Response to Reviewers (Second Round)

COATi: statistical pairwise alignment of protein-coding sequences

Juan J. García Mesa Ziqi Zhu Reed A. Cartwright

Dear Editor,

We are resubmitting our manuscript, "COATi: statistical pairwise alignment of protein coding sequences", for your consideration. This manuscript describes a new model for aligning protein coding sequences which produces more accurate alignments, while being robust to common artifacts found in genomic annotations. Our approach is implemented in the software package, COATi, which is released via an MIT license and available on GitHub. While COATi can generate multiple-sequence alignments, this paper is focused on our pairwise aligner which is has received more attention and validation. Additionally, to promote open science and reproducibility all data, procedures, and scripts used to validate COATi and compare it with alternative methods have been uploaded to GitHub.

We would like to thank the associate editor and two reviewers for their helpful comments and suggestions. We have addressed all of the concerns brought to our attention by the reviewers as described below.

Sincerely,
Reed A. Cartwright, on behalf of all authors

Reviewer 1 Comments

1. Comment: *It is now very clear and easy to read manuscript and a important edition to the alignment literature since it does well relative to the objective it pursues. It has moved away from statistical alignment formulation in the first submission toward doing well in agreeing with benchmark alignments. Some discussion of this would be useful.*

   TODO

2. Comment: *And how will the authors turn this into a multiple alignment tool?*

   COATi currently has an early-stage module for estimating pairwise alignments. However, it needs more work (and an additional round of funding) before it meets the high-quality standards exhibited by the pairwise aligner.

Review 2 Comments

1. Comment: *This is my second review of the manuscript "COATi: statistical pairwise alignment of protein-coding sequences" by García Mesa et al. and I have to say that it is practically a new manuscript. The first version really excited me, and I felt the science was both sound and very useful to the community, but the paper was difficult to understand. This version is much better; it is well-written, and it confirmed my belief that COATi is an important and useful contribution to the field of sequence alignment.*

   We appreciate your positive words.

2. Comment: *My only somewhat more conceptual comment is about the use of K2P distances. I buy the argument that "… the Kimura's 2-parameter distance is more suitable for non-coding sequences, it is straight forward to calculate and provides a quantitative measure of the evolutionary divergence between sequences" (p. 8, line 191) I am a bit mixed about its use. To be honest, I don't think there are any truly appropriate distances for*

coding regions. However, p-distances would actually be more intuitive. The authors should consider calculating p-distances and placing them in the supporting information (or even replacing the K2P data). I don't think this will change conclusions, but it would make things more intuitive for readers. Since I am certain this won't change conclusions, I don't want to push it that hard.

We calculated p-distances and have included the results in the supplement. The results were equivalent to K2P distances, which we now mention in the manuscript.

3. Comment: *I am not certain why the authors use "phase-3" indels rather than "phase-0". I find phase-0 more intuitive, and it strikes me that this could be changed easily and would simplify the writing – the authors wouldn't have to explain that they using phase-3 instead of phase-0!*

   We originally used phase-0 in our terminology. However, in our experience from presenting this research to diverse audiences, phase-3 was easier terminology to explain than phase-0. For example, "a phase-X gap occurs after the X position in a codon" is easier to explain than "a phase-X gap occurs before the X+1 position in a codon". Importantly, by using phase-3, we can succinctly explain the relationship between the frequencies of the gaps as phase-3 > phase-2 > phase-1. We have added additional language to the manuscript to emphasize these points.

4. Comment: *The authors should use "see" rather than c.f. on p. 5, line 82. The abbreviation should be cf., not c.f., and it should be used to mean "compare." Given the Ranwez et al. references I don't think the authors of this paper mean "compare."*

   We do mean "compare" here because COATi's innovation is in similar spirit to MACSE but fundamentally different.

5. Comment: *I would consider adding the fact that their "...COATi results estimate that only 41% indels in protein-coding sequences between humans and gorillas are phase-3 indels" (on p. 15, line 370) to the abstract. This is a nice empirical demonstration that their methods is producing "biologically-reasonable" alignments.*

   Thank you for the suggestion. We have reworked the abstract and including this results.

6. Comment: *I'm a bit confused by the authors' use of "frameshifts" when discussing the behavior of Clustal Omega (p. 15, line 402). A simple phase-1 or phase-2 indel that is a multiple of 3 is not a frameshift. Are they referring to separate indels that might have an intermediate that has a genuine frameshift? By "genuine frameshift" I mean a case where a presumably inactivating indel mutation that is not a multiple of 3 occurs and persists in a population long enough to have a nearby compensatory mutation that puts the sequence back in frame (i.e., the sum of then indel lengths is a multiple of 3). I think a minor rewrite for clarity would help readers.*

   In this instance, we are referring to an abiological frameshift that produces an abiological amino-acid translation. We have reworked this section for clarify our meaning.

7. Comment: *Overall, I really like this new version of the manuscript!*

   Thank you.