September 22, 2015

Dear Ron,

We thank you and the anonymous reviewer for insightful and thoughtful feedback on our paper, which we have re-titled "Computational biology in the 21st century: Scaling with compressive algorithms." We believe we have addressed all major and minor points made by both you and and the reviewer. Below, we address each point in turn.

(1) regarding the uneven background: A good solution is to include a box of basic biology and bioinformatics terms, rather than putting them inside the text. (But do keep in mind the total length constraint..)

We have incorporated a box for definitions.

(2-3) I fully agree that the focus on compressive algorithms should be improved and that title should be more specific.

We have changed the title, and attempted to narrow the focus to compressive algorithms.

a. Numerous references are made to [4], sometimes several times within the same paragraph. I realize this was a way to get around the limit on citations, but the result is odd. One solution would be to cite all the original references but split some of them to the supplement.

We believe we have improved this situation, not by moving references to the supplement, but by pushing this reference to the end when needed, and by citing the original paper when possible (we freed up some space for references by narrowing the focus of the paper).

b. page 3 right, the sentence starting "A second way is to perform homology" is vague and would be lost for the CACM audience.

Homology search is now defined in the box, and we have improved the language here.

c. page 5 right, what does "unblock many blocked reactions" mean?

As part of the re-focusing of the paper, we limit discussion of this and other network approaches to a brief mention, and this terminology is no longer present.

d. page 6 left "compressive genomics" is used here and is only defined later. B_D is not defined. The bottom paragraph was vague for me.

We have defined B_D, rewritten the paragraph, and more properly introduced the term "compressive genomics."

e. page 6 Table 1 please explain the meaning of the results in the caption.

We have expanded the caption.

f. page 7 left please discuss the issue of loss of solutions due to the compression method.

We have added a paragraph discussing the accuracy limitations of compressive omics.

g. some jargon used and not defined: transcriptome, FASTQ, caBLAST, caBLASTP

We have added definitions, either to the box (transcriptome and BLAST) or in the text (caBLAST/caBLASTP). In our opinion, the mention of FASTQ was a needless detail, so we have removed it entirely from the paper.

Review of Berger et al. "Computational biology in the 21st century: Algorithms that scale"

This manuscript reviews computational problems that are faced in biology now that that discipline is awash in huge amounts of data. The manuscript generally focuses on problems in sequence analysis (mapping, assembly, gene expression, metagenomics, etc.) but also touches on chemical structure databases and problems biological networks. This topic is of crucial importance to biology, computational biology, and — I would argue — computer science, and so the article is very timely. The authors are leading researchers in developing fast algorithms for these problems and have produced many of the techniques cited in this review.

The statements made in the manuscript appear accurate to me, and as an expert in computational methods for biology the manuscript was understandable.

However, there are a number of stylistic weaknesses and weaknesses in choice of material that could be fixed in a revised version:

[1] It is not clear from the writing who the audience is for this article. On the one hand common biological terms such as "model organism" are defined, while more complex concepts and jargon are not (16S, shaped seeds, blocked reactions, chemogenomics, de Bruijn). I suspect it would be difficult for a general reader of CACM to follow some of this manuscript. Similarly, I think the manuscript does not do enough to convince the general reader of the interest or importance of the topics described.

We have attempted to address this completely valid criticism in two ways: by re-focusing the paper, some of these concepts (blocked reactions, 16S) do not come up. We have defined those terms that are still present. Moreover, we hope we have done a better job of emphasizing the importance of the topics at hand.

[2] Though it is admirable to try to cover the breadth of biological data types, the flow of the article is often somewhat disjointed as the context is switched from genomics, to networks, to chemical databases.

We have narrowed the focus of the article; networks are only mentioned briefly in order to explain that they are outside the scope of the article, and we believe we have improved the connections between the disparate subjects (genomics, structure, and chemical databases).

[3] It isn't clear what the overall point that the authors are trying to make is. There are some clear sub-points (use the structure of the data, a hierarchical algorithmic approach

often works well), but the article lacks a strong, novel point of view. This starts from the subtitle: "Algorithms that scale". This has been a goal of computer science since its start. The article would benefit from the development of a more specific story, a more decisive point of view, and a stronger backbone. As just an example, a title or subtitle such as "The success of compressive algorithms" and the corresponding changes to make that the early and complete focus of the manuscript would result in a stronger, more interesting article.

In addition to a new title, we have endeavored to strengthen the story regarding compressive algorithms, both by reorganizing sections of the paper and by narrowing the focus.

[4] More text could be spent synthesizing what is special and different about biological "big data" than other (sometimes larger) data collections in other disciplines. Relatedly, the introduction spends a lot of time arguing for the need for faster algorithms, which I think is a settled point nearly everywhere in computer science. More interesting would be a more detailed discussion of the situations in biology where faster algorithms are most crucial (this is done with read mapping, but there must be others).

Within the space limitations, we believe we have addressed this. We have rewritten the last paragraphs of the conclusion, in particular, to discuss how biological data is unique.

Minor points:

[5] If kept, the paragraph in section 3 starting with "Given a sequenced genome..." should come before the paragraph that precedes it.

We have made this change.

[6] Burroughs-Wheeler is misspelled in section 1. (It should be Burrows)

We have corrected this error.

[7] The sample size argument of section 5, paragraph 2 [there are a possible 20^500 *proteins but only* $6.9x10^7$ *are observed] is not persuasive. Such "search-space-size" arguments are nearly always vapid. This sentence is one of a over 39841379142783065371079463001877881566518830903922673 possible strings of characters of this length. But I had to consider very few of them to write it. I would find a more specific argument about evolution, relatedness of species, or something else to argue that the low dimensionality is surprising.*

We have attempted to produce a less vapid argument, by focusing on the redundancy inherent in (for example) multiple genomes.

[8] The last two paragraphs of the manuscript are somewhat disappointing. The penultimate paragraph positions computational biology as a derivative field, an applier of ideas from the broader CS community. The last paragraph deflates the hope of computation driving science with the appeal to requiring wet-lab experiments. Neither paragraph is wrong, but their emphasis does not serve computational biology well, in my opinion.

We thank the reviewer for pointing this out; indeed, it was a rather dismal conclusion to the article. We have completely rewritten the conclusion after meditating on what distinguishes

biological "big data" from other sources of big data.

Sincerely,

Bonnie Berger