Improved Genomic Annotation using Frameshift Aware Translated Search with Profile Hidden Markov Models

Genevieve Krause, Travis Wheeler
Dept. of Computer Science, University of Montana, Missoula, MT, USA

Abstract

Accurate annotation of biological sequences is fundamental to modern molecular biology. For many sequences this is a straightforward process - tools such as BLAST and HMMER quickly and accurately annotate sequences by aligning them to known sequences or sequence models. Here, we are interested in annotation by translated alignment, in which protein-coding DNA is aligned directly to protein sequences or models. We demonstrate that the use of profile hidden Markov models (pHMMs) substantially increases annotation sensitivity relative to sequence-to-sequence comparison methods such as tblastn. Even with pHMMs, annotation of protein-coding DNA sequences containing frameshift inducing indels can be particularly troublesome, as standard models do not support alignment through frame shifts. Here we present a new tool, built within the open source HMMER software package, that produces high-quality translated alignments and accurate annotation for even heavily frameshifted DNA sequences. With a new model and a first-ever Forward-Backward dynamic programing algorithm for frameshift-aware alignment, this tool promises to increase annotation of naturally frameshifted data such as pseudogenes and transposable elements, as well as improving the annotation of indel-rich long read sequencer data.

Frameshift Aware pHMM

The use of profile hidden Markov models has been shown to increase sensitivity of sequence comparison tools by allowing position specific substitution and gap scoring. When comparing DNA to protiens a pHMM can matches three nucleotides in the target DNA sequence to each amino acid in the protein query HMM. It also allows for the insertion or deletion of amino acids and the corresponding three nucleotides, but when confronted with the insertion or deletion of only one or two nucleotides the resulting frameshift causes the model to incorrectly translate the remaining sequence. The result is that alignments are cut short or missed altogether.

This problem is not unique to pHMM based search but effects all current translated search methods. The use of pHMMs, however, provides us with a potential solution. By modifying the HMM to allow for some probability of inserting or deleting any single nucleotide, we can account for these indels without losing track of the correct frame. This is the approach we have taken in our frameshift aware pHMM.

Figure 1

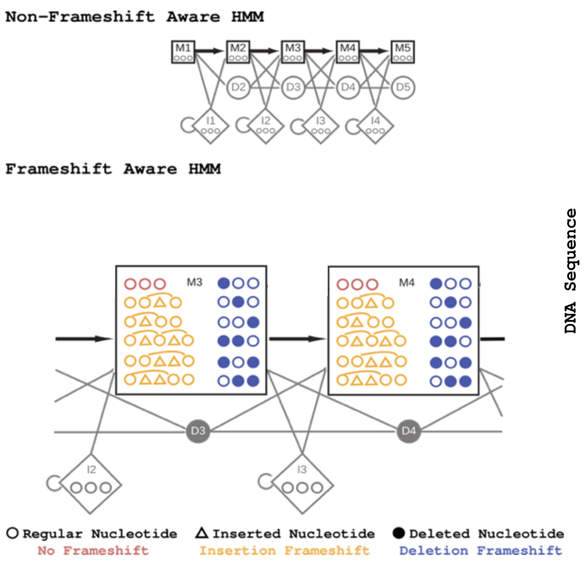


Figure 2

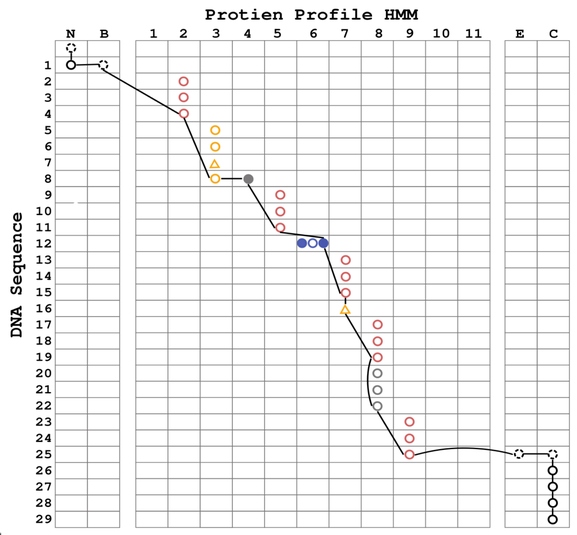


Fig. 1 displays a graphical representation of the core component of the pHMM architecture used currently in standard translated search (top) and in the frameshift aware version (bottom). The frameshift awareness comes from a greater complexity in the emissions of the match states. Rather than emitting amino acids made only from a standard three nucleotide codon the new model allows codons to be a as short as one nucleotide (having two deletions) or as long as five nucleotides (having two insertions).

An example of a dynamic programing matrix (Fig. 2) shows how an alignment can proceed through various states and codon types. The N and B columns allow the model to move through nucleotides until the probable start of the alignment is located. The core model then matches translated codons to amino acid positions, moving diagonally through the matrix. Deletions (filled grey circles) and insertions (open grey circles) of amino acids allow the alignment to move horizontally or vertically through the matrix but cannot account for frameshifts. This is made possible by nucleotide insertions (shown as yellow triangles) and nucleotide deletions (shown as filled blue circles) which allow for translation to continue in the correct frame. Once the probable end of the alignment is located the E and C columns then allow the model to account for he rest of the nucleotides, if any exist.

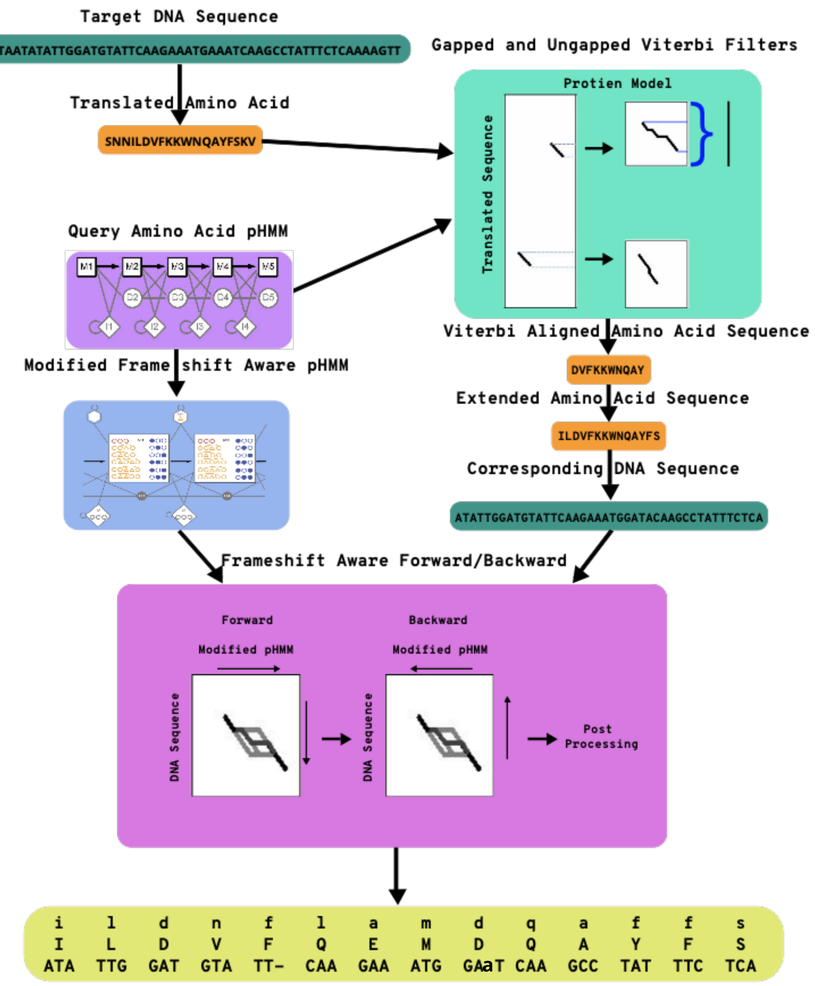
Frameshift Pipeline

The implementation of the new frameshift aware model and algorithms required that we modify the pipeline by which the HMMER software package filters and processes sequence compassions. Fig. 3 shows a representation of the new pipeline. The user provides one or more DNA target sequences and one or more query protein pHMMs.

The DNA sequence, and its reverse complement, are then translated into amino acid sequences from all available open reading frames. These amino acid sequences are then feed into the ungapped Viterbi filter which confirms the existence of significantly high scoring ungapped alignments between the translated sequence and the user provided protein pHMM. The gapped Viterbi filter then searches for longer alignments by allowing for the insertion or deletion of amino acids. By working with a standard codon translation in these filter stages we ensure that only sequences with a high probability of being genes of pseudogenes make it to the frameshift search stage.

If a translated sequence passes through the Viterbi filters the length of the amino acid Viterbi alignment is returned. The sequence is then extended passed the alignment to capture additional sequence that could have been missed in a simple translated Viterbi comparison. Finally, the translated sequence is then mapped back to the corresponding start and end coordinates from the original DNA sequence.

Figure 3



It is this subsequence of the original DNA target that is then sent off to the new forward and backward algorithms. When compared to the modified pHMM the alignment can easily switch between frames to maintain the correct alignment in the presence of nucleotide indels. After post processing the program will then display the forward score and corresponding e-value, as well as an alignment that shows the consensus sequence from the protein HMM, the translated target sequence, and the original DNA sequence split in to codons to display indels.

Results

Mathew Campbell and John McCutcheon (University of Montana) provided us with genomes of the Magicicada endosymbiont. The unique evolution of these endosymbionts has resulted in the proliferation of indel ridden pseudogenes. While several homologous protein sequences from related organisms are available no currently available software has been able to annotate these pseudogenes.

Using the new frameshift aware tool we were able to successfully align some of these pseudogenes. Fig 4. shows the limited annotation possible using standard search algorithms (grey) verses the expanded annotation possible with out new frameshift aware methodology on one of these endosymbiont genomes.

Fig. 5 highlights one of these alignments. Using standard pHMM search we able to find a short segment of the alignment but the new tool was able to expand the alignment by ~1100%. Recent improvements to runtime and memory usage have prepared out model for use on more datasets. We are looking forward to more results like these in the near future.

Figure 4

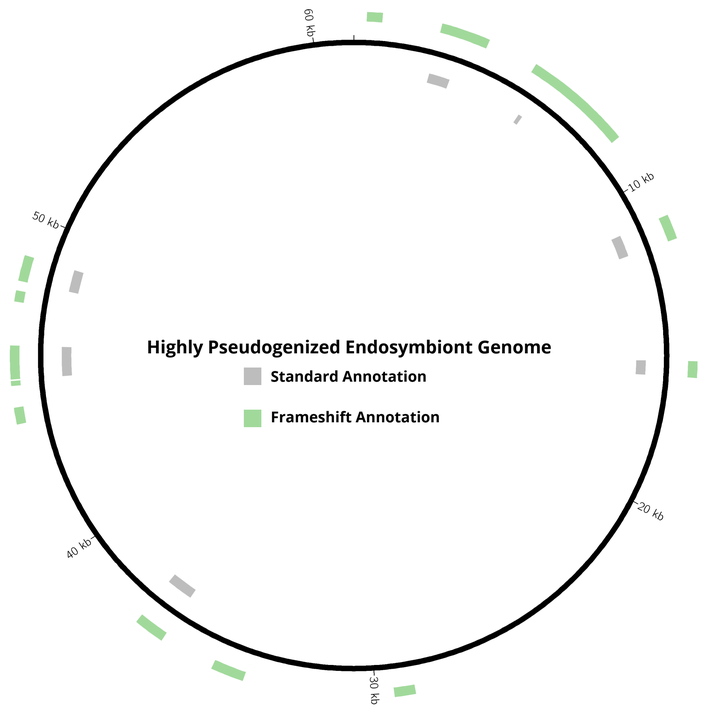


Figure 5