Quantitative Bio-Science 31(2), 95~99(2012)

A Rejoinder to "A Microbiology Primer for Pyrosequencing"

Stephen Rush, Shaun Pinder, Marcio Costa[1], Peter T. Kim*

Department of Mathematics and Statistics, University of Guelph, Canada
[1]Department of Pathobiology, University of Guelph, Canada
(Received August 10, 2012; Revised October 31, 2012; Accepted November 20, 2012)

ABSTRACT

The discussions presented have been insightful and beneficial in terms of the topic of this paper. In our rejoinder we will attempt to harmonize the various discussants points of views. This provides a future research agenda that can be pursued for the enhancement of subject matters in mathematics and statistics, in the pursuit of a better understanding of biology.

Key words : Graphical models, Topological methods, Dimension reduction, Temporal evolution

1. Introduction

We wish to express our gratitude for each discussants contribution. It is always a pleasure when such varied perspectives complement each other and act cohesively to form possible programs of research.

While the many suggestions and views may be treated independently to some extent, we have noticed that there was a directional flow of material from one topic to the next. We have organized our rejoinder to mirror this flow, with the following order:

- Graphical Models
- Topological Methods
- Dimension Reduction
- Temporal Evolution

1.1. Issues

We wish to thank the authors of [1] for their suggestions for improving readability of [2]. Most of these have been implemented in one form or another. The most notable changes are the inclusions of two sections treating pyrosequencing and DNA amplification explicitly and the addition of a few clarifying examples.

Regarding the comment on "more integration between the content of the article and the program mothur", we wish to point out that in the first paragraph of section 3, we explicitly mentioned that the format of the article follows a typical routine through the mothur pipeline, and frequently refer to what can be done in mothur and the files involved. Further, we included an appendix with the commands used, in their order of use, with their defaults.

Finally, with respect to Kruskal's nonmetric multidimensional scaling, one of the coauthors of [1] received an earlier draft of the manuscript of [2] that did not yet include our exposition of this subject; the manuscript that was sent out to all others did. A number of equivalent and alternative majorization algorithms are described in [3] along with Kruskal's original algorithm.

We would like to point to an error on our part. In our original manuscript, we referred to the divergence measures of section 6.1 [2] as metrics; however, only the 'eachgap' distance is truly a metric.

2. Graphical Models

Massam and Mudalige [1] raise the possibility of applying

* Correspondence should be addressed to Dr. Peter T. Kim, Department of Mathematics and Statistics, University of Guelph, Guelph, Ontario N1G 2W1 Canada. Tel: +1-519-824-4120(ext.58165), Fax: +1-519-837-0221, E-mail: pkim@uoguelph.ca. This research was supported by an NSERC DG 46204.

graphical models to the analysis of microbial communities, as far as the evolutionary relationships in a phylogenetic tree are concerned.

Rhodes and Sullivant [4] have done much work to address this. Sullivant presented their joint work at a Field's Institute workshop on graphical models in April of 2012, with emphasis on Eukaryota. The same principles apply of course to Bacteria, but with some further complications.

With respect to massively parallel (pyro)sequencing (MPS) data concerning Bacteria, this method cannot be applied directly, since it relies on multiple gene sites that are assumed to be independent of each other. In MPS, a region of a single gene is sequenced; the fidelity of the technology is not yet such that we can have contiguous genes sequenced, and it is difficult to determine whether contiguous genes are independent. Even were the fidelity of MPS to improve to that extent, genes in the bacterial genome have a habit of moving from one place to another, so we could not guarantee the sequencing of the same collection of genes every time.

To further complicate this approach, certain genes can be exchanged between bacteriums of the same or divergent species.

However, outside MPS, rigorous phylogenetic mixture models [4] can be developed so that the relationships they determine may be used in the analysis of MPS data.

Huckeman [5] suggests RNA folding results in a sudden drop in dimensionality. When a structure is biologically selected, any changes to the base sequence that alters conformation generally lead to loss of functionality, and subsequent cell death. The restriction on configuration space places a restriction on sequence space, possibly resulting in the restriction to lower dimensional manifolds with singularities.

The implications of preserved helices, loops, and sheets on sequence space restrictions may possibly be formulated in the conceptually appealing framework of graphical models.

3. Topological Methods

The concept of topology treating the distribution of microbial communities or sequence spaces is very interesting. The picture can change depending on method of attack, although hopefully not the results.

While considering the topology of microbial, it is important to bear in mind that our space carries a quotient topology, in the following sense. A microbial community can be described as a collection of pairs (b_i, λ_i) where λ_i is some DNA sequence from bacterium b_i and $b_i=b_j \Rightarrow i=j$. These pairs exist in some 'environment', which we treat as a manifold, M. The data we have access to lies on the quotient space of M, call it Q, under the identifications $\{b_i \sim b_j \text{ if } \lambda_i=\lambda_j\}$. Thus while no longer working directly with a manifold, we still treat Q as Huckeman suggests [5], as a manifold with singularities.

An additional complication is incurred by the process of *in vitro* DNA amplification. In this process, an initial sample of genomic DNA extracted from bacteria are duplicated; we run the risk of "observing" the same bacterium multiple times.

3.1. Metrics

Perhaps the most obvious path towards understanding this structure is through the development of a (pseudo) metric topology.

In the last section of his discussion [6], Bubenik has suggested two excellent metrics. His first suggestion is to define the distance between sequences to be the length of the path determined by a series of mutations minimizing some cost function based on some mutation penalization schema. To some some extent, this may already be addressed when using the eachgap metric on aligned sequences. To see this, look to the exposition of the Needleman-Wunsch algorithm [7] in section 4.5.1 of [2]. Each sequence alignment algorithm is based on the insertion of gaps to minimize the cost of point mutations and insertions/deletions between sequences; when we apply the eachgap metric, the penalization is already implicit to the alignment space.

The second metric proposed by Bubenik is more appealing from a theoretical point of view. It denies that the minimum cost path is always the actual path taken by recognizing that many separate series of mutations can yield the same sequence. In this case, the associated probabilities for each mutation may be informed by phylogenetic mixture models as discussed in [4].

Ideally, any metric we might choose would correspond well with a rigorously determined phylogenetic tree.

3.2. Point clouds

Heo [8] interprets MPS data as point cloud data, embeddable in some Euclidean space. The choice of embedding affects the geometric picture we form of this data.

For instance, the eachgap metric is a length adjusted Hamming distance between strings (with at least one identical entry), and hence induces an isometry between the space of aligned DNA sequences and points clouds in a manifold of globally

zero curvature; this is clear by noting that the triangle inequality of this metric is strict when restricted to the alignment space in question. This makes it computationally and visually appealing; however it suffers from the unreasonable assumption that the length of a gap records the number of mutational events.

Alternatively, the least cost metric proposed by Bubenik [6] does not exhibit zero curvature, and makes more reasonable assumptions about the genetics involved. It is not immediately clear to us whether his ‘Feynman’ metric induces isometry between sequence space and Euclidean space. The relationship between these three metrics would provide an interesting avenue of research.

Heo raises the issue of variable sequence lengths from each sequencing run, for each sample. After some denoising, the standard procedure is to align all the sequences and then trim them so that they all overlap in the same alignment region; consequently much information is discarded. One wonders whether it is right to do so. The rational is explored and empirical evidence is presented in [9,10].

The length of the sequence in our space determines what stratum it lies in. To see this, consider how MPS data is produced. The DNA sequences are all sequenced starting (or ending) at the same position in a highly conserved region of the 16S gene (more generally an arbitrary gene). So there is a stretch over which all sequences are nearly identical, there being perhaps two or three distinct mutants over this region. This is the zero dimensional stratum of our space. As the length increases, sequences move onto higher dimensional strata. By tracking how these sequences diverge from each other under a given metric, we may form some picture of the stratified space, i.e. manifold with singularities, on which our data lies. We could then begin to consider geodesics, which we will discuss in section 4.

Developing intrinsic metrics based on what sections overlap in alignment may prove more fruitful than simply disgarding the inconvenient parts.

3.3. Persistent homology

The emphasis on metrics is not without merit. Persistent homology is a sort of cross between the rigidity of geometry and plasticity of topology, and requires some variation of divergence to form a filtration on the simplicial complexes over a point cloud; metrics tend to be of greatest appeal.

We mentioned the quotient of a manifold at the beginning of section 3. The clustering of sequences into operational taxonomical units (OTUs) represents further quotients. Varying the parameter (s) we use to cluster into OTUs provides a model application of the persistent homology machinery.

There are a few things we can explore with this tool. First, we can investigate the overall structure of the sequence space, assuming a uniform distribution of all possible sequences. Second, we can probe the restrictions imposed by the distribution of a given microbial community.

Bubenik and Heo refer to two descriptors, persistence diagrams/Betti barcodes [11], and Bubenik’s persistence landscapes [12]. Persistence diagrams have a simple interpretation, but the family of Wasserstein metrics induce Fréchet means with no uniqueness guarantee, and we are currently aware of no efficient algorithm capable of finding it, even though it exists.

Bubenik landscapes are far more appealing in this regard, as the mean is unique and easily calculated. Further, Bubenik landscapes permit the clustering of distributions based on their topological features. It may be noted that the persistence landscape itself generally has no preimage in the space of persistence diagrams; however, it is a simple matter to find the closest persistence diagram to the persistence landscape, which may provide a candidate for the Fréchet mean.

4. Dimension Reduction

One of the ultimate goals of the preceding discourse is the development of dimension reduction techniques specific to MPS.

4.1. PCA

It is fairly standard in microbial ecology to count how many bacteriums there are of each type for each sample, and to perform principal component analyses (PCA) on various sample groupings. The application of PCA to MPS data suffers from several oversights.

Due to the uncertainty of taxonomic classifications using one or two hypervariable regions, a popular procedure is to cluster sequences into OTUs based on some similarity criterion. We know that, given a species of bacteria, that there are several mutants of the hypervariable regions. We typically identify the species with their dominant mutant. Chakravorty et al. [13] have shown that no single region is capable of discriminating amongst all bacteria down to the species level. Thus, when we

observe a sequence, we have a number of candidate species. These mutants are distributed differently in related species.

When we cluster into OTUs, we lose sight of these distributions, and a loss of information is incurred; we can no longer tell whether there is one dominant species or several. This is a crucial point since there exist pairs of closely related species, one of which is generally benign while the other is generally pathogenic, for example *Clostridium butyricum* and *Clostridium botulinum* [14].

4.2. Geodesic PCA

All 16S rRNA sequences bear some phylogenetic relationship with each other. This clearly extends towards OTUs as well. Thus a pair of axes spanned by OTU i and j may be intrinsically "closer" than to each other than they are to the axis spanned by OTU k; the space has nonzero curvature.

Instead of assuming a Euclidean space and treating each sample as a single point, we can use the sequence counts as densities on the stratified space in which DNA sequences lie. The previous sections suggest a program for determining the structure of this space.

Huckemann, Hotz, and Munk [15] have developed a generalization of PCA based on variance maximizing geodesics of manifolds, which is appropriate in our setting.

5. Temporal Evolution

We can in fact investigate the validity of models of temporal evolution along the lines of shape spaces and geodesics.

MPS places us in a position to capture still frames of various microbial communities over time, so we can witness the various changes due to shifts in environmental pressures, through either external (introduction of new microbes, reagents) such as antibiotics [16] or internal (population growth, resource consumption) factors.

Some caution, however, is advised. In order to use MPS, we do need to extract some portion of the community; thus, as in quantum theory, to observe is to disturb. The debate as to how large a microbial community needs to be before the disruption is deemed insignificant should be interesting.

Another application of graphical models might arise here. Starting from a sterile environment, how does the structure of a community change over time if we seed the environment with Culture A and then after some period of time inoculate with Culture B, and how would this differ with the reverse?

There is ongoing research on the divergence of a particular strain of *E. coli*, begun in 1988 at Michigan State by Lenski et al. [17]. Today there are over 50,000 generations preserved [18]. It would be interesting to perform a temporal study on these "monocultures" to form an empirically backed picture of how the 16S rRNA gene changes over time.

The analysis of temporal evolution of diversity may vary well depend on the development of other methods as expounded by our discussants.

References

1. Massam H, Mudalige N. A Comment to "A Microbiology Primer for Pyrosequencing". Quantative Bio-Science 2012;31(2): 87-89.
2. Rush S, Pinder S, Costa M, Kim P. "A Microbiology Primer for Pyrosequencing". Quantative Bio-Science 2012;31(2):53-81.
3. Borg I, Groenen PJF. Modern multidimensional scaling. New York: Springer; 2005.
4. Rhodes JA, Sullivant S. Identifiability of large phylogenetic mixture models, B Math Biol 2011;1011-4134.
5. Huckeman S. A Comment to "A Microbiology Primer for Pyrosequencing". Quantative Bio-Science 2012;31(2):83-84.
6. Bubenik P. A Comment to "A Microbiology Primer for Pyrosequencing". Quantative Bio-Science 2012;31(2):85-86.
7. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 1970;443-453.
8. Heo G. A Comment to "A Microbiology Primer for Pyrosequencing". Quantative Bio-Science 2012;31(2):91-93.
9. Schloss PD. The effects of alignment quality, distance calculation method, sequence filtering, and region on the analysis of 16S rRNA gene-based studies. PLoS Comput Biol 2010;6(7).
10. Schloss PD, Gevers D, Westcott SL. Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. PLoS ONE 2011;6(12).
11. Cohen-Steiner D, Edelsbrunner H, Harer H, Mileyko Y. Lipschitz functions have l_p-stable persistence, Found Comp Math 2010; 10:127-139.
12. Bubenik P. Statistical topology using persistence landscapes. arxiv:1207.6437, 2012.
13. Chakravorty S, Helb D, Burday M, Connell N, Alland D. A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. J Microbiol Meth 2007;330-339.
14. Wiegel J, Taner R, Rainey FA. An Introduction to the family Clostridiaceae, In: Dworkin M, Falkow S, Rosenberg E, Schleifer K-H, Stackebrandt E, editors. Prokaryotes. A handbook on the

biology of Bacteria. 3ed. Singapore: Springer: 2006. Vol 4 Ch 1.2. 20.
15. Huckeman S, Hotz T, Munk A. Intrinsic shape analysis : geodesic PCA for Riemannian manifolds modulo isometric lie group actions. Stat Sinica 2010;1-100.
16. Dethlefsen L, Huse S, Sogin ML, Relman DA. The pervasive effects of an antibiotic on the human gut microbiota, as revealed by 16S rRNA sequencing. PLoS Biol 2008;6(11):2383-2400.
17. Lenski RE, Rose MR, Simpson SC, Tadler SC. Long-term experimental evolution in Escherichia coli. I. Adaptation and divergence during 2,000 generations. Am Nat 1991;6(138):1315-1341.
18. Lenski RE. Celebrating 50,000 generations of the long term lines. In Experimental evolution: myxo.css.msu.edu/ecoli/celebrate50K.html. [August 2012]
19. Kruskal JB. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika 1964;29:1-27.