RESEARCH NOTE

Decomposition of mutational context signatures using quadratic programming methods [version 1; referees: 1 approved, 1 approved with reservations]

Andy G. Lynch

Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK

v1 **First published:** 07 Jun 2016, **5**:1253 (doi: 10.12688/f1000research.8918.1)
Latest published: 07 Jun 2016, **5**:1253 (doi: 10.12688/f1000research.8918.1)

Abstract

Methods for inferring signatures of mutational contexts from large cancer sequencing data sets are invaluable for biological research, but impractical for clinical application where we require tools that decompose the context data for an individual into signatures. One such method has recently been published using an iterative linear modelling approach. A natural alternative places the problem within a quadratic programming framework and is presented here, where it is seen to offer advantages of speed and accuracy.

This article is included in the RPackage gateway.

Open Peer Review

Referee Status: ? ✓

	Invited Referees 1	Invited Referees 2
version 1 published 07 Jun 2016	? report	✓ report

1 **Mohamed Helmy**, University of Toronto Canada

2 **Miguel Vazquez**, Spanish National Cancer Research Centre (CNIO) Spain

Discuss this article

Comments (0)

Corresponding author: Andy G. Lynch (andy.lynch@cruk.cam.ac.uk)

Competing interests: No competing interests were declared.

How to cite this article: Lynch AG. **Decomposition of mutational context signatures using quadratic programming methods [version 1; referees: 1 approved, 1 approved with reservations]** *F1000Research* 2016, **5**:1253 (doi: 10.12688/f1000research.8918.1)

Grant information: AGL was supported in this work by a Cancer Research UK programme grant [C14303/A20406] to Simon Tavaré. AGL acknowledges the support of the University of Cambridge, Cancer Research UK and Hutchison Whampoa Limited. Whole-genome sequencing of oesophageal adenocarcinoma was part of the oesophageal International Cancer Genome Consortium (ICGC) project. The oesophageal ICGC project was funded through a programme and infrastructure grant to Rebecca Fitzgerald as part of the OCCAMS collaboration.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

First published: 07 Jun 2016, **5**:1253 (doi: 10.12688/f1000research.8918.1)

Introduction

The existence of context-specific DNA mutational signatures as a response to carcinogens has been known for some time (see e.g. Pfeifer *et al.*[1]), but the last three years have seen progress to bioinformatic inference of mutational signatures from large scale cancer sequencing studies[2–4] such as TCGA (http://cancergenome.nih.gov/) and ICGC (icgc.org).

These methods of signature discovery, while important, do not translate to clinical application. First of all, they are reliant on a large corpus of samples for their efficacy, making them impractical to be run repeatedly for each new patient. Secondly, even with a large corpus, the results for one individual can theoretically change depending on the identities of the other patients in the corpus, which is undesirable in practice. Therefore there is great value in methods such as those recently presented by Rosenthal *et al.*[5] that can, for a single sample, break a vector of observed mutation counts into constituent signature components.

In the Cancer Research UK funded oesophageal adenocarcinoma ICGC project we have taken a similar view to Rosenthal *et al.*[5] for the decomposition of a single sample, but rather than decomposing mutational contexts into signatures by fitting iterative linear models (ILM), we have viewed the question as lying within the framework of quadratic programming (QP). By mutational contexts, we commonly mean the 96 trinucleotide contexts consisting of the 6 distinguishable mutations and the 16 combinations of immediately preceding and following bases. More general definitions are possible[3] and can be accommodated in both the QP and ILM approaches, but we assume the standard 96 in what follows.

Methods

In brief, we want to minimize the difference between the normalized observed vector of mutation contexts m (a 96 × 1 vector) and $\mathbf{S}w$ (where $\mathbf{S}$ is a 96 × k matrix, each column of which represents the contributions of mutational contexts to one signature, k is the number of known mutational signatures, and w is a k × 1 matrix of weights to be estimated). Our problem, then, is to:

$$\begin{aligned}\text{minimize } & (m-\mathbf{S}w)^T(m-\mathbf{S}w)\\ & = m^T m - w^T\mathbf{S}^T m - m^T\mathbf{S}w + w^T\mathbf{S}^T\mathbf{S}w\\ \text{subject to } & \sum_j w_j = 1, w_j \geq 0\end{aligned}$$

which is equivalent to:

$$\begin{aligned}\text{minimize } & -m^T\mathbf{S}w + \frac{1}{2}w^T\mathbf{S}^T\mathbf{S}w\\ \text{subject to } & \sum_j w_j = 1, w_j \geq 0\end{aligned}$$

which is the classical quadratic programming problem that can be solved quickly (given the form of $S^T S$) and easily using the core linear algebra functionality of R (version 3.2.4)[6] and the quadprog package (version 1.5-5)[7], which implements the dual method of Goldfarb and Idnani[8,9] to find the solution. Practical details of the implementation can be found in the 'Data and Software Availability' section of this note.

Results

Dataset 1. An R Markdown document that when compiled will reproduce all the results presented

http://dx.doi.org/10.5256/f1000research.8918.d124181

In most circumstances, both the ILM and QP approaches work well. Illustrating them on an example from the OCCAMS consortium's whole-genome sequencing of oesophageal adenocarcinoma[10], we see that the ILM and QP approaches are highly concordant (See Figure 1). The ILM approach has the advantages of familiarity of interpretation, and enforcement of parsimony should this be desired (while parsimony is generally desirable if building a predictor, if we are trying to model an underlying truth then it represents a strong assumption). More importantly, taking advantage of the linear modelling framework, it would be easy to generalize this approach to use other error models or to include additional structure should one e.g. wish to simultaneously investigate several related samples.

The disadvantage of the ILM approach comes from its having to define a subset of signatures to include in the model. While the signature matrix is of full rank, with noise in the system it is sometimes possible to approximate an observed vector with several different linear combinations of signatures, and an ILM approach is not guaranteed to give consideration to the correct combination of signatures. Even if the correct solution is reached, it can be a substantially slower approach. It is not difficult to simulate a combination of signatures that takes thousands of iterations and thousands of times longer to run than the QP approach.

If one simulates a flat combination of all available signatures, then the ILM approach performs worse than the QP approach. A fairer comparison would be to consider all equal combinations of just two signatures (with noise added). Of 351 possible such combinations using the Nature 2013 signature set[2,5], the majority are well inferred using both the ILM and QP approaches, while one (the combination of signatures 1B and 3) performs poorly for both methods. Aside from these, there is a definite set of combinations for which the ILM approach performs markedly worse than the QP approach (See Figure 2). Pairs involving signature 1B, or signature 5, appear to cause the most problems. It is not the case that the problematic pairs are themselves highly correlated, but the 1B and U2 signatures are, possibly explaining the outlying nature of the U2-R2 pair. This exercise took approximately 5 seconds using

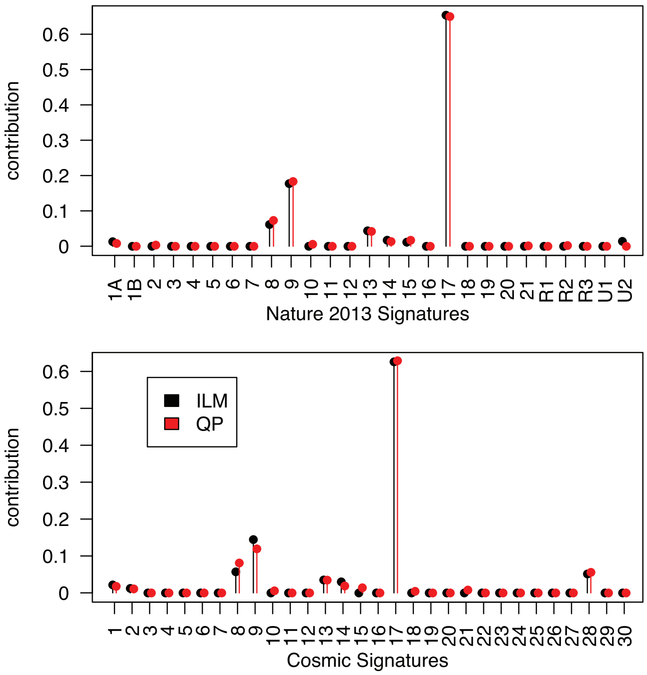

Figure 1. Performance of ILM and QP methods on oesophageal adenocarcinoma whole-genome sequencing data. 18, 916 SNVs from sequencing library SS6003314 (tumour) compared to library SS6003313 (matched normal tissue)[10] are considered. Using the two signature sets included with the deconstructSigs package (Top: the original Nature 2013 signatures[2]. Bottom: the COSMIC[11] signatures) both methods identify the same signatures as being active and produce estimates of contribution weight that are remarkably similar. Note that we are not adjusting for frequencies of contexts in the genome in these analyses.

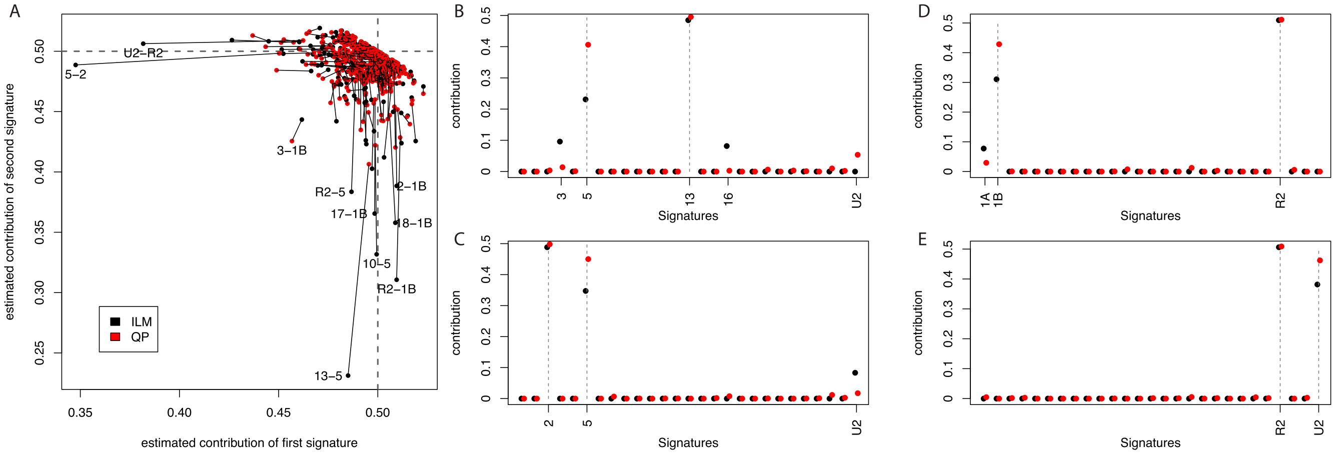

Figure 2. Performance of ILM signature deconstruction methods with simulated data. A. 351 simulated datasets were constructed, one for each possible pair of the 27 Nature 2003 signatures, with equal weighting given to both of the signatures and independent uniform errors applied to each mutational context count (ranging from –5% to +5%). The contributions for the two signatures that should be detected are illustrated, with a line linking the estimates from the ILM and QP methods. Perfect performance would see contributions of 0.5 estimated for both signatures in all cases. The identities of outlying signature-pairs are indicated. **B**. The contributions estimated from the combination of signatures 13 and 5. **C**. The contributions estimated from the combination of signatures 2 and 5. **D**. The contributions estimated from the combination of signatures 1B and R2. **E**. The contributions estimated from the combination of signatures R2 and U2. In all four cases, both methods underestimate the contribution of one signature, but the ILM method more drastically. The ILM method is also more prone to the erroneous detection of other signatures.

the QP approach, and approximately 15 minutes using the ILM approach (on a well-specified desktop).

Conclusion

Since it makes use of well-established and core R code in a classical mathematical context, no new software is required to use the QP approach (see Data and software availability and Supplementary material for details of implementation). The speed and improved performance of the QP approach makes it an attractive alternative to the ILM method and complements the additional functionality of the deconstructSigs package[5].

Data and Software Availability

F1000Research. Dataset 1: An R Markdown document that when compiled will reproduce all the results presented, 10.5256/f1000research.8918.d124181[13].

The raw oesophageal adenocarcinoma data for library SS6003314, from which some of these counts are derived, are available from the European Genome-phenome Archive (EGA; accession EGAD00001000704).

Competing interests

No competing interests were declared.

Grant information

AGL was supported in this work by a Cancer Research UK programme grant [C14303/A20406] to Simon Tavaré.

AGL acknowledges the support of the University of Cambridge, Cancer Research UK and Hutchison Whampoa Limited. Whole-genome sequencing of oesophageal adenocarcinoma was part of the oesophageal International Cancer Genome Consortium (ICGC) project. The oesophageal ICGC project was funded through a programme and infrastructure grant to Rebecca Fitzgerald as part of the OCCAMS collaboration.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

WGS sequencing pipelines for the oesophageal ICGC project are managed by Lawrence Bower.

Supplementary material

The R Markdown document (Dataset 1) compiled into a PDF file.

Click here to access the data.

References

1. Pfeifer GP, Denissenko MF, Olivier M, *et al.*: **Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers.** *Oncogene.* 2002; **21**(48): 7435–7451.
PubMed Abstract | Publisher Full Text
2. Alexandrov LB, Nik-Zainal S, Wedge DC, *et al.*: **Signatures of mutational processes in human cancer.** *Nature.* 2013; **500**(7463): 415–21.
PubMed Abstract | Publisher Full Text | Free Full Text
3. Shiraishi Y, Tremmel G, Miyano S, *et al.*: **A Simple Model-Based Approach to Inferring and Visualizing Cancer Mutation Signatures.** *PLoS Genet.* 2015; **11**(12): e1005657.
PubMed Abstract | Publisher Full Text | Free Full Text
4. Gehring JS, Fischer B, Lawrence M, *et al.*: **SomaticSignatures: Inferring mutational signatures from single-nucleotide variants.** *Bioinformatics.* 2015; **31**(22): 3673–3675.
PubMed Abstract | Publisher Full Text | Free Full Text
5. Rosenthal R, McGranahan N, Herrero J, *et al.*: **DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution.** *Genome Biol.* 2016; **17**(1): 31.
PubMed Abstract | Publisher Full Text
6. R Core Team: **R: A Language and Environment for Statistical Computing.** R Foundation for Statistical Computing, Vienna, Austria.
Reference Source
7. S original by Berwin A. Turlach R port by Andreas Weingessel: **quadprog: Functions to solve Quadratic Programming Problems.** R package version 1.5-5. 2013.
Reference Source
8. Goldfarb D, Idnani A: **Dual and Primal-Dual Methods for Solving Strictly Convex Quadratic Programs.** *Lect Notes Math.* Springer-Verlag, 1982; **909**(i): 226–239.
Publisher Full Text
9. Goldfarb D, Idnani A: **A numerically stable dual method for solving strictly convex quadratic programs.** *Math Program.* 1983; **27**(1): 1–33.
Publisher Full Text
10. Weaver JM, Ross-Innes CS, Shannon N, *et al.*: **Ordering of mutations in preinvasive disease stages of esophageal carcinogenesis.** *Nat Genet.* 2014; **46**(8): 837–43.
PubMed Abstract | Publisher Full Text | Free Full Text
11. Bamford S, Dawson E, Forbes S, *et al.*: **The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website.** *Br J Cancer.* 2004; **91**(2): 355–358.
PubMed Abstract | Publisher Full Text | Free Full Text
12. Putnam NH, O'Connell BL, Stites JC, *et al.*: **Chromosome-scale shotgun assembly using an *in vitro* method for long-range linkage.** arXiv: 1502 . 05331v1 [q-bio . GN] 18 Feb 2015. *Genome Res.* 2016; **26**(3): 342–50.
PubMed Abstract | Publisher Full Text | Free Full Text
13. Lynch A: **Dataset 1 in: Decomposition of mutational context signatures using quadratic programming methods.** *F1000Research.* 2016.
Data Source

Open Peer Review

Current Referee Status: ? ✓

Version 1

Referee Report 17 November 2016

doi:10.5256/f1000research.9596.r17440

✓ **Miguel Vazquez**
Structural Biology and BioComputing Programme, Spanish National Cancer Research Centre (CNIO), Madrid, Spain

The author tackles the problem of determining the mutational processes that were active on a tumor, and specifically in a single sample setting by leveraging already available signatures. The relevance of this approach is thus clear and was established in previous work: it allows working with signatures in a more general setting such as the clinic, and reusing already available signatures helps interpretation by the community as these become more familiar to all.

The author's contribution is limited to a technological advance, but in that is seems to surpass the previous approach in speed and accuracy (though I present some reservations below) by casting the problem into the more sophisticated framework of quadratic programming. I think this approach has benefits and I'm convinced that at no expense, and as such, I'm strongly favorable. I have however a few concerns that I'd like to raise.

Biologically I understand that mutational processes have signatures that are non-orthogonal, so a particular footprint of activity (the mutations on a sample) could in general be explained by different activation patterns of these signatures. How do these methods account for prior probabilities? e.g. mutational patterns related to smoking can be far more prevalent that exposure to a rare carcinogenic that could resemble the smoking signature in whole or in part. I can imagine the methods that extract this signature leveraging cohort data to untangle these prior probabilities, but then I think the methods presented in this paper in the deconstructSig cannot make use of this priors. In any case, I don't think current cohort methods predicting de-novo signatures are accounting for these priors since I would imagine they should be reporting these in addition to the signatures, which I believe they are not.

Coming back to the article at hand, the second paragraph in the result section seems to relate to this question in part. I find this paragraph confusing, possibly due to my own shortcomings so perhaps the author can clarify it for me, or even make it more clear on the text if need be. Let me explain. The author claims that the signature matrix is full rank. Correct me if I'm wrong, but in general it need not be, making the problem of approximating the result with different combinations is not just a result of noise and actually not specific to ILM, but to both methods. In fact the following phrase: 'an ILM approach is not guaranteed to give consideration to the correct combination of signatures' seems unfair, does the QP approach offer such guarantees? If so, perhaps this could be explained.

This paragraph was one of the main arguments for the improvement on accuracy, and I've presented my reservations. The other argument are some experiments presented on synthetic data involving the mixture of two signatures. These experiments seem too simplistic and I believe they do not address the problems presented in the previous paragraph either. However I do find that they suffice for the purposes of this article.

In conclusion, I concur with Mohamed that its mostly the performance that drives the message home at this point. Though I would not like to discourage indexing of this article, I feel that the author could improve his arguments regarding accuracy.

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Referee Report 06 October 2016

doi:10.5256/f1000research.9596.r16433

?

Mohamed Helmy
Bader Lab, Donnelly Centre for Cellular and Biomedical Research, University of Toronto, Toronto, ON, Canada

The article by Lynch presents a technical improvement of a recently published method[1] for inferring signatures of mutational contexts from large cancer sequencing data sets. The author proposes a quadratic programming (QP) approach over the iterative linear modeling (ILM) approach that was implemented in Rosenthal et al. According to the article, the presented approach provides technical improvement (speed) as well as an improvement in the accuracy.

The paper is well written and the results support the technical improvement of the QP approach over the ILM approach. The exercise provided by the author shows ~180 folds increase in the speed when using QP, comparing with the ILM approach (5 seconds vs. 15 minutes, respectively). That is a significant increase in performance that can be very useful when using such data in clinical applications, for instance. However, the current manuscript is not showing an improved accuracy for the QP approach.

Therefore, I would recommend adding more details that show the improvement in accuracy or just focus only on the improved performance of the presented approach.

Also, I have one minor comment: for consistency, Figure 1 should be A and B instead of top and bottom.

References

1. Rosenthal R, McGranahan N, Herrero J, Taylor BS, Swanton C: DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution.*Genome Biol*. 2016; **17**: 31 PubMed Abstract | Publisher Full Text

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.