July 3, 2014

Genomics (Elsevier)   
RE: Response to reviewers (Manuscript: Sequence Classification of homoPolymer Emissions)

Dr. Quackenbush:

We want to thank the reviewer for the comments on our manuscript, which led to a significant improvement in this paper. We specifically thank them for suggesting the two comparisons (of SCOPE++ versus a basic tool, and of SCOPE++ versus a near-trivial HMM), which, as the reviewer was correct in observing, significantly adds to the overall study.

In the following we address all comments point by point.

**Reviewer Comment**: “you really need to proof this manuscript much, much better, before it's ready for publication”.

**Response:** We apologize, and very much appreciate the fact that the reviewer could look past this. It looks like I inadvertently uploaded the wrong draft of the paper (complete in substance, but a draft before proofreading / editing was completed). The resubmitted version should be in closer to final form and largely free of such errors.

We will not further address points related to grammar / style / English in this response. In all cases the reviewer was correct, the problems have been fixed, and we ran it through a third party for editing for good measure.

**Addressed in paper**: Throughout.

**Reviewer Comment**: “You could, and should, do a lot better with respect to referencing appropriate concepts.”.

**Response**: We have added references to materials on HMM (the Durbin book, and the classic 1986 Rabiner overview paper), as well as a few words of explanation when first introducing Viterbi and Baum-Welch. Note that we opted to emphasize the Durbin book, as opposed to original papers. We feel that while difficult, Durbin is still easier for more biologically oriented researchers to understand than most of the technical papers on the subject.

**Addressed in paper**: In the first paragraph of Methods we explicitly refer the interested reader to *Durbin et al.* (as well as mention Rabiner). Following this we have added citations to Durbin as appropriate.

**Reviewer Comment**: “you do things like compare algorithmic results to human-results, based on an apparent assumption that the human annotation is error-free”

**Response**: We clearly did not explain this well. (Or, perhaps, did explain it well – and obscured understanding with the grammatical problems.) This is not what we are doing, and I’ve rewritten the relevant text to try to clarify. But let me provide an alternate explanation here.

We are *not* comparing our tool output to human annotations, but instead using those human annotated sets as the basis for creating simulated benchmark sets. We want to compare different tools, but in order to do that we need a set where we know the correct annotations: we cannot differentiate a true positive from a false positive if we don’t know where the tails actually lie. We were unable to find any existing benchmark sets, so the other option was to test them on simulated sequences. But we wanted to be careful that our model wasn’t over-simplified – possibly leaving out unidentified structural characteristics of the sequence which impact tool result quality. Thus we developed our “semi-synthetic” sequences as follows:

1. Take a set of actual sequences and hand-annotate them for poly(A) tails. We had six sets covering three organisms from three different sequencing technologies, and randomly picked 500 sequences from each of the six sets for the hand annotation. (Not fun; thank goodness for the existence of undergraduates.)
2. We next took each sequence and “purified” the annotated tail – meaning that we replaced each non-A character with an A[[1]](#footnote-1). What we then had was a “semi-synthetic” sequence – a real sequence with a possibly synthetic tail embedded in it. Whether or not the humans made any errors, with the “cleaning” we have introduced a tail may be biologically meaningless, but it should be identified as a poly(A) tail by any search tool. Further, any characteristics of the surrounding sequence that might have an impact on a tool are still present, as that portion was unchanged from the original.
3. From this we could generate simulated sequences with arbitrary (controllable) error that could serve as benchmarks for quality measurement.

This is clearly not perfect, and there is more we could do to increase the accuracy of the model. But we believe that this is a very good first-order approximation, providing us with the ability to test various tools despite the lack of actual benchmark sets.

**Addressed in paper**: We have rewritten the “Benchmark Sets” section of Methods (trying to describe the above more concisely).

**Reviewer Comment**: “More detail on the actual HMM would be appreciated.  For example, if you have only one "in a homopolymer" internal state, but this state accepts errors, what prevents it from accepting two long homopolymers, separated by a large number (but small proportional to the length of the two homopolymers) of non-homopolymer bases?”

**Response:** Certain potential errors are likely unavoidable for any method that is based solely on sequence analysis. (Example: if a non-tail A base happens to fall next to the poly(A) tail, it is inevitably going to be included as part of the tail – and it seems unlikely that this could even be verified as a mistake through only sequence analysis.) We argue that any tool will suffer from such problems. In the case of closely-neighboring tails in specific, in practice they do not tend to be close enough that SCOPE++ will merge them – the amount of error this will introduce is negligible.

**Addressed in paper**: We have tried to add more details, and explain the above, in what is now the last paragraph of the HMM section.

**Reviewer Comment**: “What makes the HMM approach more appropriate than a simple regular expression that matches the AA leader and trailer, wrapped around a "no more than N% non-A" match?”

**Response**: This was a great idea, and descriptions of the results have been added to the paper (paragraph 2 of Results – we included a text description only, as the results plots were uninteresting, and not worth extending the length of the paper to include). Doing this analysis also highlighted the (unsurprising) fact that identifying whether a poly(A) tail is present is fairly easy (even in the presence of obfuscating errors) – but pinning down the end-points is quite difficult. It is for the second that the complexity of the HMM is required.

**Addressed in paper**: Discussion has been added in what is now the second paragraph of the Results section.

**Reviewer Comment**: “Someone curious about the generalizability of the approach to other domains might be interested in why the Viterbi algorithm is not proving to be computationally overwhelming, and why Baum-Welch optimization shows no meaningful improvements.  A credible argument could be made that this is because the HMM is both trivial, and a poor structural match for the underlying biological phenomenon.  I don't believe that this is actually the case here, but, those observations would be consistent with a trivially simple, inappropriate HMM - refuting this wouldn't be a bad thing to do.”

There are two points here:

1. It is our belief that Baum-Welch does not help because our preliminary estimation of the HMM parameters are fairly close to the optimal values; spending time on Baum-Welch to improve the estimation is not worth the effort, as there is little improvement to be had. (To put it another way: the parameter estimation problem in this case is too simple to bother with Baum-Welch.) We have added a note to this effect.
2. As suggested, we implemented an alternative HMM model consisting of two states (“poly(A)” and (“non-poly(A)”). As predicted, it failed. The tool had a sensitivity of below 0.1. We have added a new paragraph explaining this, and drawing the conclusion that our HMM must be reflecting the biological structure in some way that the basic HMM fails to do. (We again thank the reviewer for this suggestion; it really does add something to the paper.)

**Addressed in paper**: The first point is addressed in the third paragraph of the “Parameter Estimation” section. The second point is covered in what is now the third paragraph of **Results**.

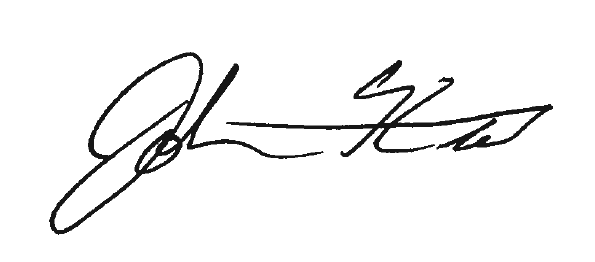
**Reviewer Comment**: “Can the approach be extended to arbitrary (non homogenous) tandem repeats?”

**Response**: I’ve added a discussion to the conclusion, but I am hesitant to say too much without actual study. Attempting to identify complex sequences (in the information-theoretic sense) could add enough complexity to the HMM to drive up the runtime to undesirable levels. However, searching for low-complexity repeated sequences (e.g. satellite sequences such as those that cause Huntington’s disease) might be more in the range of what the approach can cover. This could be worth following up on, but is beyond the scope of what we are presenting here.

**Addressed in paper**: We have added a paragraph to the conclusion in which this is discussed.

We believe this covers all points made by the reviewer, and look forward to hearing from you on this paper.

Yours Truly,



Dr. John Karro

Associate Professor, Miami University

Department of Computer Science and Software Engineering, Department of Microbiology, Department of Statistics

karroje@miamiOH.edu

1. Half our sets were from the complementary strand, thus containing poly(T) tails – where the appropriate adjustments were made. [↑](#footnote-ref-1)