalysis. We made sure that the data under investigation is the best representation of the information for function evaluation. We considered 236 (Figure 6.1 illustrating the filtration of the data in MACiE) unique CATH folds in this analysis, such that we could assign *nd* (fold age) values to the respective enzymes. Second, domains that were not involved in the reaction were discarded. There are many enzymes with more than one domain, for example MACiE entry M0124 (EC: 1.9.3.1, cytochrome-c oxidase) was associated with 16 domains, of which only one domain (CATH: 1.20.210.10 - Cytochrome C Oxidase, chain A) was used as a catalytic domain to complete the reaction. A careful filtering was done by only selecting domains that were participating in the reaction as a catalytic chain. The catalytic domain distribution of the remaining enzyme structures was as follows: 240 enzyme entries with a single catalytic domain, 63 enzymes having two different catalytic domains, four enzymes with three catalytic domains and only one enzyme entry in MACiE (M0207, EC 2.7.9.1, pyruvate-phosphate dikinase) with four domains (CATH 3.30.1490.20, nd = 0.0539; CATH 3.30.470.20, nd = 0.058; CATH 3.20.20.60, nd = 0.112; CATH 3.50.30.10, nd = 0.377) that participate in catalysis; pyruvate-phosphate dikinase is a key enzyme participating in gluconeogenesis and photosynthesis. Thus, a total of 308 MACiE enzymes were considered for further analysis. Only these H-level structures were used further to explore the evolution of biocatalytic mechanisms.

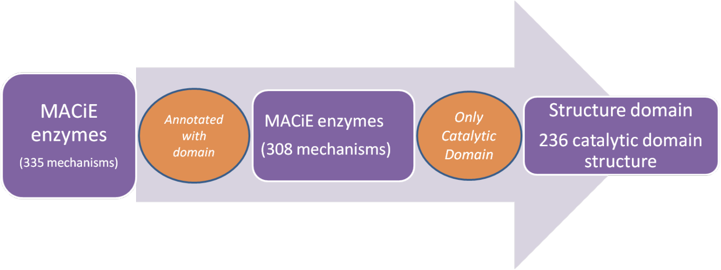


Figure 6.1: This figure illustrates the filtration process to retrieve only enzymes possessing catalytic domains. Out of 335 MACiE enzymes only 308 MACiE enzymes have 236 catalytic domains, which participate in the catalytic reaction, hence providing the mechanistic step type definitions.

6.2.2 Phylogenetic Analysis:

Using shared and derived characteristics of the protein such as fold-usage Genomic Abundance (G), Gustavo Caetano-Anollés group [116] have quantified the relative time or age of the fold (*node distance : nd*). First, the count of domains present in genomes was retrieved and normalized to compensate for the difference between the genome sizes. Second, in order to establish the evolutionary direction, the maximum states were specified as being the ancestral. Third, maximum parsimony was used to reconstruct the phylogenetic trees. This reconstruction of the phylogenetic trees is based on two assumptions: first, protein structure is far more conserved than sequence, and second, the most ancestral folds are generally most popular and abundant in nature. The relative age of individual protein folds is calculated by measuring a distance in nodes from the hypothetical ancestral fold on a relative 0-1 scale, where 0 is the most ancestral fold and 1 is the youngest fold. The *nd* depicts the number of cladogenic events along a lineage and was used as an indicator of the ancestry of each metabolic enzyme for which a protein structure is known or could be inferred.

Phylogenetic statements relate to definitions of structures that are modern and are constructed from a structural census in the proteomes of extant organisms. Consequently, retrodictions are derived from modern structural complexity and do not necessarily depict the actual structure of hypothetical ancestors, which will always remain unknown (molecules can be brought back from the past experimentally by resurrection but cannot be confirmed to be truly bona fide retrodictive constructs). However, if molecules become structurally canalized in evolution, then modern retrodictive statements truly approximate molecular history.

6.3 Findings

6.3.1 A General Approach Grounded in Protein Domain Structure

In order to test the hypothesis that the most ancestral protein domains use the greatest number of biocatalytic mechanistic step types, we assume that extant protein domain structure is the best historical archive that is available to explore ancient enzyme functions. The assumption holds good ground. At high levels of structural complexity, evolutionary change occurs at an extraordinarily slow pace. A new fold superfamily may take hundreds of thousands to millions of years to materialize in sequence space while new sequences develop on Earth in less than microseconds [209]. In fact, a recent comparative analysis of aligned structures and sequences showed that structures were 3–10 times more conserved than sequences [59]. Here we use the ages of domain structures, derived from phylogenomic reconstruction and a recent census of CATH domain structure in hundreds of genomes [117], to study how chemical mechanisms developed in protein evolution. The use of molecular structure and abundance in phylogenomic analysis offers numerous advantages over traditional methods [210], eliminating phylogenetic problems such as alignment, phylogenetic inapplicables and taxon sampling. Their use does not violate character independence, a serious problem that has not been addressed in phylogenetic sequence analysis. To our knowledge, this is the first study to explore the evolution of biocatalytic mechanisms using a timeline of CATH homologous superfamily (H-level) domain structures and data analysis. However, there is another comprehensive database, FunTree [102], that brings together sequence, structure from CATH, chemical and mechanistic information from MACiE, and phylogenetics.

6.3.2 Historical Trends Unfold a Natural History of Biocatalytic Mechanisms

In order to explore the use and reuse of biocatalytic mechanisms in evolution, we mapped the mechanistic definitions of enzymatic functions to their respective CATH H-level structures, with structures ordered according to fold age (Figures 6.2, 6.3, 6.4). For this purpose we first created a presence and absence (PA) matrix, a heat map representing the distribution of the presence (red) and absence (yellow) of the mechanistic step types (rows, y-axis) in the fold (columns, x-axis) (Figure 6.2). The rows were ordered vertically according to the first appearance of the mechanism over fold age and were indexed with the numbers of: (I) MACiE enzyme entries (shades of grey and black), (II) H-level structures (shades of grey and purple), and (III) EC classes that appeared at each age.

Remarkably, the most popular enzyme mechanistic step types were associated with the oldest H-level structures (Figure 6.2). This evolutionary trend suggests that the oldest enzymes already provided a sufficiently flexible scaffold to support many diverse mechanistic step types in order to complete their reactions. Within the early scaffolds, the mechanistic steps had more time to be adapted by the domain structures and to be further recruited in the course of evolution. The existence of late emerging structures with many mechanistic steps supports the existence of widespread recruitment processes in evolution. This trend seems to be explained in terms of the "preferential attachment principle" that guides the growth of scale-free network behaviour, and implies that the more prevalent functions are typically the earliest, as previously shown in the exploratory analysis of the ancestral fold structures [211].

We observed that 'proton transfer', 'bimolecular nucleophilic addition', 'bimolecular nucleophilic substitution', and 'unimolecular elimination by (or from) the conjugate base' (definitions are represented in Figure 3.3) are the

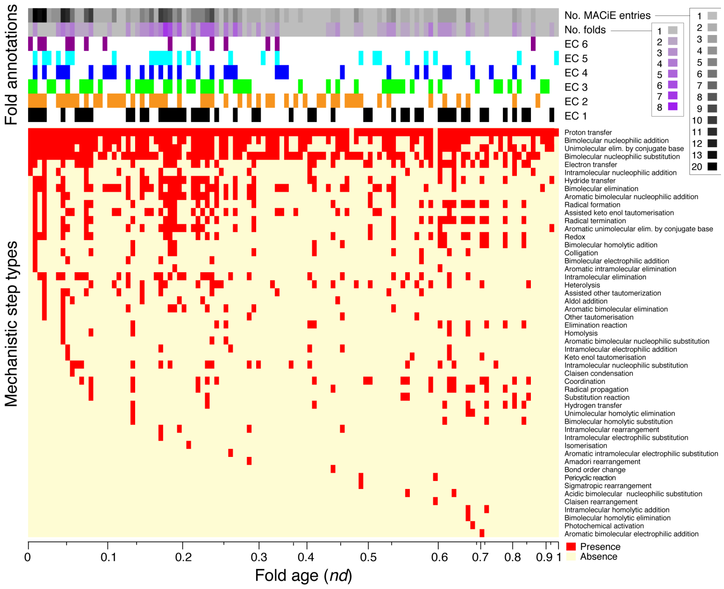


Figure 6.2: The heat map describes the distribution of presence (red) and absence (yellow) of mechanism step types (y-axis) over fold age (x-axis). Rows of the heat map (mechanisms) are ordered vertically according to the first appearance of the step type in time, with the oldest at the top. The row sidebars at the top of the heat map are used to describe the number of MACiE entries and CATH H-level domain structures (annotated as number of folds) appearing at each fold age, and presence of top-level EC classes that are associated with these H-level structures (see colour key). The x-axis scale reflects the different *nd* values found in our dataset, arranged from the oldest on the left to the youngest on the right. Every unique *nd* value forms a separate column. The non-linear scale is defined by the number of unique *nd* values falling in each interval of *nd*. There are many distinct *nd* values between 0.0 and 0.3 found in our dataset, so the scale is expanded in this region. There are few distinct *nd* values between 0.7 and 1.0, so the scale is very condensed in that region. Geological time is taken to be approximately linear with *nd*, where *nd*=0 represents the origin of the protein world approximately 3.8 billion years ago and *nd*=1 corresponds to the present.

most common mechanistic step types, in accordance with their distribution in MACiE enzyme reaction mechanisms (the prevalence of each step type is also given in Appendix A.2) [19, 106]. These types of mechanistic step are recognisably fundamental building blocks of enzyme chemistry, which is carried out in aqueous solution usually at approximately neutral pH. Several of the canonical amino acids have pKa values close to neutral, with Holliday et al. [19] having observed particularly strong propensities for His and Glu to facilitate proton transfer. The chemistry of the amino acid side chains also means that several are negatively charged at roughly neutral pH, and hence it is no surprise that the enzyme far more often acts as a nucleophile, favouring mechanisms labelled as nucleophilic, rather than as an electrophile. Furthermore, it has been noted that enzyme active sites are well suited to stabilising the charged intermediates common in addition and elimination reactions, for instance by hydrogen bonding [212]. The ubiquity of aqueous environments in enzyme chemistry restricts the repertoire of reactions available. Indeed, most enzyme reactions are composed of steps that might seem unexciting to an organic chemist. The rare occurrence of more complicated organic chemistry, 'aldol addition', 'amadori rearrangement', 'claisen condensation', 'claisen rearrangement', 'pericyclic reaction' and 'sigmatropic rearrangement', constitutes the exception rather than the rule, and enzymes sample the space of possible mechanisms notably differently from how an organic chemistry textbook would do so.

The rate of introducing new mechanistic step types at different fold ages is shown in Figure 6.3, which represents a cumulative plot where fold age is shown on the x-axis. The y-axis shows the proportion of the total number of defined step type annotations (N = 51) that have been uncovered up to that fold age on the x-axis. It is clear in this plot that the first four H-level structures (the first two increments of fold age, 0 to 0.0098) are responsible for introducing a third of the known mechanistic step types (18/51), and the first six structures (the first four increments of fold age, 0 to 0.049) are responsible for over half of them (27/51). However, the development of the other half was harder and required the unfolding of about 3/4 of the evolutionary timeline, up to $nd = 0.73$ and about 2.5 billion years of evolution [205]. The detailed information regarding the introduction of mechanistic step types is provided in Appendix A.2.

In order to look at the distribution of the mechanistic step types of an

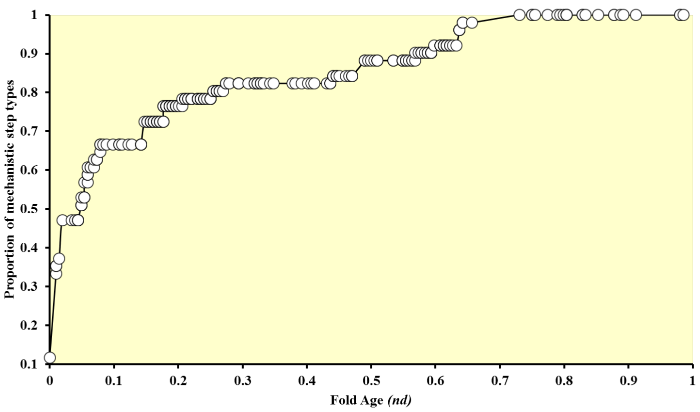


Figure 6.3: The graph shows the proportion of mechanistic step types that are present at a particular time.

enzyme in evolutionary time, we counted the number of mechanistic step types associated with H-level structures (Figure 6.4). Figure 6.4 is a heat map representing the number of mechanism step types (y-axis) used by those structures having each different discrete value of fold age (x-axis). Each cell represents the number of H-level structures with a different colour code; for example black represents 1 structure, yellow represents 2 structures and brown represents 3 structures sharing the same count of mechanistic step types. Moreover, each position indicates the number of H-level structures associated with a number of functions. For instance, black colour at column 1 row 6 means that there is one structure that uses 6 different mechanistic step types to complete its reaction. In a further section, we will discuss the patterns in detail.

6.3.3 Ancient H-level Structures are Popular, Central and Versatile

The most ancient H-level structure that appears in the MACiE database is CATH 3.40.50.300, the P-loop containing nucleotide triphosphate hydrolase. This fold has been consistently identified as the most ancestral fold structure [13, 117, 204]. The P-loop hydrolase structure consists of the most ancient and abundant topology, the Rossmann fold (CATH 3.40.50), which has the 3-layer ($\alpha\beta\alpha$) sandwich (3.40) architecture. The CATH 3.40.50.300 superfamily contains enzymes with diverse molecular functions, including

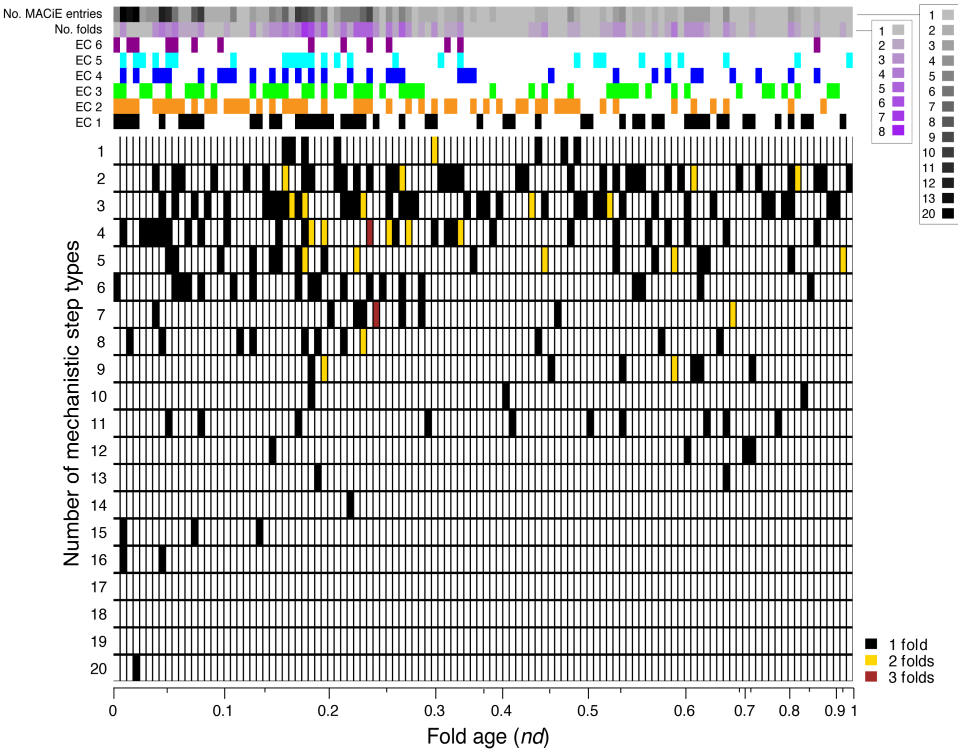


Figure 6.4: Heat map representing the number of mechanistic step types (y-axis) used by H-level structures of each different fold age (x-axis). Different colours indicate distinct structures which happen to share both the same number of mechanistic step types and an identical fold age. For example, in column 2 the black colouring of rows 4, 15 and 16 shows that four structures respectively accommodate 4, 15 and 16 different mechanistic step types to effect their reactions. The colour code for the row sidebar is similar to that in Figure 6.2; the x-axis scale is also similar to that in Figure 6.2.

signal transduction, hydrolase and transferase enzymatic activities [213]. Wang et al. previously observed [208] diverse overall functions for this structure. In the current analysis, there are only five MACiE enzyme entries that share this structure; these are associated with six mechanistic step types, 'proton transfer', 'electron transfer', 'bimolecular nucleophilic addition', 'bimolecular nucleophilic substitution', 'intramolecular nucleophilic addition' and 'unimolecular elimination by the conjugate base' (Table Appendix A.2). MACiE enzymes associated with this oldest structure are dethiobiotin synthase (EC 6.3.3.3, M0074), estrone sulfotransferase (EC 2.8.2.4, M0154), H+-transporting two-sector ATPase (EC 3.6.3.14, M0178), nitrogenase (EC 1.18.6.1, M0212, multi-domain) and adenylate kinase (EC 2.7.4.3, M0290). Except for nitrogenase, the rest of these enzyme entries each have a single catalytic domain, hence, it is straightforward to annotate the function with this fold. Nitrogenase (M0212, PDB: 1n2c) [214] is a very important enzyme of nitrogen metabolism that fixes atmospheric nitrogen (N_2) gas into the reduced forms that are usually assimilated by plants [215]. The enzyme has a complex 3D structure that is highly conserved across many different organisms and contains domains from three different homologous superfamilies. These H-level structures were discovered by evolution at different times. The ancient CATH 3.40.50.300 nitrogenase catalytic core was later accesorized with a domain from the CATH 3.40.50.1980 superfamily, which evolved at $nd = 0.401$ after the oxygenation of Earth's atmosphere [208,216,217], and a non-catalytic domain CATH 1.20.89.10, which appears to have been accreted last into the molecule (nd = 0.549). Residues from the ancient nitrogenase core with the oldest domain of the molecule are involved in the first two steps of the long 15-step reaction, which include the mechanistic step types 'bimolecular nucleophilic substitution', 'electron transfer' and 'proton transfer'. The remaining 13 steps are carried out by catalytic residues from the CATH 3.40.50.1980 domain.

The second most ancient H-level structures include CATH 3.50.50.60, the T-level topology is 3-layer $\beta\beta\alpha$ while the H-level structure, which has no specific name assigned but corresponds to the FAD/NAD(P)-binding domain FunFams definition in CATH, is found in 7 MACiE entries, CATH 3.40.50.720 (NAD(P)-binding Rossmann-like domain, found in 12 MACiE enzymes), and CATH 3.40.50.150 (Vaccinia Virus protein VP39, found in two MACiE entries). All three H-level structures appear at $nd = 0.0098$.

These structures have 16, 15, and 4 catalytic mechanistic step types (Figure 6.4), respectively, of which a total of 11 are non-overlapping with those of the first P-loop hydrolase fold structure and were therefore newly introduced at this time (see Table Appendix A.2). These newly discovered mechanistic step types include three involving aromatic groups, as well as the first involving radicals, and also ‘bimolecular electrophilic addition’, ‘bimolecular elimination’, ‘redox’, ‘colligation’ and ‘assisted keto-enol tautomerisation’. It was interesting to note that the ‘bimolecular elimination’ mechanism was shared with all the three H-level structures of the same age. There are 9 different mechanisms shared by CATH 3.40.50.720 and CATH 3.50.50.60 (shown in Table Appendix A.2). It is also noteworthy that studies by the Orengo group [218, 219] suggest there may be distant homologies between these structures based on their similarity in graph-based structure comparison and shared use of organic cofactors (NAD and FAD). The structures are functionally diverse due to the conformational change of the ligands, organic cofactors or structural plasticity of the proteins [220].

In MACiE, the ferredoxin - NADP+ reductase enzyme (M0142, EC: 1.18.1.2) combines the CATH 3.40.50.150 and CATH 3.50.50.720 H-level structures to complete its biochemical reaction. This enzyme plays a very important role in electron transfer from the flavoenzyme NADPH-adrenodoxin-reductase (AdR) to two P450 cytochromes; this process is involved in the production of steroid hormones. The two domains of this enzyme share the following functions: ‘aromatic unimolecular elimination by the conjugate base’, ‘aromatic bimolecular nucleophilic addition’, ‘redox’, ‘radical termination’, ‘radical formation’.

The third most ancient H-level structure ($nd = 0.0147$), CATH 3.40.50.620, the H-level Hups α/β layered fold, is responsible for 13 MACiE entries and introduces the novel ‘intramolecular elimination’ function. This structure supports central catalytic functions of the cell, including the amino acylation reactions of aminoacyl-tRNA synthetase (aaRSs) catalytic domains that are crucially involved in the attachment of L-amino acids to cognate tRNA molecules and are responsible for the specificity of the genetic code. The structure includes the tyrosyl-tRNA ligase EC function (M0197; EC 6.1.1.1) of the tyrosyl-RS functional family, the oldest aaRSs delimiting the process of translation [221]. The structure activates a specific amino acid by condensation with ATP to form an aminoacyladenylate intermediate, which then esterifies the 2’ or 3’-hydroxyl group of the ribose at the 3’ end of the acceptor arm of tRNA. The process, which is highly specific, involves proofreading.

6.3.4 Some Structures Hold Exceptionally Diverse Mechanistic Step Types

Some H-level structures by nature use many diverse mechanistic step types to effect their catalytic activity. A member of the TIM barrel α/β structure that is highly popular in metabolism, the CATH 3.20.20.70 superfamily (aldolase class I, $nd = 0.0196$), which immediately follows the aaRS fold in the timeline, uses steps with 20 different mechanistic step types. Five of these appeared for the first time with this fold (Table Appendix A.2). It is not surprising that the fold has such diverse functions. Based on the Hierarchic Classification of Enzyme Catalytic Mechanisms (RLCP; where R: Basic Reaction, L: Ligand group involved in catalysis, C: Catalysis type and R: Residues/cofactors located on Proteins) classification [101] analysis of functional subclasses, Nagao et al. [222] suggested that aldolase class I enzymes have various functional classifications. An interesting conserved property is that most of the ligands have at least one phosphate group. The mechanistic step types of aldolase class I (see Table 6.1) are rare in the MACiE database. Out of 335 MACiE enzyme entries, ‘aldol addition’, ‘aromatic bimolecular elimination’, ‘assisted other tautomerisation’, ‘heterolysis’ and ‘other tautomerisation’, respectively, appeared in 9, 6, 20, 25 and 9 MACiE enzyme entries in at least one stage of the reaction. This suggests that the aldolase class I superfamily contains a group of enzymes that possess very specific mechanistic step types.

Two additional H-level structures utilize 16 different mechanistic step types each, CATH 3.50.50.60 ($nd = 0.0098$) (which we have already mentioned) and CATH 3.40.50.970 ($nd = 0.049$), the second largest number of mechanistic step types associated with structures in the timeline. These structures also belong to the most popular fold topology, the Rossmann fold. Following their appearance ($nd = 0.049$), most of the basic and common mechanistic step types had already been introduced. The CATH 3.40.50.970 structure introduces ‘homolysis’, represented in only one MACiE entry (M0119 ; EC: 1.2.7.1; pyruvate: ferredoxin oxidoreducatse). We observed that two mechanistic step types, ‘homolysis’ and ‘colligation’, were introduced at the same fold age but by different H-level structures. By definition, the ‘homolysis’ mechanistic annotation is the converse of the ‘colligation’ step that was introduced by CATH 3.50.50.60; ‘homolysis’ is the cleavage of a covalent bond where each atom retains one of the two bonding electrons, whereas ‘colligation’ is when the two free radicals combine to form a covalent bond.

6.3.5 The Combinatorics of Mechanistic Steps Reveals Winners

We were also interested to see what sets of mechanistic step types described the combinations of steps used by various enzymes to effect their reactions. To do so, we looked for the combination of the different mechanistic step types, irrespective of order, and at the various H-level structures sharing each combination of biochemical steps. Instances of re-utilisation of particular mechanistic step types may shed light on evolutionary recruitment of common mechanistic steps by different structures. For this we first created “mechanistic annotation patterns”. These patterns reflect all the different combinations of the presence and absence of mechanistic step types. This kind of analysis illustrates that different H-level structures share common mechanistic annotation patterns. We found that there are 133 different mechanistic annotation patterns used by the enzymes in our dataset. Pattern 4 is most popular mechanism combination, involving ‘bimolecular nucleophilic substitution’ and ‘proton transfer’ (see Figure 6.5, H-level structures are grouped together in the white box). In MACiE, there are 42 H-level structures that use two mechanistic step types in order to complete their reactions. Out of these 42 structures, 30 use pattern 4 in order to complete their reactions. Patterns 4 and 15 suggest that there are few H-level structures that accommodate similar mechanistic step type combinations.

Pattern 15 is the second most popular pattern and includes ‘bimolecular nucleophilic addition’, ‘proton transfer’ and ‘unimolecular elimination by the conjugate base’. In MACiE, there are 46 different catalytic H-level structures that use three mechanistic step types in order to complete their reactions, out of which 22 structures use pattern 15 to effect their reactions. The enzymes of the CATH 3.20.20.70 (aldolase class I) structure use the maximum number of 20 different mechanistic step types to effect their overall reactions. These step types constitute pattern 133 (see Table 6.1), which is not shared by any other structure. These patterns suggest which mechanistic step types are compatible with one another or are preferentially combined together. There are 101 patterns unique to one structure.

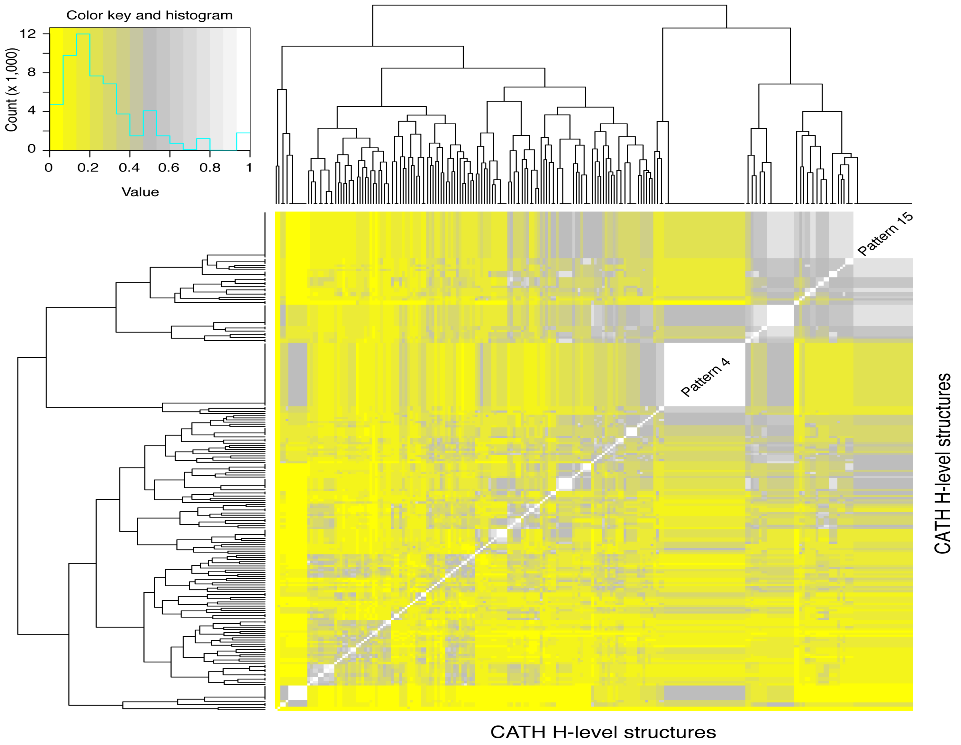


Figure 6.5: For this figure we have calculated the Jaccard similarity scores. Here the x and y axes in the plot are ordered using a hierarchical clustering algorithm in which the two most similar data points are linked together at each iteration. The colours of the heat map represent the similarity scores where yellow suggests low or no (when 0) similarity and white (1) means that identical combinations of mechanistic steps are shared between two H-level structures. The top left corner represents the colour key for the similarity scores and the distribution of the similarity scores.

To visualise the combinatorial patterns, we have plotted a heat map of similarity of the mechanistic step types between two H-level structures (Figure 6.5). We calculated the Jaccard similarity scores:

$$Jaccard = \frac{|A \bigcap B|}{|A \bigcup B|} \tag{6.1}$$

where A and B are two sets and the Jaccard coefficient of similarity is defined as the size of the intersection divided by the size of the union between the two sets. To visualize computed similarity scores, we constructed a presence and absence (PA) matrix where columns represent the mechanistic annotation as an entity and rows represent the CATH H-level structures. The score ranged from 0 to 1, with 0 signifying that no similar mechanistic step types existed between two structures and 1 signifying that the two structures shared an identical combination of mechanistic step types in order to complete their reactions. The most popular mechanism combinations, pattern 4 (‘bimolecular nucleophilic substitution’ and ‘proton transfer’) and pattern 15 (‘bimolecular nucleophilic addition’, ‘proton transfer’ and ‘unimolecular elimination by the conjugate base’), are labelled in the heat map of Figure 6.5 and are clearly distinguishable. As expected, these patterns include the most common and ancient mechanistic step types introduced with the CATH 3.40.50.300 structure.

Table 6.1: Pattern 133, the mechanistic step types associated with CATH 3.20.20.70, Aldolase class I Mechanistic

Mechanistic step types with CATH 3.20.20.70, Aldolase class 1
Unimolecular elimination by the conjugate base
Redox
Radical termination
Radical formation
Proton transfer
Other tautomerisation
Intramolecular nucleophilic addition
Intramolecular elimination
Hydride transfer
Heterolysis
Electron transfer
Bimolecular nucleophilic substitution
Bimolecular nucleophilic addition
Bimolecular elimination
Assisted other tautomerisation
Assisted keto-enol tautomerisation
Aromatic unimolecular elimination by the conjugate base
Aromatic bimolecular nucleophilc addition
Aromatic bimolecular elimination
Aldol addition

The research goals of this work are not to explore mappings of mechanistic step types along metabolic pathways, as this would require one to unfold a complex network structure with graph theoretical approaches. However, in order to make explicit the complex recruitment patterns that are expected we have mapped H-level structures in the nucleotide interconversion pathway of purine metabolism [221], the oldest of all metabolic subnetworks defined by the KEGG database [129]. Since nucleotide interconversion precedes purine biosynthesis in evolution [221], we compared mechanistic step types associated with this pathway (Table 6.2). In MACiE, we found only 8 H-level structures involved in purine metabolism, ranging in *nd* value from 0 to 0.411. Remarkably, and despite the absence of MACiE entries for the most ancient enzymes of energy interconversion (EC 2.6.1.3. and EC 3.6.4.1), the results reveal the very early rise of the highly abundant pattern 4 in evolution and complex patterns of recruitment of additional chemistries which are ultimately associated with the combinatorics of mechanistic step types of Figure 6.5.

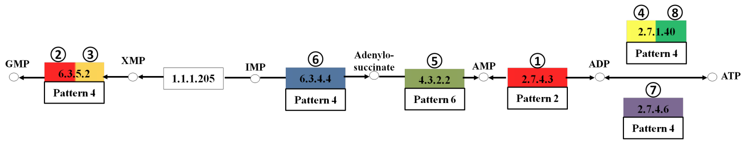
Table 6.2: MACiE Enzymes for Purine Metabolism.

MACiE	Enzyme name	EC	Subnetwork	PDB	CATH H level Structure	*nd value*	Combinatorial pattern	Mechanistic step types
M0290	adenylate kinase	2.7.4.3	INT	1zio	3.40.50.300	0	Pattern 2	Bimolecular nucleophilic substitution
M0234	GMP synthase (glutamine hydrolysing)	6.3.5.2	INT	1gpm	3.40.50.880	0.0980	Pattern 4(+2)	Proton transfer
								Bimolecular nucleophilic substitution
								Unimolecular elimination by the conjugate base
M0326	pyruvte kinase	2.7.1.40	INT	1pkn	3.20.20.60	0.1127	Pattern 4	Proton transfer
								Bimolecular nucleophilic substitution
M0326	pyruvte kinase	2.7.1.40	INT	1pkn	2.40.33.10	0.4118	Pattern 4	Proton transfer
								Bimolecular nucleophilic substituion
M0080	adenylosuccinate lyase	4.3.2.2	INT	1c3c	1.20.200.10	0.1667	Pattern 6	Proton transfer
								Bimolecular elimination
M0065	adenylosuccinate synthase	6.3.4.4	INT	1gim	3.40.440.10	0.2353	Pattern 4 (+2)	Proton Transfer
								Bimolecular nucleophilic substitution
								Assisted other tautomerisation
								Aromatic bimolecular nucleophilic substituion
M0150	nucleoside-diphosphate kinase	2.7.4.6	INT	1ndp	3.30.70.141	0.3186	Pattern 4	Proton transfer
								Bimolecular nucleophilic substitution

6.4 Conclusion

Contemporary protein structures consist of independently folding and compact domains that can be used as a fossil record of molecular evolution. We have utilised the available resources of enzyme mechanisms and the relative ages of CATH H-level domain structures to get a better insight into the natural history of biocatalytic mechanisms. Our analysis shows that the most designable structures (e.g., the α/β barrel and Rossmann fold) served as scaffolds to higher numbers of biochemical functions. The first two structures were responsible for introducing 35% (18/51) of the known catalytic step types described by the mechanistic step types. Over half of these appeared in the evolutionary timeline of domains before structures specific to Archaea, Bacteria and/or Eukarya [117], during a period of architectural diversification (nd <0.39). The most common mechanistic step types were also the most ancient and included fundamental building blocks of enzyme chemistry, 'proton transfer', 'bimolecular nucleophilic addition', 'bimolecular nucleophilic substitution', and 'unimolecular elimination by the conjugate base'. Later on in evolution, these mechanistic steps participated in a combinatorial interplay and were the highest represented in catalytic functions. The combination of 'bimolecular nucleophilic substitution' and 'proton transfer' was the most popular of all patterns of mechanistic step types. The other half of mechanistic step types appeared gradually after organismal diversification ($0.67 < nd < 1$) and during a period that spanned $\approx$ a billion years of evolutionary history.

Our phylogenomic approach is based on a census of protein domain structure in the proteomes of cellular organisms and the crucial axiom of polarization that claims that structural abundance increases in the course of evolution. This 'process' model of molecular accumulation in proteomes is based on Weston's generality criterion of homology and additive phylogenetic change [223] that in our case describes the slow and nested accumulation of homologous domain structures in the branches (proteome lineages) of the tree of life. A careful phylogenetic reconstruction analysis reveals that while both gains and losses of domain structures are frequent events, gains always overshadow losses in evolution [224]. They found that domain gains occurred throughout the evolutionary timeline albeit at a non-uniform rate. Noticeable, Nasir et al. [224] found that the gains-to-loss ratios increased with evolutionary time (fold age*nd*) and were relatively higher in the late evolutionary periods. The process has advantage to ensure that more domain gains availability to use combinatorial interplay that is responsible for the generation of novel domain architecture and further novel functions. This supports the general proportionality of domain abundance and evolutionary time of phylogenetic argumentation and the principle of continuity.



	CATH H-level Structure	Combinatorial patterns	*nd* values
1	3.40.50.300	Pattern 2	0
2	3.40.50.620	Pattern 4 (+2)	0.01471
3	3.40.50.880	Pattern 4 (+2)	0.09804
4	3.20.20.60	Pattern 4	0.11275
5	1.20.200.10	Pattern 6	0.16667
6	3.40.440.10	Pattern 4 (+2)	0.23529
7	3.30.70.141	Pattern 4	0.31863
8	2.40.33.10	Pattern 4	0.41177

Figure 6.6: Early evolution of mechanistic step types in the most ancient of all metabolic pathways. The diagram describes structural and functional innovation and recruitment of enzymes participating in the nucleotide interconversion (INT) pathway of the purine metabolism subnetwork of KEGG. The diagram shows that pattern 4 of possible mechanistic step type combinations is the most popular choice among the enzymes of this ancient pathway. Among the mechanistic step types in pattern 4, “Proton Transfer” is used by almost all the enzymes in the subnetwork (see Table 6.2). Annotated H-level structures associated with enzymatic activities are traced in the pathways with a color code according to their nd value, which is also given in table format together with CATH H-level code and mechanistic step type patterns. The most ancient enzymes exhibit a number of additional mechanistic step types that add to those of pattern 4. These additional mechanistic step types are listed in parentheses (+x, where x represents the number of additional types). For details of H-level structure and pattern association, see Table 6.1

6.5 Summary

In this chapter, we looked into the patterns using the definition of mechanistic step types from MACiE mapped onto the relative age of CATH H-level structure. A significant portion of this chapter is reprinted with permission from all co-authors in [4]. The analysis was performed by Neetika Nath, John BO Mitchell and Gustavo Caetano-Anollés.

7

Conclusion and Discussion

THE aim of the research described in this thesis is to investigate general trends in the diversity of biocatalytic mechanisms and their application towards function prediction, and understanding the evolution of biocatalysis. For this, we use chemoinformatics descriptors encoded by enzyme reactions. The data was retrieved from the MACiE database and was quantitatively represented as chemoinformatics descriptors, as discussed in Chapter 3, Data and Databases.

The final chapter of this thesis provides a discussion of the three major contributions to this work, which are explained in Chapters 4, 5 and 6. First, investigation of the functional properties of the enzyme mechanistic clusters is discussed in Section 7.1. Second, how well the enzyme mechanism can be used for function prediction is discussed in Section 7.2. Also, we discuss the results from the regression analysis in Section 7.3. Finally, we consider the adaptive nature of enzymatic reactions in Section 7.4.

7.1 Global Analysis of Enzyme Reaction Mechanisms

Determining the number of clusters and validating the results of various clustering algorithms are challenging tasks, especially when no prior information is provided. As no prior classification strategy was available for the enzyme mechanisms, we used different clustering methods to determine the number of clusters. Getting no obvious output from our analysis, we used an in-house algorithm: PFClust, designed by Mavridis, Nath and Mitchell [1] (discussed in Chapter 4) for clustering analysis of enzyme reaction mechanisms. To demonstrate the important functional attributes associated with members within clusters, we text-mined some of the biological attributes, such as mechanistic annotation from MACiE, metal cofactors from Metal MACiE. We compared the results between *OBC* clusters and *CBC* clusters to test for a strong association between the enzyme reactions and important biological attributes.

The result of this work suggests that the *CBC* descriptors cluster the data into significantly fewer clusters than *OBC* descriptors, suggesting that different functions tend to share similar mechanisms. We observe that enzymes often use different mechanistic steps to perform similar functions.

In this section, first we discuss the comparison of results from PFClust with six other state-of-the-art algorithms, and next we discuss the results of PFClust to cluster enzymatic reactions.

7.1.1 PFClust: Results and Discussion

We show that PFClust is able to cluster the CATH datasets a little better, on average, than any of the other algorithms, and furthermore is able to do this without the need to specify any external parameters. It is shown that PFClust can accurately group data according to their similarities without the need for any parameter tuning. Our clustering results on the CATH datasets show that PFClust provides structurally meaningful clusters. Also, that it performs best when compared to six other well-known clustering algorithms. Clustering protein domains using a density representation gives excellent agreement with the CATH part-manually curated classification.

7.1.2 Results From Mechanistic Annotation

Our motivation, here, is to perform a global study of enzyme reaction mechanisms and seek biological properties enriching clusters. This may provide important insights for better understanding of the diversity of chemical reactions of enzymes. We describe how the chemical mechanisms of enzyme reactions cluster in a space defined by chemoinformatics descriptors, using unsupervised global analysis.

In order to determine the number of clusters we designed a workflow (see Figure 4.7, Chapter 4) using *OBC* and *CBC* datasets. Ideally, one would expect to get an agreement between the algorithms used, which is supported by external or internal validation. Contrary to what is expected we got equivocal results from various state-of-the-art clustering algorithms; the results are shown in Figure 4.8. Thus, we decided to use the PFClust algorithm to seek patterns in the enzymatic reaction dataset because PFClust does not require any prior information as an input and its in-built validation step is efficient to optimise the results.

We found that when the enzyme reactions were described by *OBC* descriptors they formed a larger number of clusters than when clustered by *CBC* descriptors. For *OBC*, PFClust suggests 39 clusters and 57 singletons, whereas, *CBC* groups into 13 clusters and 18 singletons. There were eight MACiE reactions found to be singletons when clustered using either overall or composite enzymatic reaction.

Although the MACiE reactions are grouped into different clusters depending on the descriptor used, we found some biological features exclusively associated with particular clusters. For example, cluster 2 of *OBC* consists of enzymes belonging to the oxidoreducatase (EC 1) class of reactions. Another example is cluster 13 of *CBC*, where the members of this cluster use iron as a metal cofactor. The metal ion binding and heme binding GO molecular function occur in all the enzymes present in cluster 13 of *CBC*. All the clusters and their features are discussed in Appendix A.1.

The relationships between enzyme mechanisms and biological features within enzyme clusters help us to avoid the over-prediction of enzyme function, as well as guiding our decision-making in enzyme engineering. Also, such studies can generate more hypotheses to improve our knowledge of function annotation. This study can lead to very interesting questions, such as which pathways are these enzymes involved in, and do they have similar mechanistic steps or overall reactions?

7.2 Enzyme Function Prediction

Here we discuss the results from Chapter 5, based on the machine learning prediction. Our results strongly suggest that different enzymes typically bring about similar chemical transformations by dissimilar mechanisms. We conclude this because we find that the use of mechanistic information as a set of descriptors diminishes the EC prediction performance compared to descriptors encoding [31] information only on the overall transformation [33].

Almonacid et al. showed that, for convergently evolved pairs of enzymes sharing an EC sub-subclass, *OS* was almost universally higher than *MS* [33]. In this work, *OBC* descriptors can predict EC class with up to 68% accuracy, compared to 62% for the mechanism dependent *CBC* description of the reaction. Overall reaction similarity gives 71% prediction accuracy, mechanistic similarity only up to 60%. Thus, we find that the descriptor definitions based on overall reaction tend to be better predictors than those based on chemical mechanism, though *CBC* does well on the external test set. Since EC numbers are defined on the basis of the overall chemical transformation catalysed, the strong performance of overall reaction-based measures is reassuring - albeit that some questions arise over the congruence of the EC sub-subclass-based and descriptor-based definitions of a “similar reaction”. In this work, we are necessarily looking at predicting the top level EC class, since MACiE contains insufficiently many examples of each category at the subclass or lower levels. Hence, the overall reactions sharing the same label in this study are considerably less similar than those sharing third or fourth level labels.

Overall we found that *HD* descriptors strongly outperformed all other descriptors in predicting the classes EC 1, oxidoreductase, and EC 5, isomerase, as these descriptors were engineered to capture specific features from the respective EC classes. For example, in EC 5, out of 30 MACiE entries, 27 enzymes catalyse the simple stoichiometry of one starting material being transformed into one isomeric product. Interestingly the ranking of the descriptors, performances matches with the overall prediction performance (see Table 5.1). It was difficult to predict EC 2 when descriptors were encoded based on overall reaction as compared to when mechanistic information was added. This is in contrast to EC 5, where prediction performance was worst among all when descriptors represented mechanistic information. This suggested that mechanistic information based descriptors did not have enough information to be able to detect enzymes catalysing isomerase activity.

Nonetheless, the superiority of the predictions made using descriptors based on overall chemical transformation is entirely in accordance with the conclusions from [33]; mechanisms of analogous reactions tend to be less similar than are the overall reactions.

EC class 1, the oxidoreductases, covers a diversity of chemistry - the unifying feature being that all are redox reactions. Within the breadth of biological redox reactions, there are some recognisable clusters: significant proportions of these reactions involve interconversion of NAD and NADH, or NADP and NADPH (ten and 14, respectively, out of the 84 oxidoreductases in MACiE 3.0). Overall, the oxidation and reduction reactions of EC class 1 seem to leave recognisable chemical signatures in the descriptors; for instance C-H cleavage, O-H formation and C-C bond order changes are all common, and class 1 is generally well-recognised. Our chemical interpretations here, and indeed for all six classes, are based on analysis of descriptor values.

EC class 2, transferases, encompasses any chemical reaction that transfers a functional group from one molecule to another; quite commonly phosphate moieties, in the cases of kinases and phosphatases, or methyl groups are transferred. The 63 MACiE 3.0 entries in EC class 2 are diverse reactions, seeming to lack clear chemical patterns. Unsurprisingly, they are poorly predicted.

EC class 3, the hydrolases, is more tightly defined than many of the other classes, since it consists of reactions where water is used to hydrolyse a chemical bond. In fact, two of our 65 hydrolases are exceptions to this rule: M0226 is annotated as the reverse reaction, while M0172 is presented as utilising a hydroxide ion rather than neutral water. Almost half of the hydrolases in MACiE 3.0, 33 out of 73, catalyse the hydrolysis of biopolymers such as peptides, proteins, DNA or RNA. Hydrolases are well-predicted by all the descriptor sets, though less so for composite bond change descriptors. Hydrolysis leads to simple repeated and recognisable patterns of bond making and breaking. An example for the overall bond change is C-N single bond cleavage, combined with C-O and N-H single bond formation, for amide or peptide hydrolysis. These are recognisable from both overall reaction and, to a slightly lesser extent, mechanistic data. In the mechanistic case, the corresponding patterns also include bond changes which occur in one step of the mechanism and are subsequently undone in a later step.

EC class 4, lyases, includes those enzymes that catalyse the breaking of a covalent bond, other than by redox or hydrolysis reactions. While a typical textbook definition of a lyase may specify that there should be one substrate and two products, only 28 of the 49 lyases in MACiE 3.0 obey this rule. Six are presented as the reverse reaction, and as many as 15 present an assortment of stoichiometric or other complexities. Despite these extra challenges, lyases are generally well-recognised as overall reactions, primarily due to the prevalence of C-C single bond cleavage and C-H single bond formation.

EC class 5, isomerases, comprises enzymes that catalyse a reaction in which the product is an isomer of the starting material. Twenty seven of the 30 examples in MACiE 3.0 have the simple stoichiometry of one starting material being transformed into one isomeric product; 19 of the 30 enzymes catalyse constitutional isomerisation, seven are epimerases or racemases, two topoisomerases catalyse winding or unwinding of DNA, one enzyme is a cis - trans isomerase and one a tautomerase. We find that isomerases are well predicted by overall reaction descriptors, but are extremely hard for the mechanistic descriptors to predict. We interpret the lack of a relationship between membership of EC class 5 and mechanism as indicating that the class comprises a diversity of reactions, united only by the feature that the product is an isomer of the starting material. Thus, our results support the hypothesis that isomerisation reactions can evolve from mechanistically diverse starting points. The two overall reaction based descriptors do rather better, possibly because the reactions often involve formation or cleavage of O-H single bonds. Given knowledge of the definition of an isomerase as an enzyme whose substrate and product are isomers, it is a simple matter for a human to design a cheminformatics descriptor or descriptors to capture isomerisation reactions. The human designed descriptors were deliberately engineered to include the change in molecular mass between the largest substrate and the largest product. This descriptor is zero for all but one of the isomerases and allows this descriptor set to recognise isomerases with high accuracy. The isomerase most often incorrectly predicted by the human designed predictions is M0196, where the starting material and product are a (trivial) protonation state away from being isomers.

EC class 6, ligases, is composed of enzymes that catalyse the joining together of two molecules coupled with the conversion of ATP to AMP, or ATP to ADP. The human designed descriptors are chosen so that they specifically include a feature recognising ATP hydrolysis; this allows them to recognise ligases accurately. Ligases are characterised by both the formation and cleavage of P-O single bonds and we suggest this as the reason why both the overall and composite bond change descriptors do well in recognising ligases.

7.3 Application of Machine Learning Method

Here, we used cheminformatics descriptors to predict the solubility of drug-like molecules. Overall, results suggests that RF or SVR can provide a marginally better prediction of log S than the machine learning methods when cheminformatics descriptors are the sole input. We noticed that fitting the RF model on data that are scaled to a given mean and standard deviation produces a statistically significant improvement in its prediction with cheminformatics descriptors alone rather than theoretical energies (for detail see [3]). This suggests that slightly more useful information about the molecules' log S values is conveyed by the cheminformatics descriptors than by the theoretical descriptors alone. The joint results do present a statistically significant improvement for PLS and RF, once scaled by the mean/ standard deviation, compared to those for the theoretical energies alone. Additionally, we note that the RF method has produced promising predictions in this work, with relatively low RMSE.

7.4 History of Biocatalytic Mechanisms

Phylogenomic analysis of the occurrence and abundance of protein domains in proteomes has recently showed that the α/β architecture is probably the oldest fold design. This holds important implications for the origins of biochemistry. Here we explore structure-function relationships addressing the use of chemical mechanisms by ancestral enzymes. We test the hypothesis that the oldest folds used the most mechanisms. We start by tracing biocatalytic mechanisms operating in metabolic enzymes along a phylogenetic timeline of the first appearance of homologous superfamilies of protein domain structures from CATH. A total of 335 enzyme reactions were retrieved from MACiE and were mapped over fold age. We define a mechanistic step type as one of the 51 mechanistic annotations given in MACiE, and each step of each of the 335 mechanisms was described using one or more of these annotations. We find that the first two folds, the P-loop containing nucleotide triphosphate hydrolase and the NAD(P)-binding Rossmann-like homologous superfamilies, were α/β architectures responsible for introducing 35% (18/51) of the known mechanistic step types. We find that these two oldest structures in the phylogenomic analysis of protein domains introduced many mechanistic step types that were later combinatorially spread in catalytic history. The most common mechanistic step types included fundamental building blocks of enzyme chemistry: 'Proton transfer', 'Bimolecular nucleophilic addition', 'Bimolecular nucleophilic substitution', and 'Unimolecular elimination by the conjugate base' (for definition see Figure 3.3 in Chapter 3). They were associated with the most ancestral fold structure typical of P-loop containing nucleotide triphosphate hydrolases. Over half of the mechanistic step types were introduced in the evolutionary timeline before the appearance of structures specific to diversified organisms, during a period of architectural diversification. The other half unfolded gradually after organismal diversification and during a period that spanned $\approx$2 billion years of evolutionary history.

In these studies we trust the CATH classification scheme of domain structure, assignments of known structures to sequences, and current understanding of metabolic networks and associated chemical reactions. We note that it is highly likely that there is an 'underground' metabolism of weak catalytic specificities that is not annotated and involves a multiplicity of substrates and perhaps mechanistic step types. Our analysis is unable to capture this aspect of enzymatic function at this time. Similarly, our analysis does not explore biases in the distribution of annotations of molecular functions among structures and structures among functions nor the distribution of mechanisms across enzymatic reactions. Instead, it reveals patterns of accumulation of mechanistic step types in evolution.

7.5 Future Work

This thesis contains three main contributions are: a clustering algorithm, a machine learning method to annotate enzyme function, and an exploration of structure function relationship. Although the results presented in this thesis have demonstrated the effectiveness however, there is always extension of the method for new application in different domains. Here, we will discuss future extensions of these methods to other domains and also the challenges of their evaluation.

There is clearly much work to be done in the area of virtual screening. Perhaps the most direct extension of this work is by the means of using a more enzyme mechanistic knowledge analysed through clustering methods and evolutionary studies to express the properties of the vessels which helps the existing enzyme engineering methods. In following, I suggests the individually where I can extend my thesis work. And also I am suggesting the overall extension of my work in the industry and enzyme engineering studies.

- A clustering algorithm: a PFClust (Chapter 4) allows complete control over simulation and could be extended to explore what factors affects of the knowledge collected. In this work, we used mechanistic annotations of enzymes from MACiE database and clustered the enzymatic reactions. Such results have potential to create multiple hypothesis which will further generate multiple interesting enzyme function relationship. Also, PFClust is not limited to enzymatic reactions dataset, this algorithm can be used on any similarity matrix where the prior knowledge about cluster number is not available.

- A machine learning algorithm: Regarding potential extensions of the machine learning schema presented in Chapter 5, one can add additional information such as features from structure or sequence to improve the prediction pattern. Also, to make this method available publicly and easily available through web interface. Our machine learning method can also be extended to learning all gene products annotations, for example in the form of Gene Ontology terms and can be used to study the relationship between EC top class and enzyme function features from structure, sequence or GO terms.

- Evolution study of structure-function relationship: In this work (Chapter 6), we looked into the patterns using the definition of mechanistic step types from MACiE mapped onto the relative age of CATH H-level structure. Our analysis can be extended to understand patterns from metal as co-factors from the database metal-MACiE, in order to study the trend of metal as co-factor use in the enzymatic reaction before the Great Oxidation Event (GOE) [225].

One of the major extension to our work would be using this knowledge towards the improvement of enzyme engineering and incorporating such information into enzyme virtual screening and potential mapping of the location. For example, we know which catalytic properties/ entities in the reactions are common using clustering analysis also study of evolutionary suggests which folds provide store house for which mechanistic steps. Such information could be fruitful in shortlisting the potential candidates for mutation based enzyme engineering and fasten the process of screening.

In industry, many advancements occur, and are being currently developed, that take advantage of new computer architecture, database infrastructure and high throughput screening data. Combinatorial libraries and data mining are leading to new information being generated, often from old data. This kind of development may lead to new empirical models capable of fast, accurate predictions using existing applications.

7.6 Summary

Refinement of the methodologies of protein function can yield a massive amount of important information for better protein function prediction. Our knowledge and understanding of diversities in enzyme function will definitely improve the precision of function annotation with fewer false positives. This in turn will help to improve the computational methods for fast and accurate prediction methods.

To summarise this chapter, we have discussed our main findings on enzyme function. First, we found that different enzymes typically bring about similar chemical transformations by dissimilar mechanisms. Second, evidence was presented to support the assumption made for enzymatic mechanistic step types, that the older functions are most likely to be the most popular mechanistic steps.

I would like to conclude with a summary of the applicability of the methods developed in this thesis to other domains. The machine learning method is potentially relevant to any structured curation process in any domain. However, we designed our method specific to enzyme function annotation to understand the relationship between enzyme function with top level EC class. This method could also be of interest to other domains having regression analysis, pools of readily available unlabelled instances, identifying genes as a biomarker using RF method and where RF, SVN, KNN is the algorithm of choice.

A

Data and Tables

A.1 Results from Quantitative Global Analysis of Enzyme Reaction Mechanisms

Discussion on Clusters of *OBC* Descriptors

Here, we list the important features associated with each cluster output from PFClust for *OBC* descriptors. To evaluate the attributes in each cluster, we created a heatmap (see Figures 4.10 and 4.11) of 'mechanistic profile', which is discussed in Chapter 4; section 4.5 Evaluation Clustering Solutions. To understand the meaning in these clusters we used various biological resources, such as GO, KEGG pathways etc. We found that every cluster has special features that are distinct from the other clusters. Especially, we found that in most of the cases, GO annotation agreed with the cluster's EC top class function.

In this section, we discuss each cluster and what we find to be interesting features in each one. For *OBC* descriptor, PFClust produced 39 clusters and 57 singletons. The observation of clustering functions between enzymes that have been classified as structurally unrelated provides some of the most striking consequences of the evolutionary mechanisms.

These examples suggest that finding related information by nonhomologous enzymes can have an important practical application. For example, such information can be useful to identify off-targets for pharmaceuticals.

Characteristics for the *OBC* Cluster 1: this group consists of 38 MACiE enzymes that includes all entries of β-lactamase present in MACiE. In this cluster, most enzymes prefer water (KEGG C00001) during the reaction. This is an important feature of EC 3, hydrolases, thus most of the enzymes belongs to the hydrolases class. The second dominating EC class in this cluster are EC 2, transferases. In total, 38 MACiE enzymes including β-lactamase (M0015, M0016, M0210, M0257, M0258) are present. When investigated using GO annotations for molecular functions in this group, we found that hydrolase activity (GO:0016787) was very popular in this group. Mostly, they catalyse hydrolytic cleavage of carbon-oxygen (C.O), carbon-nitrogen (C.N), and carbon-carbon (C.C). No clear feature of metal-cofactor was registered, however, we found that 12 enzyme members in this group prefer zinc as a metal-cofactor for reaction. In this group, most popular mechanistic reactions found are bimolecular nucleophilic addition, proton transfer and unimolecular elimination by the conjugate base, which are among the top ten popular mechanistic step types in MACiE [4, 212].

Cluster 2: this group possesses six MACiE enzymes which are all oxidoreductase enzymatic reactions, EC 1. According to the EC classification system they share sub-subclass (1.1.1._), except M0100. This group's function is in full agreement with the GO annotation biological process: oxidoreductase activity (GO:0016491). Following is the list of MACiE enzyme entries in this group.

- M0007 EC 1.1.1.42 isocitrate dehydrogenase (NADP+)
- M0021 EC 1.1.1.38 malate dehydrogenase (oxaloacetate-decarboxylating)
- M0092 EC 1.1.1.22 UDP-glucose 6-dehydrogenase
- M0100 EC 1.2.1.8 betaine-aldehyde dehydrogenase
- M0255 EC 1.1.1.1 alcohol dehydrogenase
- M0256 EC 1.1.1.1 alcohol dehydrogenase

The outstanding feature noted in this group is the use of NADH (KEGG: C00004) as a substrate in the reaction. The common overall steps catalysed in this group are related to carbon, hydrogen, oxygen: C.C_2.1, C.H_0.1, C.C_1.2, C.O_1.2, C.N_2.1. Also, the common mechanistic step types are aromatic bimolecular nucleophilic addition , bimolecular elimination and hydride transfer.

Cluster 3: this cluster consists of 40 MACiE enzymes that are distributed mostly between EC 2, transferase, and EC 3, hydrolase. We found that this group is very diverse in terms of biological function; however, lipid metabolic process (GO:0006629) is the most popular choice of biological function. Out of 40, 36 MACiE enzymes preferred magnesium as a metal cofactor to affect the reaction. This is important regarding the effect of magnesium upon lipid metabolism as it is suggested [226] that acute deficiency of magnesium can effect the cholesterol levels. Most of the reactions use balance between formation or breaking of oxygen - hydrogen bonds and between oxygen - phosphate bonds, those are O.H_0.1, O.H_1.0, P.O_1.0, P.O_0.1. Mostly enzymes possess either ATP (KEGG: C00002), water (KEGG: C00001), DNA (KEGG: C00039) or RNA (KEGG: C00046) as a substrate to effect the reaction.

Cluster 4: consists of five enzymes belonging to the family of transferases (EC 2) listed below:

- M0022 EC 2.3.1.87 aralkylamine N-acetyltransferase: is an enzyme that is involved in the day/night rhythmic production of melatonin, by modification of serotonin.
- M0023 EC 2.1.1.20 glycine N-methyltransferase
- M0046 EC 2.1.1.72 site-specific DNA-methyltransferase
- M0048 EC 2.4.2.8 hypoxanthine phosphoribosyltransferase
- M0224 EC 2.3.1.48 histone

This group consists of enzymes with similar overall reactions that prefer reaction elements like nitrogen, hydrogen, oxygen; N.H_1.0 , C.N_0.1, S.H_0.1 and C.S_1.0. No similar GO annotation molecular or biological functions were found within this cluster.

Cluster 5: consists of 15 enzymes that are distributed between oxidoreductase reaction EC 1, hydrolase EC 3 and lyases EC 4 top class. In this group, most common GO molecular function annotation is lyase activity (GO:0016829), which is in concordance with 11 members of this cluster.

Cluster 6: possesses eight MACiE enzymes that are either oxidoreductase, EC 1 or lyases, EC 3. Only four out of eight enzymes have metal cofactor to effect the reaction, which are magnesium, iron and calcium (M0136 - acyl desaturase, M0184- pectate lyase, M0281 - electron transferring flavoprotein dehydrogenases and M0311 - phosphopyruvate hydratase). Notably, all of these enzymatic reactions use assisted keto enol tautomerisation mechanistic step type to complete the reaction.

Cluster 7: possesses four MACiE enzymes belonging to isomerase, EC 5, apart from M0204 (uroporphyrinogen-III synthase) which is a lyase EC 4. The GO molecular function annotations are : GO:0016866 intramolecular transferase activity, GO:0016853 isomerase activity, and GO:0003824 catalytic activity. This set of enzymes uses colligation and homolysis mechanistic step types.

Cluster 8: contains four enzymes where two of them belong to transferase, EC 2, and the rest to lyases, EC 4. This group also possesses assisted keto enol tautomerisation and aldol addition mechanistic steps. Enzymes in this group are involved in glycolysis (KEGG ec00010). The fructose aldolase enzymes of EC class 4 (M0052, fructose bisphosphate aldolase (Class II) and M0222 EC 4, fructose-bisphosphate aldolase (Class I)) are grouped with M0053 (malate synthase) and M0078 (citrate (Si)-synthase).

Cluster 9: consists of four enzymes, mostly transferase, EC 2, and hydrolases, EC 3. Members of the group of mechanistic step types claisen condensation are present with other popular enzymatic mechanistic step types: bimolecular nucleophilic addition and unimolecular elimination by the conjugate base. No other biological features were found to be significantly dominating in this group.

Cluster 10: group of five oxidoreductase, EC 1, reactions, which participates in different catalytic activity such as electron transport chain (GO:0022904), tetrahydrofolate metabolic process (GO:0046653). These enzymes are annotated with more than ten mechanistic step types among which seven are the mostly preferred step types such are electron transfer and aromatic bimolecular nucleophilic addition.

Cluster 11: consists of ten MACiE enzymes primarily involved in oxidoreductase, EC 1. Two of the members belong to isomerase activity, EC 5 (M0051, phosphoenolpyruvate carboxykinase (ATP) and M0190, isopentenyl-diphosphate delta-isomerase). According to GO molecular function annotation, two most popular annotations are oxidoreductase activity (GO:0016491) and FMN binding (GO:0010181). It is interesting to find that all the enzymes here catalyse on NAD (KEGG C00005) as a substrate. Mostly the reactions in this group use hydride transfer and aromatic unimolecular elimination by the conjugate base reaction step types.

Cluster 12: consists of only three members but are very interesting, as this is a collection of different overall reactions; M0082 (EC 2) glutamine-fructose-6-phosphate transaminase, M0095 (EC 5) arabinose isomerase and M0146 (EC 1) pyrogallol hydroxytransferase. One common factor among this group is that all enzymes are annotated with keto enol tautomerisation mechanistic step types to effect the reaction.

Cluster 13: is a collection of three enzymes which perform transfrases, EC 2, overall reaction. These enzymes participated in different biological processes such as blood coagulation and tissue regeneration. The mechanistic step types shared between these enzymes are the most popular mechanistic step types in the MACiE database such as proton transfer and bimolecular nucleophilic addition.

Cluster 14: is a set of three enzymes performing different overall reactions, EC top class oxidoreductase EC 1, transferase EC 2 and lyases EC 3. The reason why these enzymes are together is that all of them are participating in a reaction which requires for either breakage, formation or charge change of bonds of sulfur and oxygen; S.O_1.0, S.O_0.1, S.O_1.2.

Cluster 15: consists of six different enzymes, three of them are oxidoreductase EC 1, one transferase EC 2, and two isomerase EC 5. Among this, M0111 (glutamate synthase (ferredoxin)) and M0304 (glutamate synthase (NADPH)) have common biological processes, and those are glutamine metabolic and glutamate biosynthetic processes. Most common mechanistic step type is are intramolecular elimination.

Cluster 16: consists of nine ligase EC 6 and one transferase EC 2. The common factor in this group is that in their catalysis they use magnesium as a metal cofactor. The popular molecular function among all is ATP binding (GO: 0005524) and this also suggests why this group contains common substrate ATP (KEGG C00002). These enzymatic reaction types are aromatic bimolecular nucleophilic substitutions and bimolecular nucleophilic substitution.

Cluster 17: consists of 12 MACiE enzymes, mostly belonging to hydrolases EC 3 with two entries of transferase, EC 2 and one of lyases EC 4. This set of enzymes includes cyclomaltodextrin glucanotransferase, phospholipase A2. According to the GO biological process, carbohydrate metabolic and lipid catabolic are the most popular annotations. Such enzymes mostly catalyse reactions involving hydrogen, carbon and oxygen; H.O_0.1 , H.O_1.0 or C.O_0.1 and C.O_1.0. All the enzymatic reaction carries either proton transfer or unimolecular elimination by the conjugate base mechanistic step types. It is not a surprise to see that water (KEGG: C00001) is a common factor in this group as this cluster is dominated by hydrolases EC 3.

Cluster 18: a group of a four members of lyases EC 4. All of these enzymes perform different functions where the nucleotide sugar or fucose metabolic process is involved. In this cluster, proton transfer and assisted keto enol tautomerisation played vital roles for completing the reaction.

Cluster 19: all the four enzymes that are involved in peroxidase are grouped together in this cluster. This group of enzymes is important for understanding response to oxidative stress (GO: 0006979), according to GO biological process. For molecular function GO, we found all the enzymes are participated in peroxidase (GO: 0004601) and oxidoreductase activity (GO: 0016491). We found that six mechanistic step types were included in this set including acidic bimolecular nucleophilic substitution, bond order change, heterolysis, intermolecular nucleophilic substitution, redox and substitution reaction. Also, over represented substrates in this group are chloride (KEGG C00698) and hydrogen peroxidase (KEGG: C00027).

Cluster 20: consists of seven MACiE enzymes that are distributed among transferases EC 2 and lyases EC 4. In this group, magnesium is used by all the enzymes for the reactions. This group of enzymes mostly performs reactions on C-C or C-H bond formation. The mechanistic step types annotated in this group are intramolecular reactions such as intramolecular electrophilic addition/ substitution or elimination. In this group of enzymes most of them act on 2- trans-6- trans- farnesyl diphosphate (KEGG: C00448) substrate also supported by KEGG [1]. It is interesting to note that the PF-Clust was able to recognise them to be together.

Cluster 21: groups five enzymatic reactions where two enzymes belong to oxidoreductase EC 1, two are isomerase EC 5 and one EC 4 lyases. In

[1] http://www.genome.jp/dbget-bin/www$_{b}get?C$00448

this group, “assisted keto enol tautomerisation” step types are the popular choice for the enzymatic reaction.

Cluster 22: is a group of two MACiE enzymes, M0123 adenylyl sulfate reductase and M0279 phosphoadenylyl sulfate reductase (thioredoxin) belonging to oxidoreductase reaction. Both of these enzymes are involved in sulfur metabolism (KEGG ec00920)[2]. Where, M0123 participates in dissimilatory sulfate reduction and oxidation pathway, whereas M0279 in assimilarity sulfate reduction.

Cluster 23: all the eight members in this group belong to ligases family EC 6, and use zinc or magnesium as a metal cofactor. According to GO biological process, most of the enzymes in this group are involved in tRNA aminoacylation for protein translation (GO: 0006418)and according to molecular function it is ATP binding (GO:0005524). The most abundant mechanistic step types are assisted other tautomerisation and bimolecular nucleophilic substitution that are involved in bond breaking and formation between P.O, O.H and C.O. The most common substrate used within this group is ATP (C00002).

Cluster 24: consists of two members which belong to oxidoreductase, EC 1: M0130 (naphthalene 1,2-dioxygenase) and M0131 (4-hydroxybenzoate 3-monooxygenase). Both of the participants acquire same GO biological process annotation: aromatic compound catabolic process (GO:0019439). Common mechanistic step types are aromatic bimolecular nucleophilic addition, aromatic unimolecular elimination by the conjugate base and bimolecular nucleophilic substitution.

Cluster 25: consists of two enzymes belonging to EC 1 oxidoreductase reaction: M0067 (alanine dehydrogenase) and M0139 (xanthine dehydrogenase). In this group, aromatic bimolecular addition and bimolecular nucleophilic addition step types are annotated with both of the enzymes.

Cluster 26: this clusters two enzymes together, for both of which the overall reaction is oxidoreductase reaction EC 1: M0104 (quinoprotein glucose dehydrogenase) and M0208 (ubiquinol-cytochrome-c reductase), where M0104 uses calcium and M0208 uses iron as a metal cofactor to support the chemical transformation. In this group, only proton transfer is common in both the enzymatic reactions. These two enzymes share a substrate, namely ubiquinone (KEGG: C00399), to complete the reaction.

[2] http://www.genome.jp/kegg-bin/show$_{p}athway?ec$00920

Cluster 27: group of four enzymes which all belong to oxidoreductase reaction EC 1: M0003 (NAD(P)H dehydrogenase (quinone)), M0093 (hydroxymethylglutaryl-CoA reductase (NADPH)), M0109 (dihydroorotate oxidase) and M0227 (GDP-L-fucose synthase). In this group hydride transfer and proton transfer are common mechanistic step types used in the chemical reaction.

Cluster 28: group of two enzymes of EC 1 oxidoreductase reaction, M0106 (pyruvate dehydrogenase (acetyl-transferring)) and M0280 (3-methyl - 2 - oxobutanoate dehydrogenase (2- methylpropanoyl - transferring)). Both of these enzymes use magnesium as a metal cofactor to complete the reaction. From molecular function two distinct activities stand out, and these are protein binding (GO: 0005515) and oxidoreductase activity, acting on the aldehyde or oxo group of donors, disulfide as acceptor (GO:0016624).

Cluster 29: group of three enzymes belonging to transferases EC 2, M0148 (transaldolase) and M0289 (acetolactate synthase), and lyases EC 4, M0298 (tartronate - semialdehyde synthase). A common factor in this cluster is that both (M0289 and M0298) use magnesium as a metal cofactor. The most commonly used mechanistic step types are bimolecular nucleophilic addition and proton transfer.

Cluster 30: group of two different overall function enzymes, from transferase, M0031 (EC 2, thymidylate synthase) and isomerases, M0056 (EC 5, tRNA-pseudouridine synthase I). Interestingly, we found no common features that could be used to group these two enzymes together apart from PFClust suggestion.

Cluster 31: group of two enzymes, one belongs to transferases: M0030 (EC 2, formate C-acetyltransferase) and oxidoreductase: M0119 (EC 1, pyruvate). Similar to enzyme members in Cluster 30, we could not find any similar features in this cluster either.

Cluster 32: group of 19 enzymes mostly belonging to isomerase top class (EC 5), oxidoreductase (EC 1), hydrolase (EC 3) and ligases (EC 6). This group was also difficult to determine any common biological factors shared between all of its 19 enzymes.

Cluster 33: is a group of two enzymes belonging to EC top class 1, oxidoreductase, M0125 (catechol oxidase) and M0135 (peptidylglycine monooxygenase). Both of these enzymes preferred copper as a metal cofactor to complete the reaction. In this group, many common mechanistic step types were found and these are bimolecular homolytic addition, which is also a rare step in the MACiE database, radical formation, radical propagation, radical termination.

Cluster 34: group of four enzymes from different overall transaction oxidoreductase EC 1, hydrolase EC 3 and lyases EC 4. Out of four, three enzymes use iron metal-cofactor and M0284 (EC 3) uses cobalt as a cofactor.

Cluster 35: group of two enzymes belonging to the transferase (EC 2) class, M0008 (nicotinate-nucleotide diphosphorylase (carboxylating)) and M0147 (glycine hydroxymethyltransferase).

Cluster 36: group of two enzymes belongs to hydrolases EC 3 and lyases EC 4 class: M0060 (glucosamine-6-phosphate deaminase) and M0185 (DNA-(apurinic or apyrimidinic site) lyase) respectively. Both enzymes are annotated with similar mechanistic step types, and these are bimolecular electrophilic addition and bimolecular elimination.

Cluster 37: is a group of four MACiE enzymes; M0004 (nitrite reductase (NO-forming)), M0124 (cytochrome-c oxidase), M0138 (superoxide dismutase) and M0276 (nitrate reductase), all of which belong to EC 1, oxidoreductase reaction. The common metal-cofactors in this group are copper and iron. The common mechanistic step types in this group are electron transfer and proton transfer. In addition, these enzymes catalytic reactions on protons (KEGG: C00080).

Cluster 38: group of two enzymes, M0006 (glutathione-disulfide reductase) and M0277 (mercury(II) reductase), belonging to oxidoreductase reaction, EC 1. The specific biological process they both possess (GO:0045454) is cell redox homeostasis: by definition, it means any process that maintains the redox environment of a cell or compartment within a cell. In this cluster, we found two GO molecular functions terms common: flavin adnine dinucleotide binding and NADP binding.

Cluster 39: group of four enzymes; M0155 (formyl-CoA transferase, EC 2), M0198 (long-chain-fatty-acid-CoA ligase, EC 6), M0295 (methylated - DNA – [protein] - cysteine S - methyltransferase, EC 2) and M0307 (Ubiquitin transfer cascade (E1, E2, E3), EC 6).

Characteristics for the *CBC* When *CBC* descriptors were used for PF-Clust 15 clusters were found. Using mechanistic profile we found there are many strong signals suggesting association between the clusters and mechanistic annotation. For example, Cluster 13 have radial termination, radial formation, electron transformation and bond order change to be highly expressed annotations between M0124 (EC 1) cytochrome - c oxidase , M0239 (EC 1) peroxidase.

Here, we will discuss the important biological features found in the clusters of *CBC*.

Cluster 1: consists of 105 MACiE enzymes which are widely distributed among all six top classes of overall reaction. Most of the enzyme reactions in this group use the following reaction types: O.H_1.0, O.H_0.1, N.H_0.1, C.O_1.2, C.O_2.1, C.O_0.1, C.O_1.0, C.N_1.0, P.O_1.0, P.O_0.1, C.Cl_1.0, Mn.O_0.1, Mn.O_1.0. Some mechanistic step types such as bimolecular nucleophilic addition, bimolecular nucleophilic substitution, proton transfer and unimolecular elimination by the conjugate base also showed strong expression in this group. This group consists of 76% of the hydrolases activity (EC 3) from the MACiE database. Here, 34 MACiE enzymes possess metal ion binding GO molecular function, our analysis suggests that enzyme members in this group use magnesium and zinc as a metal co factor.

Cluster 2: group consists of 88 MACiE enzymes where mostly the enzymes belong to oxidoreductase (EC 1). In this group, 46% of EC 1 overall reactions from MACiE database are present. This group consists of aromatic bimolecular elimination and nucleophilic addition, mechanistic step types. Out of 88 MACiE enzymes, here 34 use metal cofactors which are magnesium, zinc and iron.

Cluster 3: consists of 29 enzymes mostly belonging to isomerase (EC 5). We found that, 40% of the isomerase EC 5 are present in this group. Most of the reaction use metal cofactor such as magnesium.

Cluster 4: possesses 21 MACiE enzymes. In this set of enzymes 90% of them perform proton transfer, bimolecular nucleophilic addition and unimolecular elimination by the conjugate base, whereas bimolecular nucleophilic substitution is only used by 33% of the enzymes. Moreover, clasien condenstation annotated in 4% of the enzymes in this cluster.

Cluster 5: consists of four enzymes in total, out of which three belong to oxidoreductase reaction EC 1 M0125 (catechol oxidase), M0135 (peptidylglycine monooxygenase), M0136 (acyl-[acyl-carrier-protein] desaturase), and one acts as isomerase, EC 5 M0192 (prostaglandin-E synthase). In this cluster, most distinct reaction steps involve oxygen with copper, iron and sulfur, and those are Fe.O_1.0, Fe.O_0.1, O.O_2.1, O.O_1.0, Cu.O_1.0, Cu.O_0.1, S.O_1.0, S.O_0.1. This set of enzymes is important for plant bioinformatics studies.

We found that, M0192 Prostaglandin E synthase (or PGE synthase) is an enzyme involved in eicosanoid and glutathione metabolism, a member of MAPEG family. The M0136 (acyl-[acyl-carrier-protein] desaturase) enzyme belongs to the family of oxidoreductases, specifically those acting on paired donors, with O_2 as oxidant and incorporation or reduction of oxygen. This enzyme class plays a critical role in the biosynthesis of unsaturated fatty acids in plants.

The M0125 (catechol oxidase), catechol is present in small quantities in the vacuoles of cells of many plant tissues. Catechol oxidase is present in the cell cytoplasm. If the plant tissues are damaged, the catechol is released and the enzyme converts the catechol to ortho-quinone, which is a natural antiseptic. Also, we found that M0125, M0135 and M0136 share similar substrate i.e. oxygen (C00007).

Cluster 6: consists of six MACiE enzymes distributed among oxidoreductases EC 1 (M0122: protein-methionine-S-oxide reductase, M0143: arsenate reductase and M0279: phosphoadenylyl-sulfate reductase (thioredoxin)), transferase EC 2 (M0153: thiosulfate sulfurtransferase and M0156: coenzyme-B sulfoethylthiotransferase) and isomerases EC 5 (M0191: protein disulfide-isomerase). Notably, these enzymes are annotated with bimolecular and intramolecular nucleophilic substitution mechanistic step types in order to execute the following reaction entities: S.H_0.1, S.H_1.0, S.S_0.1, S.S_1.0, S.O_1.0, S.O_2.1, S.O_0.1, As.O_1.0, As.O_2.1, As.S_0.1, As.S_1.0, Ni.C_0.1, Ni.C_1.0.

Most of the enzymes in this group possess thioredoxin (KEGG C00342) as a substrate to act on. Interestingly, all enzymes in this group work one of the following substrates: thioredoxin, thiosulfate, methylthio or disulfide. Apart from M0279 MACiE enzyme (appeared in *OBC* Cluster 22), the rest of the enzymes were found to be singletons when clustered using *OBC*.

Cluster 7: group of three MACiE enzymes including two with isomerase activity (EC 5, M0062(methylmalonyl-CoA mutase), M0063 (methylaspartate mutase)) and one transferase (EC 2, M0268 (methionine synthase)). Each of these three enzymes uses cobalt as a metal co-factor in their reactions. According to GO molecular function annotation, this cluster possesses cobalamin binding and metal ion binding molecular function.

Cluster 8: is a group of 29 MACiE entries where most of the enzymes act on phosphorus, oxygen, nitrogen and sulfur in order to effect the reaction: P.O_1.0, P.O_0.1, P.O_1.2, P.O_2.1, P.N_0.1, P.N_1.0, C.P_0.1, P.S_0.1, P.S_1.0. No oxidoreductase reaction was found in this class. The metal cofactors assisting the biochemical transformation for reactions in this group are either magnesium or zinc. Following are some important features, such as KEGG pathways, that are annotated by each enzyme in this group.

- M0023 EC Number: 2.1.1.20 glycine N-methyltransferase: is a foliate binding protein and it is found in abundant quantity in the liver [227]. This enzyme participates in glycine, serine and threonine metabolism.

- M0040 EC Number: 2.7.2.3 phosphoglycerate kinase: is an important ATP generating step in glycolysis. This enzyme catalyzes the reversible transfer of a phosphate group from 1 ,3-bisphosphoglycerate (1,3-BPG) to ADP producing 3-phosphoglycerate (3-PG) and ATP [228].

- M0042 EC Number: 3.1.30.2 nuclease: catalyzes the hydrolytic cleavage of DNA and RNA in the presences of metal cofactors [229].

- M0043 EC Number: 3.1.3.2 acid phosphatase : mostly present in lysosome [230].

- M0046 EC Number: 2.1.1.72 site-specific DNA-methyltransferase: an enzyme responsible for producing a species-characteristic methylation pattern on adenine residues in a specific short base sequence in the host cell DNA.

- M0047 EC Number: 3.1.3.48 protein-tyrosine-phosphatase: are a group of enzymes that remove phosphate groups from phosphorylated tyrosine residues on proteins [231].

- M0051 EC Number: 4.1.1.49 phosphoenolpyruvate carboxykinase (ATP): is an enzyme in the lyase family used in the metabolic pathway of gluconeogenesis [232].

- M0058 EC Number: 4.6.1.1 adenylate cyclase: All classes of AC catalyze the conversion of ATP to 3',5'-cyclic AMP (cAMP) and pyrophosphate [233].

- M0079 EC Number: 2.4.2.21 nicotinate - nucleotide - dimethylbenzimidazole phosphoribosyl transferase : it is one of the enzymes of the anaerobic pathway of cobalamin biosynthesis [234].

- M0088 EC Number: 2.7.7.12 UDP -glucose-hexose-1-phosphate uridylyl transferase: This enzyme belongs to the family of transferases, specifically those transferring phosphorus-containing nucleotide groups [235].

- M0101 EC Number: 3.6.1.29 bis(5'-adenosyl)-triphosphatase: specifically act on acid anhydrides in phosphorus-containing anhydrides [236].

- M0150 EC Number: 2.7.4.6 nucleoside-diphosphate kinase: are enzymes that catalyze the exchange of phosphate groups between different nucleoside diphosphates [237].

- M0152 EC Number: 2.7.8.7 holo-[acyl-carrier-protein] synthase: specifically those transferring non-standard substituted phosphate groups [238].

- M0157 EC Number: 3.1.2.6 hydroxyacylglutathione hydrolase: specifically the class of thioester lyases [239].

- M0178 EC Number: 3.6.3.14 H+-transporting two-sector ATPase: is one of the most putative proteins and its function as oxidative phosphorylation using ATP cofactor [211].

- M0179 EC Number: 3.6.4.9 chaperonin ATPase: is an enzyme with system name ATP phosphohydrolase (polypeptide-unfolding) which assists in protein folding [240].

- M0194 EC Number: 5.4.2.8 phosphomannomutase / phosphoglucomutase: This enzyme belongs to the family of isomerases, specifically the phosphotransferases (phosphomutases), which transfer phosphate groups within a molecule [241]. According to KEGG analysis, it is suggested that this enzyme is involved in Fructose and mannose metabolism (ec00051).

- M0198 EC Number: 6.2.1.3 long-chain-fatty-acid-CoA ligase: member of the ligase family that activates the breakdown of complex fatty acids [242].

- M0202 EC Number: 6.5.1.1 DNA ligase (ATP): specific type of enzyme, a ligase, (EC 6.5.1.1) that facilitates the joining of DNA strands together by catalyzing the formation of a phosphodiester bond [243].

- M0206 EC Number: 5.4.2.6 beta-phosphoglucomutase: to the family of isomerases, specifically the phosphotransferases (phosphomutases), which transfer phosphate groups within a molecule [244].

- M0242 EC Number: 3.1.4.41 sphingomyelin phosphodiesterase D: hydrolase the ester linkage between Cer phosphate and choline.

- M0246 EC Number: 2.7.10.1 receptor protein-tyrosine kinase

- M0271 EC Number: 5.4.2.9 phosphoenolpyruvate mutase: This enzyme belongs to the family of isomerases, which transfer phosphate groups within a molecule [245].

- M0287 EC Number: 2.7.7.4 sulfate adenylyltransferase: This enzyme belongs to the family of transferases, specifically those transferring phosphorus-containing nucleotide groups (nucleotidyltransferases) [246].

- M0290 EC Number: 2.7.4.3 adenylate kinase: this enzyme is good for study for extremophilic adaptive nature [247].

- M0295 EC Number: 2.1.1.63 methylated-DNA–[protein]-cysteine S-methyltransferase: This enzyme belongs to the family of transferases, specifically those transferring one-carbon group methyltransferases [248].

- M0296 EC Number: 2.7.7.39 glycerol-3-phosphate cytidylyltransferase: This enzyme belongs to the family of transferases, specifically those transferring phosphorus-containing nucleotide groups [249].

- M0299 EC Number: 2.7.7.3 pantetheine-phosphate adenylyltransferase : specifically those transferring phosphorus-containing nucleotide groups

- M0310 EC Number: 6.1.2.1 D-alanine-(R)-lactate ligase

Cluster 9: is a group of three enzymes belonging to oxidoreductase family (EC 1), including two chloride peroxidase (M0248 , M0250) enzymes and one alkanal monooxygenase M0132.

Cluster 10: includes five enzymes belonging to the oxidoreductase (EC 1) family, enzymes M0034 (catechol 2,3-dioxygenase), M0129 (taurine dioxygenase), M0133 (camphor 5-monooxygenase), M0134 (tyrosine 3-monooxygenase), M0137 (deacetoxycephalosporin-C synthase). All of these enzymes possess the following reaction steps: C.O_0.1, Fe.O_1.0, Fe.O_0.1, O.O_2.1 O.O_1.0, Fe.O_1.2, Fe.O_2.1, S.O_2.1, Fe.S_1.0.

M0034 EC 1 catechol 2,3-dioxygenase; This isolate exhibited important characteristics such as broad range of pH, temperature and time course for enzyme activity. M0129 EC 1 taurine dioxygenase. M0133 EC 1 camphor 5-monooxygenase. These enzymes possess iron as metal cofactors and moreover, catalysis on oxygen (KEGG C00007) is present in all these enzymes. When clustered with OBC, these enzymes were found to be singletons.

Cluster 11: is a group of three enzymes belonging to oxidoreductase reaction (EC 1); M0121 (sulfite oxidase), M0144 (arsenite oxidase) and M0276 (nitrate reductase). All these enzymes use molybdenum metal co-factor for completing the reactions. The common mechanistic step types in this group of enzymes are electron and proton transfer.

Cluster 12: this group of oxidoreductase overall reaction EC 1 of four MACiE enzymes; M0037 (prostaglandin-endoperoxide synthase), M0110 (D-amino-acid oxidase), M0113 (sarcosine oxidase) and M0130(naphthalene 1,2-dioxygenase).

Cluster 13: this group of two enzymes of oxidoreductase reaction (EC 1); M0124 (cytochrome-c oxidase) and M0239 (peroxidase). Both of these enzymes use iron as a metal cofactor to support the reaction. Also, according to GO annotation, metal ion binding and heme binding molecular function was retrieved from the cluster. Both of the enzymes were annotated with following mechanistic step types - bimolecular nucleophilic substitution, electron transfer, proton transfer, radical formation and redox.

A.2 Table for Chapter 6: Enzyme Function Evolution: *Chemolution Study*

Fold Age	CATH	Description	Mechanisms Discovered
0	3.40.50.300	P-loop containing nucleotide triphosphate hydrolases	Bimolecular nucleophilic addition
			Bimolecular nucleophilic substitution
			Intramolecular nucleophilic addition
			Proton transfer
			Unimolecular elimination by the conjugate base
			Electron transfer
0.0098	3.40.50.150	Vaccinia Virus protein VP39	Bimolecular elimination
0.0098	3.40.50.720	NAD(P)-binding Rossmann-like Domain	Bimolecular elimination
			Aromatic bimolecular nucleophilic addition
			Aromatic unimolecular elimination by the conjugate base
			Assisted keto-enol tautomerisation
			Aromatic intramolecular elimination
			Bimolecular homolytic addition
			Radical formation
			Radical termination
			Redox
			Bimolecular electrophilic addition
0.0098	3.50.50.60	FAD/NAD(P)-binding domain	Bimolecular elimination

			Aromatic bimolecular nucleophilic addition
			Aromatic unimolecular elimination by the conjugate base
			Assisted keto-enol tautomerisation
			Aromatic intramolecular elimination
			Bimolecular homolytic addition
			Radical formation
			Radical termination
			Colligation
			Redox
0.0147	3.40.50.620	HUPs	Intramolecular elimation
0,0196	3.20.20.70	Aldolase class 1	Heterolysis
			Aldol addition
			Assisted other tautomerisation
			Aromatic bimolecular elimination
			Other tautomerisation
0.0490	3.40.50.970	Not Assigned (1-deoxy D-xylulose-5-phosphate synthase -like domain 1/2/3)	Homolysis
			Elimination reaction
0.0490	3.40.190.10	Periplasmic binding protein-like 11	Aromatic bimolecular nucleophilic substitution
0.539	3.90.226.10	2-enoyl-CoA hydratase; chain A domain 1	Keto-Enol tautomerisation
			Intramolecular electrophilic addition
0.0588	3.40.47.10	Peroxisomal Thiolase; Chain A, domain 1	Claisen condensation
0.0588	3.40.30.10	Glutaredoxin	Intramolecular nucleophilic substitution
0.0686	3.60.21.10	Purple acid phosphatase; Chain A, domain 2	Coordination
0.0784	2.60.120.10	Jelly Rolls	Radical propagation

0.0784	3.40.50.1820	Not Assigned, 4,9-DSHA hydrolase activity, (Carboxyesterase-related protein -like domain 1)	Substitution reaction
0.1471	3.20.70.20	Anaerobic ribonucleotide-triphosphate reductase large chain	Bimolecular homolytic substitution
			Hydrogen transfer
			Unimolecular homolytic elimination
0.1765	1.10.600.10	Farnesyl diphosphate synthase	Intramolecular electrophilic substitution
			Intramolecular rearrangement
0.2059	2.40.100.10	Cyclophilin	Isomerisation
0.2549	3.40.50.10090	Not Assigned (Urophorphyrinogen -111 synthase - like domain 1/2)	Aromatic intramolecular electrophilic substitution
0.2745	3.30.1130.10	GTP Cyclohydrolase 1, domain 2	Amadori rearrangement
0.4412	1.10.520.10	Not Assigned (Catalase-preoxidase- like domain 1/2)	Bond order change
0.4902	3.40.50.10230	Precorrin-BX methylmutase CbiC/CobH	Sigmatrophic rearrangement
			Pericyclic reaction
0.5686	1.10.606.10	Vanadium-containing Chloroperoxidase domain 2	Acidic bimolecular nucleophilic substitution
0.5980	1.10.590.10	Chorismate Mutase subunit A	Claisen rearrangement
0.6373	3.20.20.240	TIM Barrel	Intramolecular homolytic addition
			Bimolecular homolytic elimination
0.6422	1.25.40.80	Serine threonine protein phosphatase 5, tetratricopeptide repeat	Photochemical activation
0.7304	1.10.800.10	Phenylalanine Hydroxylase	Aromatic bimolecular electrophilic addition

B

Data and R Code

All results are reproducible and the complete R scripts and example dataset are available as a supporting information for this thesis.

Folder in CD:

SI_Chapter_3

- mechanistic_step_type_proportion : This table contains information about mechanistic step type definitions, and the numbers and proportions of MACiE mechanisms that include each step type. The counts are from the complete MACiE data set (335 reaction mechanisms).
- Metal_macie : This table contains information about metal as a cofactor participating in enzymatic reactions. The counts are from the complete Metal MACiE data set (188 entries). The data is represented in Figure 3.5.
- Descriptors_Classification
  - Additional_File_1 : Details of the five sets of descriptors used in this work are listed in this file.
  - Additional_File_2_human_designed : The values of the *Human Designed* descriptors used in this work.
  - Additional_File_3_overall_bond_change : The values of the *Overall Bond Change* descriptors used in this work.
  - Additional_File_4_overall_reaction_similarity : The values of the *Overall Reaction Similarity* descriptors used in this work.
  - Additional_File_5_composite_bond_change : The values of the *Composite Bond Change* descriptors used in this work.

  - Additional_File_6_mechanistic_similarity : The values of the *Mechanistic Similarity* descriptors used in this work.

SI_Chapter_4

- Clustering_Analysis_cbc_obc : This table consists of assignment of MACiE enzymes to a cluster when using *CBC* or *OBC*, in addition to the detailed information of MACiE entries, such as EC top class, Enzyme names, Species.
- PFClust algorithm: also available for download from the Mitchell group web server[1].

SI_Chapter_5

- Informatics_Solubilty_datasets_and_scripts: Full descriptors and algorithm is also available for download from the Mitchell group web server[2].
  - Datasets
  - R_scripts_and_test_SI

SI_Chapter_6

- Dataset S1 : The complete data set used in our analysis in Chapter 6, where the first column represents the fold age (*nd* values),the second column is the H-level CATH code, and subsequent columns contain the CATH description, MACiE entry number, EC number, enzyme name. The MACiE entry numbers highlighted in red are the enzymes possessing metal cofactors.
- Table S1 : Patterns of mechanistic step types present in at least one entry in MACiE.
- Table S2 : Association between the CATH H-level structures and patterns of mechanistic step types. Patterns shared by more than one structure have their pattern numbers highlighted in green; patterns that are unique to one structure are not highlighted.

[1] http://chemistry.st-andrews.ac.uk/staff/jbom/group/PFClust.html
[2] http://chemistry.st-andrews.ac.uk/staff/jbom/group/Informatics_Solubility.html

Bibliography

[1] L. Mavridis, N. Nath, and J. B. O. Mitchell, "PFClust: a novel parameter free clustering algorithm," *BMC Bioinformatics*, vol. 14, p. 213, 2013.

[2] N. Nath and J. B. O. Mitchell, "Is EC class predictable from reaction mechanism?," *BMC Bioinformatics*, vol. 13, p. 60, Jan. 2012.

[3] J. L. McDonagh, N. Nath, L. De Ferrari, T. van Mourik, and J. B. O. Mitchell, "Uniting cheminformatics and chemical theory to predict the intrinsic aqueous solubility of crystalline druglike molecules," *Journal of Chemical Information and Modeling*, vol. 54, pp. 844–56, Mar. 2014.

[4] N. Nath, J. B. O. Mitchell, and G. Caetano-Anollés, "The natural history of biocatalytic mechanisms," *PLoS Computational Biology*, vol. 10, p. e1003642, May 2014.

[5] O. Nomenclature, C. Sciences, and Q. Mary, "International Union Of Biochemistry And Molecular," *Academic Press, London*, pp. 12–14.

[6] T. Hemalatha, T. UmaMaheswari, G. Krithiga, P. Sankaranarayanan, and R. Puvanakrishnan, "Enzymes in clinical medicine: an overview," *Indian Journal of Experimental Biology*, vol. 51, pp. 777–88, Oct. 2013.

[7] D. Petrey and B. Honig, "Is protein classification necessary? Toward alternative approaches to function annotation," *Current Opinion in Structural Biology*, vol. 19, pp. 363–8, June 2009.

[8] P. J. O. Brien and D. Herschlag, "Catalytic promiscuity and the evolution of new enzymatic activities," *Chemistry and Biology*, vol. 6, pp. R91–R105, April 1999.

[9] O. Khersonsky and D. S. Tawfik, "Enzyme promiscuity: a mechanistic and evolutionary perspective," *Annual Review of Bochemistry*, vol. 79, pp. 471–505, Jan. 2010.

[10] R. A. Jensen, "Enzyme recruitment in evolution of new function," *Annual Review Microbiology*, vol. 30, pp. 409–25, 1976.

[11] L. C. James and D. S. Tawfik, “Conformational diversity and protein evolution a 60-year-old hypothesis revisited,” *Trends in Biochemical Sciences*, vol. 28, pp. 361–368, July 2003.

[12] M. Yčas, “On earlier states of the biochemical system,” *Journal of Theoretical Biology*, vol. 44, pp. 145–160, Mar. 1974.

[13] G. Caetano-Anollés and D. Caetano-Anollés, “An evolutionarily structured universe of protein architecture,” *Genome Research*, vol. 13, pp. 1563–1571, Jan 2003.

[14] A. Mohammed and C. Guda, “Computational approaches for automated classififcation of enzyme sequence,” *Joournal of Proteomic Bioinformatics*, no. 402, pp. 147–152, 2011.

[15] R. Alterovitz, A. Arvey, S. Sankararaman, C. Dallett, Y. Freund, and K. Sjölander, “ResBoost: characterizing and predicting catalytic residues in enzymes,” *BMC Bioinformatics*, vol. 10, p. 197, Jan. 2009.

[16] G. J. Bartlett, C. T. Porter, N. Borkakoti, and J. M. Thornton, “Analysis of catalytic residues in enzyme active sites,” *Journal of Molecular Biology*, vol. 324, pp. 105–121, Nov. 2002.

[17] J. P. Nilmeier, D. a. Kirshner, S. E. Wong, and F. C. Lightstone, “Rapid catalytic template searching as an enzyme function prediction procedure,” *PloS ONE*, vol. 8, p. e62535, Jan. 2013.

[18] J. Kraut, “Serine proteases: structure and mechanism of catalysis,” *Annual Review of Biochemistry*, vol. 46, pp. 331–58, Jan. 1977.

[19] G. L. Holliday, J. B. O. Mitchell, and J. M. Thornton, “Understanding the functional roles of amino acid residues in enzyme catalysis,” *Journal of Molecular Biology*, vol. 390, pp. 560–77, July 2009.

[20] N. M. O. Boyle, G. L. Holliday, D. E. Almonacid, and J. B. O. Mitchell, “Using reaction mechanism to measure enzyme similarity,” *Journal of Molecular Biology*, pp. 1484–1499, 2007.

[21] N. Furnham, G. L. Holliday, T. A. P. de Beer, J. O. B. Jacobsen, W. R. Pearson, and J. M. Thornton, “The Catalytic Site Atlas 2.0: cataloging catalytic sites and residues identified in enzymes,” *Nucleic Acids Research*, vol. 42, pp. D485–9, Jan. 2014.

[22] S. Hunter, R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bork, U. Das, L. Daugherty, L. Duquenne, R. D. Finn, J. Gough, D. Haft, N. Hulo, D. Kahn, E. Kelly, A. Laugraud, I. Letunic, D. Lonsdale, R. Lopez, M. Madera, J. Maslen, C. McAnulla, J. McDowall, J. Mistry, A. Mitchell, N. Mulder, D. Natale, C. Orengo, A. F. Quinn, J. D. Selengut, C. J. A.

Sigrist, M. Thimma, P. D. Thomas, F. Valentin, D. Wilson, C. H. Wu, and C. Yeats, “InterPro: the integrative protein signature database,” *Nucleic Acids Research*, vol. 37, pp. D211–5, Jan. 2009.

[23] L. De Ferrari, S. Aitken, J. van Hemert, and I. Goryanin, “EnzML: multi-label prediction of enzyme classes using InterPro signatures,” *BMC Bioinformatics*, vol. 13, p. 61, Jan. 2012.

[24] C. Andreini, I. Bertini, G. Cavallaro, G. L. Holliday, and J. M. Thornton, “Metal-MACiE: a database of metals involved in biological catalysis,” *Bioinformatics (Oxford, England)*, vol. 25, pp. 2088–9, Aug. 2009.

[25] J. D. Fischer, G. L. Holliday, and J. M. Thornton, “The CoFactor database: organic cofactors in enzyme catalysis,” *Bioinformatics (Oxford, England)*, vol. 26, pp. 2496–7, Oct. 2010.

[26] K. N. Ferreira, T. M. Iverson, K. Maghlaoui, J. Barber, and S. Iwata, “Architecture of the photosynthetic oxygen-evolving center,” *Science (New York, N.Y.)*, vol. 303, pp. 1831–1838, Mar. 2004.

[27] J. J. Marsh and H. G. Lebherz, “Fructose-bisphosphate aldolases: an evolutionary history,” *Monographs of the Society for Research in Child Development*, vol. 79, pp. 138–46, Sept. 1992.

[28] A. Roy, J. Yang, and Y. Zhang, “COFACTOR: an accurate comparative algorithm for structure-based protein function annotation,” *Nucleic Acids Research*, vol. 40, pp. W471–7, July 2012.

[29] Y. Wang, K.-Y. San, and G. N. Bennett, “Cofactor engineering for advancing chemical biotechnology,” *Current Opinion in Biotechnology*, vol. 24, pp. 994–9, Dec. 2013.

[30] H.-F. Ji, L. Chen, and H.-Y. Zhang, “Organic cofactors participated more frequently than transition metals in redox reactions of primitive proteins,” *BioEssays : News and Reviews in Molecular, Cellular and Developmental Biology*, vol. 30, pp. 766–71, Aug. 2008.

[31] G. L. Holliday, C. Andreini, J. D. Fischer, S. A. Rahman, D. E. Almonacid, S. T. Williams, and W. R. Pearson, “MACiE: exploring the diversity of biochemical reactions,” *Nucleic Acids Research*, vol. 40, pp. D783–9, Jan. 2012.

[32] J. A. Gerlt, P. C. Babbitt, M. P. Jacobson, and S. C. Almo, “Divergent evolution in enolase superfamily: strategies for assigning functions,” *The Journal of Biological Chemistry*, vol. 287, pp. 29–34, Jan. 2012.

[33] D. E. Almonacid, E. R. Yera, J. B. O. Mitchell, and P. C. Babbitt, “Quantitative comparison of catalytic mechanisms and overall reactions in convergently evolved enzymes : implications for classification of enzyme function,” *PLoS Computational Biology*, vol. 6, no. 3, 2010.

[34] O. Sacher, M. Reitz, and J. Gasteiger, "Investigations of enzyme-catalyzed reactions based on physicochemical descriptors applied to hydrolases," *Journal of Chemical Information and Modeling*, vol. 49, pp. 1525–34, June 2009.

[35] D. A. R. S. Latino and J. Aires-de Sousa, "Assignment of EC numbers to enzymatic reactions with MOLMAP reaction descriptors and random forests," *Journal of Chemical Information and Modeling*, vol. 49, pp. 1839–46, July 2009.

[36] Q.-N. Hu, H. Zhu, X. Li, M. Zhang, Z. Deng, X. Yang, and Z. Deng, "Assignment of EC numbers to enzymatic reactions with reaction difference fingerprints," *PLoS ONE*, vol. 7, p. e52901, Jan. 2012.

[37] N. M. Allewell, "Thematic minireview series on enzyme evolution in the post-genomic era," *The Journal of Biological Chemistry*, vol. 287, pp. 1–2, Jan. 2012.

[38] A. E. Todd, C. A. Orengo, and J. M. Thornton, "Evolution of protein function, from a structural perspective," *Current Opinion in Chemical Biology*, vol. 3, pp. 548–56, Oct. 1999.

[39] O. Khersonsky, S. Malitsky, I. Rogachev, and D. S. Tawfik, "Role of chemistry versus substrate binding in recruiting promiscuous enzyme functions," *Biochemistry*, pp. 2683–2690, 2011.

[40] G. Dodson and A. Wlodawer, "Catalytic triads and their relatives," *Reviews*, vol. 0004, no. September, pp. 347–352, 1998.

[41] J. G. Duman, N. Li, D. Verleye, F. W. Goetz, D. W. Wu, C. A. Andorfer, T. Benjamin, and D. C. Parmelee, "Molecular characterization and sequencing of antifreeze proteins from larvae of the beetle Dendroides canadensis," *Journal of Comparative Physiology B*, vol. 168, no. 3, pp. 225–232, 1998.

[42] C. P. Ponting and R. R. Russell, "The natural history of protein domains," *Annual Review of Biophysics and Biomolecular Structure*, vol. 31, pp. 45–71, Jan. 2002.

[43] R. Rentzsch and C. A. Orengo, "Protein function prediction–the power of multiplicity," *Trends in Biotechnology*, vol. 27, pp. 210–9, Apr. 2009.

[44] M. E. Glasner, J. A. Gerlt, and P. C. Babbitt, "Evolution of enzyme superfamilies," *Current Opinion in Chemical Biology*, vol. 10, pp. 492–7, Oct. 2006.

[45] J. A. Gerlt and P. C. Babbitt, "Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies," *Annual Review of Biochemistry*, vol. 70, pp. 209–46, july 2001.

[46] E. V. Koonin, Y. I. Wolf, and G. P. Karev, “Scale-Free Evolution,” in *Power Laws, Scale-Free Networks and Genome Biology*, Molecular Biology Intelligence Unit, pp. 86–105, Boston, MA: Springer US, 2006.

[47] T. Linsky and W. Fast, “Mechanistic similarity and diversity among the guanidine-modifying members of the pentein superfamily,” *Biochimica et Biophysica Acta 1804*, vol. 1804, pp. 1943–53, Oct. 2010.

[48] S. D. Brown, J. A. Gerlt, J. L. Seffernick, and P. C. Babbitt, “A gold standard set of mechanistically diverse enzyme superfamilies,” *Genome Biology*, vol. 7, p. R8, Jan. 2006.

[49] E. V. Koonin, “Are there laws of genome evolution?,” *PLoS Computational Biology*, vol. 7, p. e1002173, Aug. 2011.

[50] B. Haegeman and J. S. Weitz, “A neutral theory of genome evolution and the frequency distribution of genes,” *BMC Genomics*, vol. 13, p. 196, Jan. 2012.

[51] J. Qian, N. M. Luscombe, and M. Gerstein, “Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model,” *Journal of Molecular Biology*, vol. 313, pp. 673–81, Nov. 2001.

[52] P. Sobrado, “Teaching principles of enzyme structure, svolution, and catalysis using bioinformatics,” *KBM Journal of Science Education*, vol. 1, pp. 7–12, 2010.

[53] Y. Ofran, M. Punta, R. Schneider, and B. Rost, “Beyond annotation transfer by homology: novel protein-function prediction methods to assist drug discovery,” *Drug Discovery Today*, vol. 10, pp. 1475–82, Nov. 2005.

[54] P. de Matos, J. A. Cham, H. Cao, R. Alcántara, F. Rowland, R. Lopez, and C. Steinbeck, “The Enzyme Portal: a case study in applying user-centred design methods in bioinformatics,” *BMC Bioinformatics*, vol. 14, p. 103, Jan. 2013.

[55] S. F. Altschup, C. Science, T. Pennsylvania, S. University, and U. Park, “Basic Local Alignment Search Tool,” *Journal of Molecular Biology*, no. 215, pp. 403–410, 1990.

[56] A. M. Schnoes, S. D. Brown, I. Dodevski, and P. C. Babbitt, “Annotation error in public databases: misannotation of molecular function in enzyme superfamilies,” *PLoS Computational Biology*, vol. 5, p. e1000605, Dec. 2009.

[57] B. Rost, “Enzyme function less conserved than anticipated.,” *Journal of molecular biology*, vol. 318, pp. 595–608, Apr. 2002.

[58] W. a. Koppensteiner, P. Lackner, M. Wiederstein, and M. J. Sippl, “Characterization of novel proteins based on known protein structures,” *Journal of Molecular Biology*, vol. 296, pp. 1139–52, Mar. 2000.

[59] K. Illergå rd, D. H. Ardell, and A. Elofsson, "Structure is three to ten times more conserved than sequence–a study of structural response in protein cores," *Proteins*, vol. 77, pp. 499–508, Nov. 2009.

[60] D. E. Almonacid and P. C. Babbitt, "Toward mechanistic classification of enzyme functions," *Current Opinion in Chemical Biology*, vol. 15, pp. 435–42, June 2011.

[61] P. F. Gherardini, M. N. Wass, M. Helmer-Citterich, and M. J. E. Sternberg, "Convergent evolution of enzyme active sites is not a rare phenomenon," *Journal of Molecular Biology*, vol. 372, pp. 817–45, Sept. 2007.

[62] L. De Ferrari and J. B. O. Mitchell, "From sequence to enzyme mechanism using multi-label machine learning," *BMC Bioinformatics*, vol. 15, no. 1, p. 150, 2014.

[63] Y. Loewenstein, D. Raimondo, O. C. Redfern, J. Watson, D. Frishman, M. Linial, C. Orengo, J. Thornton, and A. Tramontano, "Protein function annotation by homology-based inference," *Genome Biology*, vol. 10, p. 207, Jan. 2009.

[64] I. Friedberg, "Automated protein function prediction–the genomic challenge," *Briefings in Bioinformatics*, vol. 7, pp. 225–42, Sept. 2006.

[65] M. Y. Galperin and E. V. Koonin, *Functional genomics and enzyme evolution: Homologous and analogous enzymes encoded in microbial genomes*. Genetica, 2000.

[66] V. N. Bhatia, D. H. Perlman, C. E. Costello, and M. E. McComb, "Software tool for researching annotations of proteins: open-source protein annotation software with data visualization," *Analytical Chemistry*, vol. 81, pp. 9819–23, Dec. 2009.

[67] J. Minshull, J. E. Ness, C. Gustafsson, and S. Govindarajan, "Predicting enzyme function from protein sequence," *Current Opinion in Chemical Biology*, vol. 9, pp. 202–9, Apr. 2005.

[68] J. B. O. Mitchell, "Machine learning methods in chemoinformatics," *Wiley Interdisciplinary Reviews: Computational Molecular Science*, pp. n/a–n/a, Feb. 2014.

[69] J. C. Triviño and F. Pazos, "Quantitative global studies of reactomes and metabolomes using a vectorial representation of reactions and chemical compounds," *BMC Systems Biology*, vol. 4, p. 46, Jan. 2010.

[70] L. Han, J. Cui, H. Lin, Z. Ji, Z. Cao, Y. Li, and Y. Chen, "Recent progresses in the application of machine learning approach for predicting protein functional class independent of sequence similarity," *Proteomics*, vol. 6, pp. 4023–37, July 2006.

[71] D. Lee, O. Redfern, and C. Orengo, “Predicting protein function from sequence and structure,” *Nature Reviews. Molecular Cell Biology*, vol. 8, pp. 995–1005, Dec. 2007.

[72] J. Cheng, A. N. Tegge, P. Baldi, and S. Member, “Machine learning methods for protein structure prediction,” *IEE Reviews in Biomedical Engineering*, vol. 1, pp. 41–49, 2008.

[73] K. Kadam, S. Sawant, and U. Kulkarni-kale, “Prediction of protein function based on machine learning methods : an overview,” *In: Introduction to Sequence and Genome Analysis, iConcept Press Ltd., Hong Kong. (Accepted for publication)*, 2013.

[74] B. J. Lee, M. S. Shin, Y. J. Oh, H. S. Oh, and K. H. Ryu, “Identification of protein functions using a machine-learning approach based on sequence-derived properties,” *Proteome Science*, vol. 7, p. 27, Jan. 2009.

[75] K. Trivodaliev, A. Bogojeska, and L. Kocarev, “Exploring function prediction in protein interaction networks via clustering methods.,” *PloS one*, vol. 9, p. e99755, Jan. 2014.

[76] D. W. Huang, B. T. Sherman, and R. a. Lempicki, “Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources,” *Nature Protocols*, vol. 4, pp. 44–57, Jan. 2009.

[77] A. E. Todd, C. A. Orengo, and J. M. Thornton, “Plasticity of enzyme active sites,” *Trends in Biochemical Sciences*, vol. 27, pp. 419–26, Aug. 2002.

[78] B. Sterner, R. Singh, and B. Berger, “Predicting and annotating catalytic residues: an information theoretic approach,” *Journal of Computational Biology : A Journal of Computational Molecular Cell Biology*, vol. 14, pp. 1058–73, Oct. 2007.

[79] K.-C. Chou and Y.-d. Cai, “A novel approach to predict active sites of enzyme molecules,” *Proteins*, vol. 55, pp. 77–82, Apr. 2004.

[80] R. Greaves and J. Warwicker, “Active site identification through geometry-based and sequence profile-based calculations: burial of catalytic clefts,” *Journal of Molecular Biology*, vol. 349, pp. 547–57, June 2005.

[81] D. Devos and A. Valencia, “Practical limits of function prediction,” *Proteins*, vol. 41, pp. 98–107, Oct. 2000.

[82] C. Chothial and A. M. Lesk, “The relation between the divergence of sequence and structure in proteins,” *The EMBO Journal*, vol. 5, no. 4, pp. 823–826, 1986.

[83] P. C. Babbitt, “Definitions of enzyme function for the structural genomics era,” *Current Opinion in Chemical Biology*, vol. 7, pp. 230–237, April 2003.

[84] L. C. Borro, S. R. M. Oliveira, M. E. B. Yamagishi, A. L. Mancini, J. G. Jardine, I. Mazoni, E. H. D. Santos, R. H. Higa, P. R. Kuser, and G. Neshich, "Predicting enzyme class from protein structure using Bayesian classification," *Genetics and Molecular Research : GMR*, vol. 5, pp. 193–202, Jan. 2006.

[85] P. D. Dobson and A. J. Doig, "Predicting enzyme class from protein structure without alignments," *Journal of Molecular Biology*, vol. 345, pp. 187–99, Jan. 2005.

[86] J. Damborsky and J. Brezovsky, "Computational tools for designing and engineering enzymes," *Current Opinion in Chemical Biology*, vol. 19, pp. 8–16, Apr. 2014.

[87] T. M. Penning and J. M. Jez, "Enzyme redesign," *Chemical Reviews*, vol. 101, pp. 3027–46, Oct. 2001.

[88] J. Damborsky and J. Brezovsky, "Computational tools for designing and engineering biocatalysts," *Current Opinion in Chemical Biology*, vol. 13, pp. 26–34, Feb. 2009.

[89] D. A. Suplatov, W. Besenmatter, V. K. Svedas, and A. Svendsen, "Bioinformatic analysis of α/β-hydrolase fold enzymes reveals subfamily-specific positions responsible for discrimination of amidase and lipase activities," *Protein Engineering, Design & Selection: PEDS*, vol. 25, pp. 689–97, Nov. 2012.

[90] R. G. Alderson, L. D. Ferrari, L. Mavridis, J. L. Mcdonagh, B. O. John, and N. Nath, "Enzyme Informatics," *Current Topics in Medicinal Chemistry*, vol. 12, no. 17, pp. 1911–1923, 2012.

[91] N. U. Nair, C. A. Denard, and H. Zhao, "Engineering of enzymes for selective catalysis," *Current Organic Chemistry*, vol. 14, no. 217, pp. 1870–1882, 2010.

[92] S. J. Barrett and W. B. Langdon, "Advances in the application of machine learning techniques in drug discovery , design and development," tech. rep., 2005.

[93] D. Wishart, *Bioinformatics in drug development and assessment*, vol. 37. Mar. 2005.

[94] N. S. Buchan, D. K. Rajpal, Y. Webster, C. Alatorre, R. C. Gudivada, C. Zheng, P. Sanseau, and J. Koehler, "The role of translational bioinformatics in drug discovery.," *Drug Discovery Today*, vol. 16, pp. 426–34, May 2011.

[95] K. M. Koeller and C. H. Wong, "Enzymes for chemical synthesis," *Nature*, vol. 409, pp. 232–40, Jan. 2001.

[96] L. H. Greene, T. E. Lewis, S. Addou, A. Cuff, T. Dallman, M. Dibley, O. Redfern, F. Pearl, R. Nambudiry, A. Reid, I. Sillitoe, C. Yeats, J. M. Thornton, and C. A. Orengo, "The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution," *Nucleic Acids Research*, vol. 35, pp. D291–7, Jan. 2007.

[97] L. Lo Conte, B. Ailey, T. J. Hubbard, S. E. Brenner, A. G. Murzin, and C. Chothia, "SCOP: a structural classification of proteins database," *Nucleic Acids Research*, vol. 28, pp. 257–9, Jan. 2000.

[98] G. L. Holliday, D. E. Almonacid, G. J. Bartlett, N. M. O'Boyle, J. W. Torrance, P. Murray-Rust, J. B. O. Mitchell, and J. M. Thornton, "MACiE (Mechanism, Annotation and Classification in Enzymes): novel tools for searching catalytic mechanisms," *Nucleic Acids Research*, vol. 35, pp. D515–20, Jan. 2007.

[99] I. Schomburg, A. Chang, C. Ebeling, M. Gremse, C. Heldt, G. Huhn, and D. Schomburg, "BRENDA, the enzyme database: updates and major new developments," *Nucleic Acids Research*, vol. 32, pp. D431–3, Jan. 2004.

[100] E. Akiva, S. Brown, D. E. Almonacid, A. E. Barber, A. F. Custer, M. a. Hicks, C. C. Huang, F. Lauck, S. T. Mashiyama, E. C. Meng, D. Mischel, J. H. Morris, S. Ojha, A. M. Schnoes, D. Stryke, J. M. Yunes, T. E. Ferrin, G. L. Holliday, and P. C. Babbitt, "The Structure-Function Linkage Database," *Nucleic Acids Research*, vol. 42, pp. D521–30, Jan. 2014.

[101] N. Nagano, "EzCatDB: the Enzyme Catalytic-mechanism Database," *Nucleic Acids Research*, vol. 33, pp. D407–12, Jan. 2005.

[102] N. Furnham, I. Sillitoe, G. L. Holliday, A. L. Cuff, S. A. Rahman, R. A. Laskowski, C. a. Orengo, and J. M. Thornton, "FunTree: a resource for exploring the functional evolution of structurally defined enzyme superfamilies," *Nucleic Acids Research*, vol. 40, pp. D776–82, Jan. 2012.

[103] H. S. Kim, J. E. Mittenthal, and G. Caetano-anollés, "MANET : tracing evolution of protein architecture in metabolic networks," *BMC Bioinformatics*, vol. 13, pp. 1–13, 2006.

[104] N. M. Luscombe, D. Greenbaum, and M. Gerstein, "What is bioinformatics ? An introduction and overview," *Yearbook of Medical Informatics*, pp. 83–100, 2001.

[105] P. Li, J. O. Dada, D. Jameson, I. Spasic, N. Swainston, K. Carroll, W. Dunn, F. Khan, N. Malys, H. L. Messiha, E. Simeonidis, D. Weichart, C. Winder, J. Wishart, D. S. Broomhead, C. a. Goble, S. J. Gaskell, D. B. Kell, H. V. Westerhoff, P. Mendes, and N. W. Paton, "Systematic integration of experimental data and models in systems biology," *BMC Bioinformatics*, vol. 11, p. 582, Jan. 2010.

[106] G. L. Holliday, D. E. Almonacid, J. B. O. Mitchell, and J. M. Thornton, “The chemistry of protein catalysis,” *Journal of Molecular Biology*, vol. 372, pp. 1261–77, Oct. 2007.

[107] P. C. Babbitt and J. A. Gerlt, “Understanding enzyme superfamilies. Chemistry As the fundamental determinant in the evolution of new catalytic activities,” *The Journal of Biological Chemistry*, vol. 272, pp. 30591–4, Dec. 1997.

[108] G. L. Holliday, C. Andreini, J. D. Fischer, S. A. Rahman, D. E. Almonacid, S. T. Williams, and W. R. Pearson, “MACiE Version 3.0,” vol. 44, no. 0, p. 2013, 2013.

[109] B. Smith and A. Kumar, “Controlled vocabularies in bioinformatics: a case study in the gene ontology,” *Drug Discovery Today: BIOSILICO*, vol. 2, pp. 246–252, Nov. 2004.

[110] J. B. L. Bard and S. Y. Rhee, “Ontologies in biology: design, applications and future challenges,” *Nature Reviews. Genetics*, vol. 5, pp. 213–22, Mar. 2004.

[111] P. Khatri and S. Dr?ghici, “Ontological analysis of gene expression data: Current tools, limitations, and open problems,” *Bioinformatics*, vol. 21, no. 18, pp. 3587–3595, 2005.

[112] R. Caspi, T. Altman, R. Billington, K. Dreher, H. Foerster, C. a. Fulcher, T. a. Holland, I. M. Keseler, A. Kothari, A. Kubo, M. Krummenacker, M. Latendresse, L. a. Mueller, Q. Ong, S. Paley, P. Subhraveti, D. S. Weaver, D. Weerasinghe, P. Zhang, and P. D. Karp, “The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases,” *Nucleic Acids Research*, vol. 42, no. D1, pp. 459–471, 2014.

[113] C. Andreini, I. Bertini, G. Cavallaro, G. L. Holliday, and J. M. Thornton, “Metal ions in biological catalysis: from enzyme databases to general principles,” *Journal of Biological Inorganic Chemistry : JBIC : A Publication of the Society of Biological Inorganic Chemistry*, vol. 13, pp. 1205–18, Nov. 2008.

[114] A. C. R. Martin, “PDBSprotEC: a Web-accessible database linking PDB chains to EC numbers via SwissProt,” *Bioinformatics (Oxford, England)*, vol. 20, pp. 986–8, Apr. 2004.

[115] T. J. P. Hubbard, B. Ailey, S. E. Brenner, A. G. Murzin, and C. Chothia, “SCOP, Structural Classification of Proteins Database: applications to evaluation of the effectiveness of sequence alignment methods and statistics of

protein structural data," *Acta Crystallographica Section D Biological Crystallography*, vol. 54, pp. 1147–1154, Nov. 1998.

[116] G. Caetano-Anollés, H. S. Kim, and J. E. Mittenthal, "The origin of modern metabolic networks inferred from phylogenomic analysis of protein architecture," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, pp. 9358–63, May 2007.

[117] S. A. Bukhari and G. Caetano-Anollés, "Origin and evolution of protein fold designs inferred from phylogenomic analysis of CATH domain structures in proteomes," *PLoS Computational Biology*, vol. 9, p. e1003009, Mar. 2013.

[118] L. Alvarez-Cohen, M. Ashburner, I. K. O. Cann, E. F. DeLong, W. F. Doolittle, C. Fraser-Liggett, A. Godzik, J. Gordon, M. Riley, and M. Schmid, *The new science of metagenomics: revealing the secrets of our microbial planet.* The National Academies Press, 2007.

[119] P. Tarczy-Hornoch and M. Minie, *Medical informatics: bioinformatics challenges and opportunities.* New York, NY: Springer, 2005.

[120] M. Zakarya, I. U. Rahman, N. Dilawar, and R. Sadaf, "An integrative study on bioinformatics computing concepts, issues and problems," *International Journal of Computer Science*, vol. 8, pp. 330–339, nov 2011.

[121] N. Furnham, T. A. de Beer, and J. M. Thornton, "Current challenges in genome annotation through structural biology and bioinformatics," *Current Opinion in Structural Biology*, vol. 22, pp. 594–601, Oct. 2012.

[122] J. A. Gerlt and P. C. Babbitt, "Mechanistically diverse enzyme superfamilies: the importance of chemistry in the evolution of catalysis," *Current Opinion in Chemical Biology*, vol. 2, pp. 607–12, Oct. 1998.

[123] V. Egelhofer, I. Schomburg, and D. Schomburg, "Automatic assignment of EC numbers," *PLoS Computational Biology*, vol. 6, p. e1000661, Jan. 2010.

[124] W. Shannon, R. Culverhouse, and J. Duncan, "Analyzing microarray data using cluster analysis," *Pharmacogenomics*, vol. 4, pp. 41–52, Jan. 2003.

[125] D. Meyer and C. Buchta, *proxy: distance and similarity measures*, 2014. R package version 0.4-12.

[126] R Core Team, *R: a language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria, 2012. ISBN 3-900051-07-0.

[127] A. Cakmak, M. Kirac, M. R. Reynolds, Z. M. Ozsoyoglu, and G. Ozsoyoglu, "Gene Ontology-based annotation analysis and categorization of metabolic pathways," *19th International Conference on Scientific and Statistical Database Management (SSDBM 2007)*, pp. 33–33, July 2007.

[128] D. Binns, E. Dimmer, R. Huntley, D. Barrell, C. O'Donovan, and R. Apweiler, "QuickGO: a web-based tool for Gene Ontology searching," *Bioinformatics (Oxford, England)*, vol. 25, pp. 3045–6, Nov. 2009.

[129] M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, and M. Hattori, "The KEGG resource for deciphering the genome," *Nucleic Acids Research*, vol. 32, pp. D277–80, Jan. 2004.

[130] N. Grira, M. Crucianu, and N. Boujemaa, "Unsupervised and semi-supervised clustering: a brief survey," in *In A Review of Machine Learning Techniques for Processing Multimedia Content, Report of the MUSCLE European Network of Excellence (FP6)*, 2004.

[131] O. Gascuel and O. Martin, "Using repeated measurements to validate hierarchical gene clusters," *Bioinformatics*, pp. 1–7, 2007.

[132] D. Breitkreutz and K. Casey, "Clusterers : a comparison of partitioning and density-based algorithms and a discussion of optimisations,"

[133] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognition Letters*, vol. 31, pp. 651–666, June 2010.

[134] H.-P. Kriegel, P. Kroger, J. Sander, and A. Zimek, "Density-based clustering," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, pp. 231–240, May 2011.

[135] J. Handl, J. Knowles, and D. B. Kell, "Computational cluster validation in post-genomic data analysis," *Bioinformatics (Oxford, England)*, vol. 21, pp. 3201–3212, Aug. 2005.

[136] S. Datta and S. Datta, "Comparisons and validation of statistical clustering techniques for microarray gene expression data," *Bioinformatics*, vol. 19, pp. 459–466, Mar. 2003.

[137] P. Willett, J. M. Barnard, G. M. Downs, B. C. Information, U. Road, and S. Sheffield, "Chemical similarity searching," *Journal of Chemical Informatics Science*, vol. 38, pp. 983–996, Feb 1998.

[138] P. C. Boutros and A. B. Okey, "Unsupervised pattern recognition: an introduction to the whys and wherefores of clustering microarray data," *Briefings in Bioinformatics*, vol. 6, pp. 331–43, Dec. 2005.

[139] S. Galbraith, J. A. Daniel, and B. Vissel, "A study of clustered data and approaches to its analysis," *The Journal of Neuroscience*, vol. 30, pp. 10601–10608, Aug. 2010.

[140] S. Patel and K. Patnaik, "Analysis of clustering algorithms for MRImage segmentation using IQI," *Procedia Technology*, vol. 6, pp. 387–396, Jan. 2012.

[141] O. M. Rivera-borroto, M. Rabassa-gutiérrez, R. C. Grau-ábalo, Y. Marrero-ponce, J. Manuel, and G.-d. Vega, “Dunn s index for cluster tendency assessment of pharmacological data sets,” *Canadian Journal of Physiology and Pharmacology*, vol. 433, pp. 425–433, 2012.

[142] L. Rokach and O. Maimon, “Clustering Methods,” in *Data Mining and Knowledge Discovery Handbook*, ch. 15, pp. 322–351, Springer New York, second ed., 2010.

[143] H. Backlund, A. Hedblom, and N. Neijman, “DBSCAN a density-based spatial clustering of application with noise,” *Data Mining*, pp. 1–8, 2011.

[144] K. Musayeva, T. Henderson, J. B. Mitchell, and L. Mavridis, “PFClust: an optimised implementation of a parameter-free clustering algorithm,” *Source Code for Biology and Medicine*, vol. 9, no. 1, p. 5, 2014.

[145] G. Brock, V. Pihur, S. Datta, and S. Datta, “clValid , an R package for cluster validation,” *Journal Of Statistical Software*, no. March 2008, pp. 1–32, 2011.

[146] C. Hennig, *fpc: Flexible procedures for clustering*, 2014. R package version 2.1-7.

[147] V. Estivill-Castro, “Why so many clustering algorithms: a position paper,” *SIGKDD Explorations*, vol. 4, pp. 65–75, June 2002.

[148] S. Wagner and D. Wagner, “Comparing clusterings - an overview,” *Technical Report 2006-04, Faculty of Informatics, Universit?t Karlsruhe (TH)*, no. 001907, pp. 1–19, 2007.

[149] N. Bolshakova and F. Azuaje, “Cluster validation techniques for genome expression data,” *Signal Processing*, vol. 83, pp. 825–833, Apr. 2003.

[150] P. J. Rousseeuw, “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis,” *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, Nov. 1987.

[151] J. C. Dunn, “A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters,” *Journal of Cybernetics*, vol. 3, pp. 32–57, Jan. 1973.

[152] A. Agresti, *An introduction to categorical data analysis.* Wilet Interscience, seocnd edition ed., 2007.

[153] T. Chau and A. Wong, “Pattern discovery by residual analysis and recursive partitioning,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 11, no. 6, pp. 833–852, 1999.

[154] J. F. R. Iii, “Adjusted Chi-Square Statistics : application to clustered binary data in primary care,” *Analysis of Family Medicine*, vol. 2, pp. 201–203, June 2004.

[155] G. N. Lance and W. T. Williams, “A general theory of classificatory sorting strategies: hierarchical systems,” *The Computer Journal*, vol. 9, pp. 373–380, Feb. 1967.

[156] C.-p. Wei, Y.-h. Lee, and C.-m. Hsu, “Empirical comparison of fast clustering algorithms for large data sets,” *Proceedings of the 33rd Hawaii International Conference on System Science*, vol. 00, no. c, pp. 1–10, 2000.

[157] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, “A Density-Based algorithm for discovering clusters in large spatial databases with noise,” in *2nd Interntional Conference on Knowledge Discovery and Data Mining*, 1996.

[158] C. Fraley and A. E. Raftery, “Model-Based Clustering, discriminant analysis, and density estimation,” *Journal of the American Statistical Association*, vol. 97, pp. 611–631, June 2002.

[159] W. A. Fenton, R. A. Gravel, and D. S. Rosenblatt, “Disorders of propionate and methylmalonate metabolism,”

[160] A. M. Upton and J. D. McKinney, “Role of the methylcitrate cycle in propionate metabolism and detoxification in Mycobacterium smegmatis,” *Microbiology (Reading, England)*, vol. 153, pp. 3973–82, Dec. 2007.

[161] H. Renata, Z. J. Wang, and F. H. Arnold, “Expanding the Enzyme Universe: Accessing Non-Natural Reactions by Mechanism-Guided Directed Evolution,” *Angewandte Chemie International Edition*, no. 150, pp. n/a–n/a, 2015.

[162] T. Bray, A. J. Doig, and J. Warwicker, “Sequence and structural features of enzymes and their active sites by EC class,” *Journal of Molecular Biology*, vol. 386, pp. 1423–36, Mar. 2009.

[163] C. Z. Cai, L. Y. Han, Z. L. Ji, and Y. Z. Chen, “Enzyme family classification by Support Vector Machines,” *Proteins*, vol. 55, pp. 66–76, Apr. 2004.

[164] Y. Yamanishi, M. Hattori, M. Kotera, S. Goto, and M. Kanehisa, “E-zyme: predicting potential EC numbers from the chemical transformation pattern of substrate-product pairs,” *Bioinformatics (Oxford, England)*, vol. 25, pp. i179–86, June 2009.

[165] N. Furnham, J. S. Garavelli, R. Apweiler, and J. M. Thornton, “Missing in action: enzyme functional annotations in biological databases,” *Nature Chemical Biology*, vol. 5, pp. 521–5, Aug. 2009.

[166] J. A. Gerlt and P. C. Babbitt, “Can sequence determine function?,” *Genome Biology*, vol. 1, pp. 1–10, Nov 2000.

[167] A. Llinàs, R. C. Glen, and J. M. Goodman, “Solubility challenge: can you predict solubilities of 32 molecules using a database of 100 reliable measurements?,” *Journal of Chemical Information and Modeling*, vol. 48, pp. 1289–303, July 2008.

[168] The Goodman group, 2014.

[169] RSC.

[170] T. Howley, M. G. Madden, M.-L. OConnell, and A. G. Ryder, “The effect of principal component analysis on machine learning accuracy with high-dimensional spectral data,” *Knowledge-Based Systems*, vol. 19, pp. 363–370, Sept. 2006.

[171] H. Abdi, “Partial Least Squares (PLS) Regression,” *Encyclopedia for Research Methods for the Social Sciences*, pp. 1–7, 2003.

[172] S. Varma and R. Simon, “Bias in error estimation when using cross-validation for model selection,” *BMC Bioinformatics*, vol. 7, p. 91, Jan. 2006.

[173] R. Kohavi, “A study of cross-validation and bootstrap for accuracy estimation and model selection,” *International Joint Conference on Artificial Intelligence*, vol. 2, pp. 1137–1143, 1995.

[174] V. N. Vapnik, “An overview of statistical learning theory,” *IEEE Transactions on Neural Networks*, vol. 10, pp. 988–99, Jan. 1999.

[175] D. S. Palmer, N. M. O’Boyle, R. C. Glen, and J. B. O. Mitchell, “Random Forest models to predict aqueous solubility,” *Journal of Chemical Information and Modeling*, vol. 47, no. 1, pp. 150–8, 2006.

[176] Y.-d. Cai, X.-j. Liu, X.-b. Xu, and K.-c. Chou, “Support Vector Machines for prediction of protein subcellular location by incorporating quasi-sequence-order effect,” *Sciences-New York*, vol. 348, 2002.

[177] W. S. Noble and P. Street, “What is a Support Vector Machine?,” *Computational Biology*, vol. 24, no. 12, pp. 1565–1567, 2006.

[178] B. Laboratories, L. Technologies, and C. J. C. Burges, “A tutorial on Support Vector Machines for pattern recognition,” *Pattern Recognition*, vol. 167, pp. 121–167, 1997.

[179] B. Ustün, W. J. Melssen, and L. M. C. Buydens, “Visualisation and interpretation of Support Vector Regression models,” *Analytica Chimica Acta*, vol. 595, pp. 299–309, July 2007.

[180] D. Wu, V. N. Vapnik, and R. Bank, “Support Vector Machine for text categorization,” *Learning*, pp. 1–16, 1998.

[181] D. Basak, S. Pal, and D. C. Patranabis, “Support Vector Regression,” *Neural Information Processing Letters and Reviews*, vol. 11, no. 10, pp. 203–224, 2007.

[182] D.-S. Cao, Q.-S. Xu, Y.-Z. Liang, X. Chen, and H.-D. Li, “Prediction of aqueous solubility of druglike organic compounds using partial least squares, back-propagation network and support vector machine,” *Journal of Chemometrics*, no. April, pp. n/a–n/a, 2010.

[183] O. Ivanciuc, “Applications of Support Vector Machines in chemistry,” *Biochemistry*, vol. 23, pp. 291–400, 2007.

[184] A. Ben-Hur, C. S. Ong, S. Sonnenburg, B. Schölkopf, and G. Rätsch, “Support Vector Machines and kernels for computational biology,” *PLoS Computational Biology*, vol. 4, p. e1000173, Oct. 2008.

[185] L. Breiman, “Random Forests,” *Machine Learning*, pp. 5–32, 2001.

[186] A. Statnikov, L. Wang, and C. F. Aliferis, “A comprehensive comparison of Random Forests and Support Vector Machines for microarray-based cancer classification,” *BMC Bioinformatics*, vol. 9, p. 319, Jan. 2008.

[187] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston, “Random Forest: a classification and regression tool for compound classification and QSAR modeling,” *Journal of Chemical Information and Computer Sciences*, vol. 43, no. 6, pp. 1947–58, 2003.

[188] H. Pang, A. Lin, M. Holford, B. E. Enerson, B. Lu, M. P. Lawton, E. Floyd, and H. Zhao, “Pathway analysis using Random Forests classification and regression,” *Bioinformatics (Oxford, England)*, vol. 22, pp. 2028–36, Aug. 2006.

[189] A.-l. B. S. Janitza, “Overview of Random Forest methodology and practical guidance with emphasis on computational biology and bioinformatics,” *Technical Report*, no. 129, 2012.

[190] P. Cunningham and S. J. Delany, “k-Nearest Neighbour classifiers,” *Technical Report*, Aug. 2007.

[191] S. Wold and M. Sjostrom, “PLS-regression : a basic tool of chemometrics,” *Chemometrics and Intelligent Laboratory Systems*, pp. 109–130, 2001.

[192] A. Krishnan, L. J. Williams, A. R. McIntosh, and H. Abdi, “Partial Least Squares (PLS) methods for neuroimaging: a tutorial and review,” *Neuro Image*, vol. 56, pp. 455–75, May 2011.

[193] J. M. Andrade-garda, R. Boque-Marti, J. Ferre-Baldrich, and A. Carlosena-Subieta, *Partial Least-Squares Regression.* No. 10, 2 ed., 2009.

[194] M. Kuhn, “Building predictive models in R using the caret package,” *Journal Of Statistical Software*, vol. 28, no. 5, 2008.

[195] S. W. A. W. C. K. A. E. T. C. Z. M. Max Kuhn, Jed Wing and the R Core Team, *caret: Classification and Regression Training*, 2014. R package version 6.0-24.

[196] A. Liaw and M. Wiener, “Classification and regression by randomforest,” *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[197] A. Karatzoglou, A. Smola, K. Hornik, and A. Zeileis, “kernlab – an S4 package for kernel methods in R,” *Journal of Statistical Software*, vol. 11, no. 9, pp. 1–20, 2004.

[198] B.-H. Mevik, R. Wehrens, and K. H. Liland, *PLS: Partial Least Squares and Principal Component regression*, 2013. R package version 2.4-3.

[199] S. V. Stehman, “Selecting and interpreting measures of thematic classification accuracy,” *Remote Sensing of Environment*, vol. 62, pp. 77–89, Oct. 1997.

[200] J. Gorodkin, “Comparing two K-category assignments by a K-category correlation coefficient,” *Computational Biology and Chemistry*, vol. 28, pp. 367–74, Dec. 2004.

[201] J. Menke and T. Martinez, “Using permutations instead of student’s t distribution for p-values in paired-difference algorithm comparisons,” *IEEE International Joint Conference on Neural Networks*, vol. 2, pp. 1331–1335, 2004.

[202] A. Lusci, G. Pollastri, and P. Baldi, “Deep architectures and deep learning in chemoinformatics: the prediction of aqueous solubility for drug-like molecules,” *Journal of Chemical Information and Modeling*, vol. 53, pp. 1563–75, July 2013.

[203] C. L. Dupont, S. Yang, B. Palenik, and P. E. Bourne, “Modern proteomes contain putative imprints of ancient shifts in trace metal geochemistry,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, pp. 17822–7, Nov. 2006.

[204] H. F. Winstanley, S. Abeln, and C. M. Deane, “How old is your fold?,” *Bioinformatics (Oxford, England)*, vol. 21 Suppl 1, pp. i449–58, June 2005.

[205] M. Wang, Y.-Y. Jiang, K. M. Kim, G. Qu, H.-F. Ji, J. E. Mittenthal, H.-Y. Zhang, and G. Caetano-Anollés, “A universal molecular clock of protein folds and its power in tracing the early history of aerobic metabolism and planet oxygenation,” *Molecular Biology and Evolution*, vol. 28, pp. 567–82, Jan. 2011.

[206] E. Ferrada and A. Wagner, “Evolutionary innovations and the organization of protein functions in genotype space,” *PLoS ONE*, vol. 5, p. e14172, Nov. 2010.

[207] K. M. Kim and G. Caetano-Anollés, “Emergence and evolution of modern molecular functions inferred from phylogenomic analysis of ontological data,” *Molecular Biology and Evolution*, vol. 27, pp. 1710–33, July 2010.

[208] M. Wang, S. M. Boca, R. Kalelkar, J. E. Mittenthal, and G. Caetano-Anollés, “A phylogenomic reconstruction of the protein world based on a genomic census of protein fold architecture,” *Complexity*, vol. 12, no. 1, pp. 1–3, 2006.

[209] G. Caetano-Anollés, M. Wang, D. Caetano-Anollés, and J. E. Mittenthal, “The origin, evolution and structure of the protein world,” *The Biochemical Journal*, vol. 417, pp. 621–37, Feb. 2009.

[210] G. Caetano-Anollés and A. Nasir, “Benefits of using molecular structure and abundance in phylogenomic analysis,” *Frontiers in Genetics*, vol. 3, p. 172, Sep 2012.

[211] B.-G. Ma, L. Chen, H.-F. Ji, Z.-H. Chen, F.-R. Yang, L. Wang, G. Qu, Y.-Y. Jiang, C. Ji, and H.-Y. Zhang, “Characters of very ancient proteins,” *Biochemical and Biophysical Research Communications*, vol. 366, pp. 607–11, Feb. 2008.

[212] G. L. Holliday, J. D. Fischer, J. B. O. Mitchell, and J. M. Thornton, “Characterizing the complexity of enzymes on the basis of their mechanisms and structures with a bio-computational analysis,” *The FEBS Journal*, vol. 278, pp. 3835–45, Oct. 2011.

[213] Y. Kawamura, “Systematic analyses of P-loop containing nucleotide triphosphate hydrolase superfamily based on sequence, structure and function,” *Genome Informatics*, vol. 582, pp. 581–582, 2003.

[214] J. Raymond, J. L. Siefert, C. R. Staples, and R. E. Blankenship, “The natural history of nitrogen fixation,” *Molecular Biology and Evolution*, vol. 21, pp. 541–54, Mar. 2004.

[215] N. Latysheva, V. L. Junker, W. J. Palmer, G. A. Codd, and D. Barker, “The evolution of nitrogen fixation in cyanobacteria,” *Bioinformatics (Oxford, England)*, vol. 28, pp. 603–6, Mar. 2012.

[216] D. E. Canfield, A. N. Glazer, and P. G. Falkowski, “The evolution and future of Earth’s nitrogen cycle,” *Science (New York, N.Y.)*, vol. 330, pp. 192–6, Oct. 2010.

[217] Y. Yung and M. McElroy, “Fixation of nitrogen in the prebiotic atmosphere,” *Science, New Series*, vol. 203, no. 4384, pp. 1002–1004, 1979.

[218] A. J. Reid, C. Yeats, and C. A. Orengo, “Methods of remote homology detection can be combined to increase coverage by 10% in the midnight zone,” *Bioinformatics (Oxford, England)*, vol. 23, pp. 2353–60, Oct. 2007.

[219] A. Harrison, F. Pearl, R. Mott, J. Thornton, and C. Orengo, “Quantifying the similarities within fold space,” *Journal of Molecular Biology*, vol. 323, pp. 909–926, Nov. 2002.

[220] G. R. Stockwell and J. M. Thornton, “Conformational diversity of ligands bound to proteins,” *Journal of Molecular Biology*, vol. 356, pp. 928–44, Mar. 2006.

[221] G. Caetano-Anollés, M. Wang, and D. Caetano-Anollés, “Structural phylogenomics retrodicts the origin of the genetic code and uncovers the evolutionary impact of protein flexibility,” *PLoS ONE*, vol. 8, p. e72225, Aug. 2013.

[222] C. Nagao, N. Nagano, and K. Mizuguchi, “Relationships between functional subclasses and information contained in active-site and ligand-binding residues in diverse superfamilies,” *Proteins*, vol. 78, pp. 2369–84, Aug. 2010.

[223] P. H. Weston, *Indirect and direct methods in systematics.* New York: Columbia University Press, 1988.

[224] A. Nasir, K. M. Kim, and G. Caetano-Anollés, “Global patterns of protein domain gain and loss in superkingdoms,” *PLoS Computational Biology*, vol. 10, p. e1003452, Jan. 2014.

[225] C. L. Dupont, A. Butcher, R. E. Valas, P. E. Bourne, and G. Caetano-Anollés, “History of biological metal utilization inferred through phylogenomic analysis of protein structures,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, pp. 10567–72, June 2010.

[226] Y. Rayssiguier, E. Gueux, and D. Weiser, “Effect of magnesium deficiency on lipid metabolism in rats fed a high carbohydrate diet,” *The Journal of Nutrition*, vol. 111, pp. 1876–83, Nov. 1981.

[227] R. J. Cook and C. Wagner, “Glycine N-methyltransferase is a folate binding protein of rat liver cytosol,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 81, pp. 3631–4, June 1984.

[228] L. R. Chiarelli, S. M. Morera, P. Bianchi, E. Fermo, A. Zanella, A. Galizzi, and G. Valentini, “Molecular insights on pathogenic effects of mutations causing phosphoglycerate kinase deficiency,” *PloS ONE*, vol. 7, p. e32065, Jan. 2012.

[229] P. Friedhoff, B. Kolmes, O. Gimadutdinow, W. Wende, K. L. Krause, and A. Pingoud, “Analysis of the mechanism of the Serratia nuclease using site-directed mutagenesis,” *Nucleic Acids Research*, vol. 24, pp. 2632–9, July 1996.

[230] H. Bull, P. G. Murray, D. Thomas, A. M. Fraser, and P. N. Nelson, “Acid phosphatases,” *Molecular Pathology*, vol. 55, pp. 65–72, Apr. 2002.

[231] A. Alonso, J. Sasin, N. Bottini, I. Friedberg, I. Friedberg, A. Osterman, A. Godzik, T. Hunter, J. Dixon, T. Mustelin, L. Jolla, and L. Jolla, “Protein Tyrosine in the Human Genome,” *Cell Review*, vol. 117, pp. 699–711, 2004.

[232] A. Sudom, R. Walters, L. Pastushok, D. Goldie, L. Prasad, L. T. J. Delbaere, H. Goldie, A. Sudom, R. Walters, L. Pastushok, D. Goldie, L. Prasad, L. T. J. Delbaere, and H. Goldie, “Mechanisms of activation of phosphoenolpyruvate carboxykinase from escherichia coli by Ca 2 + and of desensitization by trypsin,” *Journal of Bacteriology*, vol. 185, 2003.

[233] C. Metabolic and L. Jolla, “Regulation of adenylate cyclase,” *Proceedings of the Nutrition Scociety*, vol. 44, pp. 157–165, 1985.

[234] B. Cameron, F. Blanche, M. C. Rouyez, D. Bisch, A. Famechon, M. Couder, L. Cauchois, D. Thibaut, L. Debussche, and J. Crouzet, “Genetic analysis, nucleotide sequence, and products of two Pseudomonas denitrificans cob genes encoding nicotinate-nucleotide: dimethylbenzimidazole phosphoribosyltransferase and cobalamin (5’-phosphate) synthase.,” *Journal of Bacteriology*, vol. 173, pp. 6066–73, Oct. 1991.

[235] A. Arabshahi, F. J. Ruzicka, S. Geeganage, and P. A. Frey, “Standard free energies for uridylyl group transfer by hexose-1-P uridylyltransferase and UDP-hexose synthase and for the hydrolysis of uridine 5’-phosphoimidazolate,” *Biochemistry*, vol. 35, pp. 3426–8, Mar. 1996.

[236] C. Hurtado, A. Ruiz, A. Sillero, and M. A. G. Sillero, “Triphosphate pyrophosphohydrolase in escherichia coli,” *Journal of Bacteriology*, vol. 169, no. 4, pp. 1718–1723, 1987.

[237] E. H. Postel and B. M. Abramczyk, “Escherichia coli nucleoside diphosphate kinase is a uracil-processing DNA repair nuclease,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, pp. 13247–52, Nov. 2003.

[238] R. S. Flugel, Y. Hwangbo, R. H. Lambalot, J. E. Cronan, and C. T. Walsh, “Holo-(acyl carrier protein) synthase and phosphopantetheinyl transfer in Escherichia coli,” *The Journal of Biological Chemistry*, vol. 275, pp. 959–68, Jan. 2000.

[239] A. M. Martins, P. Mendes, C. Cordeiro, and A. P. Freire, “In situ kinetic analysis of glyoxalase I and glyoxalase II in Saccharomyces cerevisiae,” *European Journal of Bochemistry / FEBS*, vol. 268, pp. 3930–6, July 2001.

[240] A. J. Afzal, A. Natarajan, N. Saini, M. J. Iqbal, M. Geisler, H. A. El Shemy, R. Mungur, L. Willmitzer, and D. A. Lightfoot, "The nematode resistance allele at the rhg1 locus alters the proteome and primary metabolism of soybean roots," *Plant Physiology*, vol. 151, pp. 1264–80, Nov. 2009.

[241] M. D. Small and M. Matheson, "Phoshpomannomutaase and phosphoglucomutase developing cassia corymbosa seeds," *Phytochemistry*, vol. 18, pp. 1147–1150, 1976.

[242] H. Li, E. M. Melton, S. Quackenbush, C. C. DiRusso, and P. N. Black, "Mechanistic studies of the long chain acyl-CoA synthetase Faa1p from Saccharomyces cerevisiae," *Biochimica et Biophysica Acta*, vol. 1771, pp. 1246–53, Sept. 2007.

[243] A. B. Shapiro, "Complete steady-state rate equation for DNA ligase and its use for measuring product kinetic parameters of NAD?-dependent DNA ligase from Haemophilus influenzae," *BMC Research Notes*, vol. 7, p. 287, Jan. 2014.

[244] J. Dai, L. Wang, K. N. Allen, P. Radstrom, and D. Dunaway-mariano, "Conformational cycling in -phosphoglucomutase catalysis : reorientation of the B-D-glucose 1.6-(bis) phosphate interediate," *Biochemistry*, vol. 45, pp. 7818–7824, april 2006.

[245] M. Sarkar, C. J. Hamilton, and A. H. Fairlamb, "Properties of phosphoenolpyruvate mutase, the first enzyme in the aminoethylphosphonate biosynthetic pathway in Trypanosoma cruzi," *The Journal of Biological Chemistry*, vol. 278, pp. 22703–8, June 2003.

[246] Z. Reuveny, D. K. Dougall, and P. M. Trinity, "Regulatory coupling of nitrate and sulfate assimilation pathways in cultured tobacco cells," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 77, pp. 6670–2, Nov. 1980.

[247] M. Davlieva and Y. Shamoo, "Structure and biochemical characterization of an adenylate kinase originating from the psychrophilic organism Marinibacillus marinus," *Acta Crystallographica. Section F, Structural Biology and Crystallization Communications*, vol. 65, pp. 751–6, Aug. 2009.

[248] K. Lim, T. J. Park, and W. K. Paik, "Phosphorylation of methylated-DNA-protein-cysteine S-methyltransferase at serine-204 significantly increases its resistance to proteolytic digestion," *Biochemical Journal*, vol. 808, pp. 801–808, 2000.

[249] A. N. Mericl and J. A. Friesen, "Comparative kinetic analysis of glycerol 3-phosphate cytidylyltransferase from Enterococcus faecalis and Listeria monocytogenes," *Medical Science Monitor : Iternational Medical Journal of Experimental and Cinical Research*, vol. 18, pp. BR427–34, Nov. 2012.