Frameshift Aware pHMM for Translated DNA to Protein Alignment

Genevieve Krause, Wheeler Lab, University of Motnana

Introduction: As the size of biological sequence datasets has grown, so too has the computational power and algorithmic sophistication used to annotate those sequences. The use of probabilistic models, such as profile hidden Markov models (pHMMs), allows a closer approximation of evolutionary relationships and improves the sensitivity of the sequence database searches which underlie annotation (5). When annotating protein coding DNA, the primary units of information are not single nucleotides, but rather codons (made up of three consecutive nucleotides), each of which is translated into one amino acid. The bounds of these codons are known as the frame, with each protein being translated form only one frame at a time. However, there are some DNA sequences in which frameshifts (the insertion or deletion of nucleotides that shift the bounds of codons) result in an incorrect amino acid translation, obscuring the evolutionary relationships and preventing annotation.

My research focuses on annotating protein coding DNA sequence which contain frameshifts using pHMMs and novel modifications of the forward and backward algorithms. By allowing for variable length codons, each position of the pHMM can be aligned to anywhere from one to five nucleotides in the DNA sequence. This promises to improve annotation of sequences containing either naturally caused frameshifts such as pseudogenes, or those caused by sequencing errors in long read technologies. The ability to annotate pseudogenes despite the accumulation of frameshifts can be vital to tracking the evolutionary history of genomes (3,6). Long read DNA sequencing technologies, such as PacBio and Nanopore, have tremendous implications for sequencing de novo genomes and metagenomic samples, but the high error rates of these technologies can lead to frameshifts in protein coding regions that prevent proper annotation (1,2). I hope to provide researchers with a straightforward approach to taming these nuisance sequences.

Hypothesis: If a DNA sequence is protein coding and contains frameshift errors then is can be aligned to the correct protein product by allowing each amino acid to align to anywhere from one to five nucleotides.

Methodology: I will build my software within the HMMER suite of sequence alignment tools. This will allow me to draw on the robust set of optimized functions that allow HMMER to offer superior sensitivity and comparable run times to non-probabilistic alignment tools such as BLAST (CITE). However, the core algorithms in HMMER (forward/backward), and all of the downstream calculations which allow an alignment to be retrieved and accessed for significance, will need to be reworked to allow for variable length codons. The validity of this method will be shown though tests with simulated and real-world data.

Intellectual Merit: Previous attempts to align frameshifted DNA to proteins such as LAST and DIAMOND (CITE) have been based on the scoring matrix method of alignment which has been sown to be less sensitive compared to the use of pHMM (CITE). Furthermore, these methods only allow for codons to vary in length from two to four nucleotides. This decreases the number of

calculations but it also forces unnatural limitations on the way frameshifts can occur. When an indels is of length mod 3 = 2 these alignments will need to spread the indel over two codons to retrieve the correct frame. My software will reveal whether this decreases the accuracy of the alignments since the nucleotides in the frameshifted codon cannot be properly aligned. When sequences contain frameshifts, we can also expect them to contain non-synonymous substitutions, further obscuring the correct protein translation. By allowing mod 3 = 2 indels to be contained in a single codon I hope to preserve as much of the correct translation as the sequence errors will allow and therefore be able to annotate with greater sensitivity and accuracy. Combined with the power of pHMM based alignments this software could be an incredibly powerful tool for researchers wanting to decode frameshifted DNA sequences.

Broader Impact: (This is an outline of the points I will want to make in this section)

Third generation long-read sequencing technologies have great potential but high error rates

Error correction has improved long read annotations but frameshifts are still present and problematic. Also, these error corrections rely on costly multiple runs of the sequence technology.

My tool could deliver better annotation at lower cost.

Researchers interested in naturally frameshifted sequences, such as pseudogenes and viral insertions, will also find my software useful.

References

(1) Biosciences P. Errors in long-read assemblies can critically affect protein prediction. 2019;37(February):1246.

(2) Goodwin S, Mcpherson JD, Mccombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Publ Gr [Internet]. 2016;17(6):33351. Available from: http://dx.doi.org/10.1038/nrg.2016.49

(3) Harrison PM, Gerstein M, Avenue W. Studying Genomes Through the Aeons: Protein Families, Pseudogenes and Proteome Evolution. 2002;2836(02):115574.

(4) Madera M, Gough J. A comparison of profile hidden Markov model procedures for remote homology detection. 2002;30(19).

(5) Mccutcheon JP, Moran NA. in symbiotic bacteria. Nat Publ Gr [Internet]. 2011;10(1):1326. Available from: http://dx.doi.org/10.1038/nrmicro2670

(6) Pearson WR, Wood T, Zhang Z, Miller W. Comparison of DNA Sequences with Protein Sequences. 1997;36(46):2436.