

# 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 programming 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.

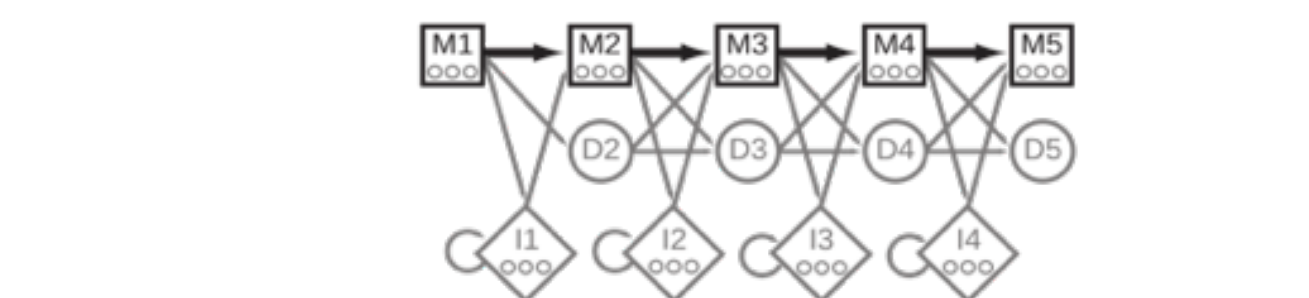
## Frameshift Aware pHMM

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 proteins a pHMM can match 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.

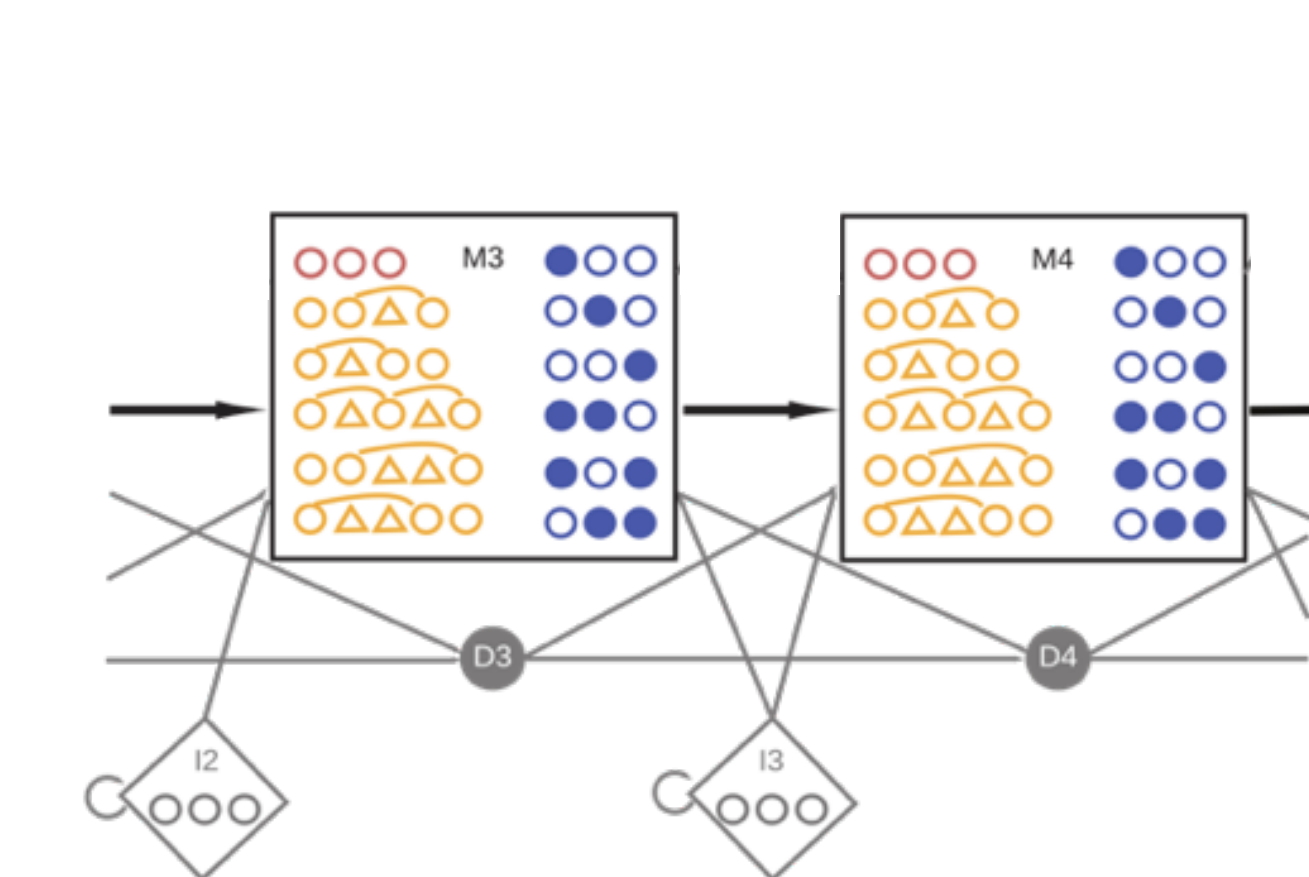
This problem is not unique to pHMM based search but effects all current translated search methods. The use of pHMMs, however, provides us with a potential solution. By modifying the HMM to allow for some probability of inserting or deleting any single nucleotide, we can account for these indels without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

Figure 1

Non-Frameshift Aware HMM



Frameshift Aware HMM

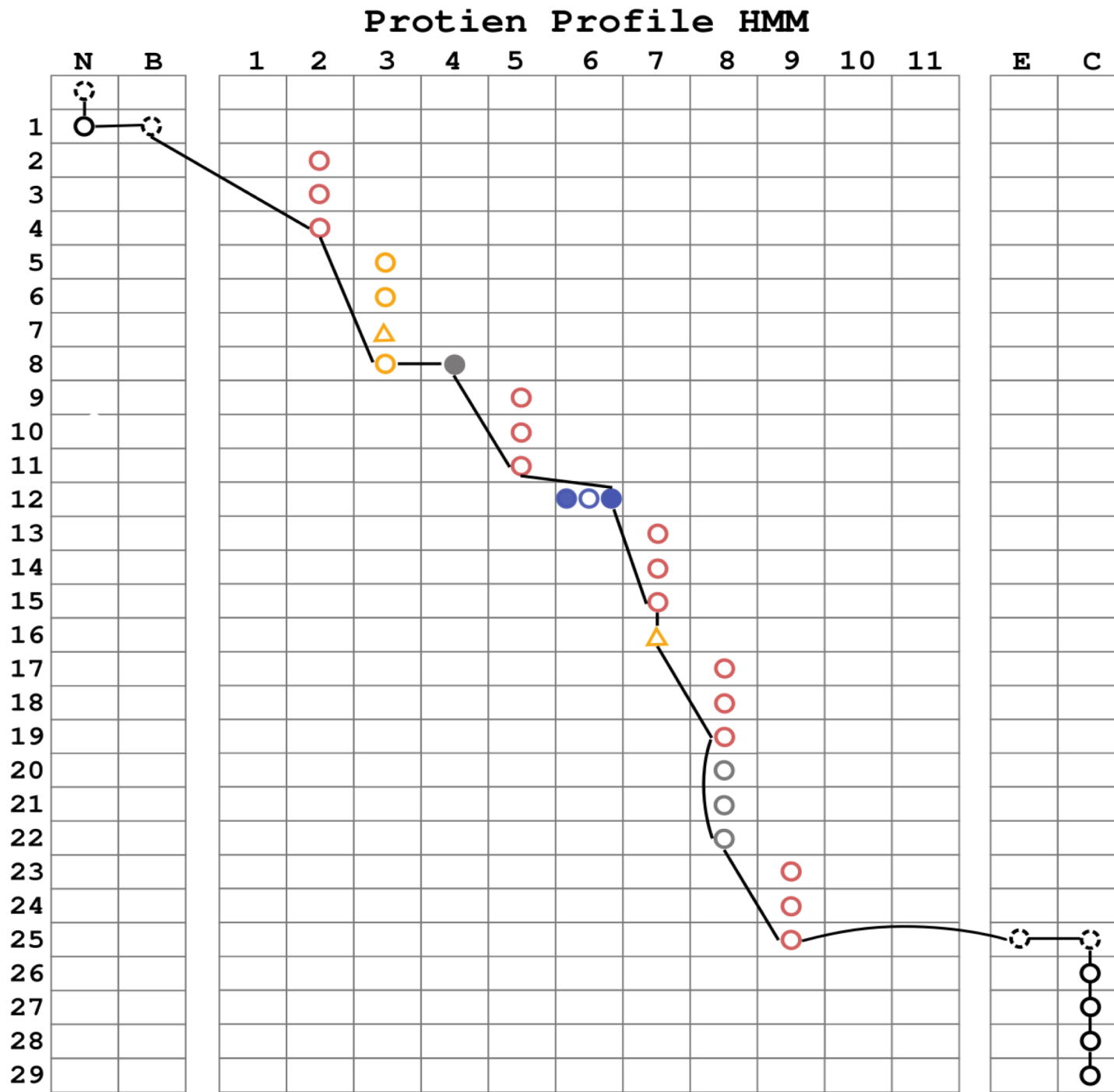


○ Regular Nucleotide    △ Inserted Nucleotide    ● Deleted Nucleotide  
No Frameshift    Insertion Frameshift    Deletion Frameshift

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 as short as one nucleotide (having two deletions) or as long as five nucleotides (having two insertions).

An example of a dynamic programming 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 the rest of the nucleotides, if any exist.

Figure 2



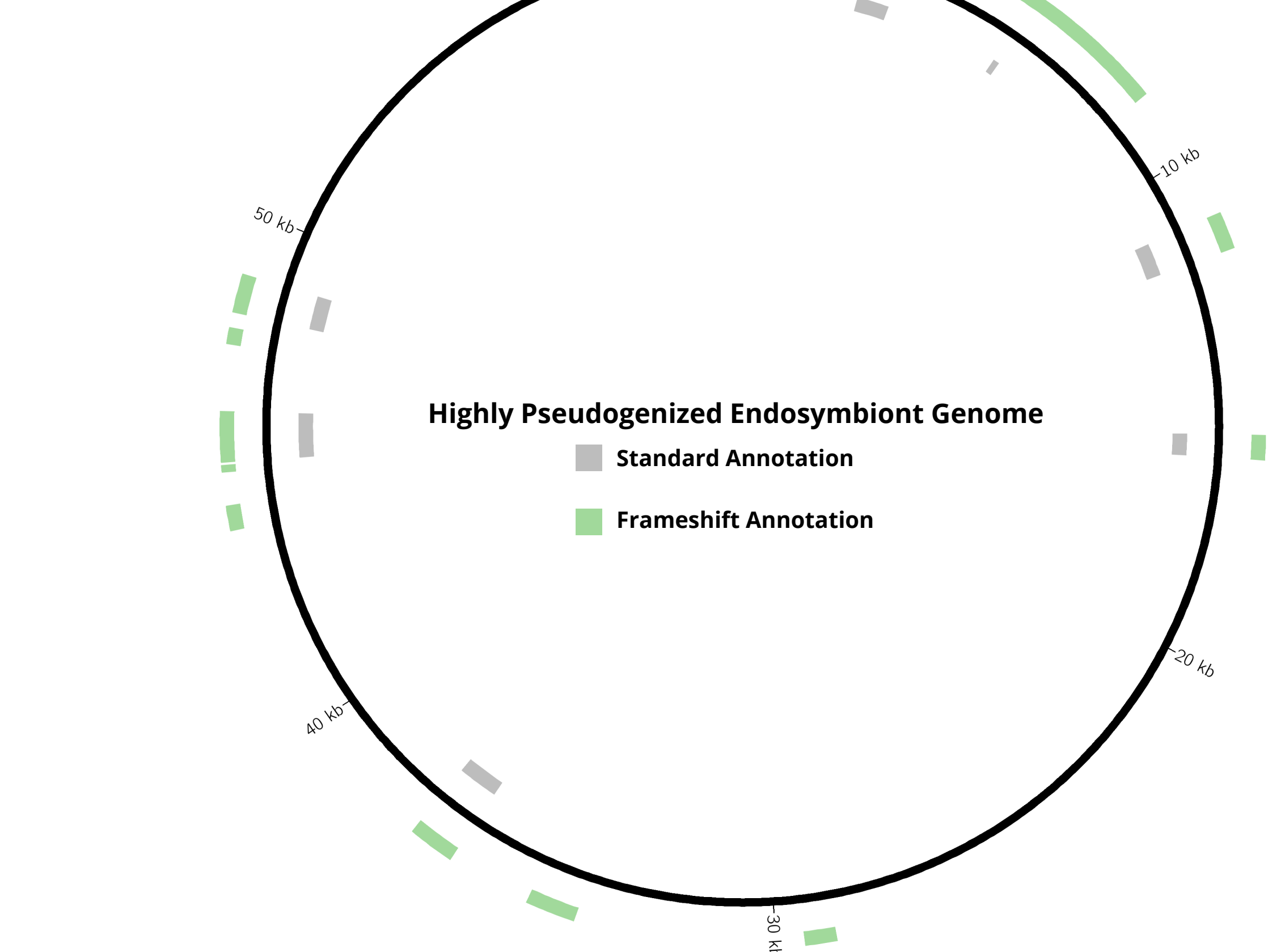
## Results

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 our model for use on more datasets. We are looking forward to more results like these in the near future.

Figure 4



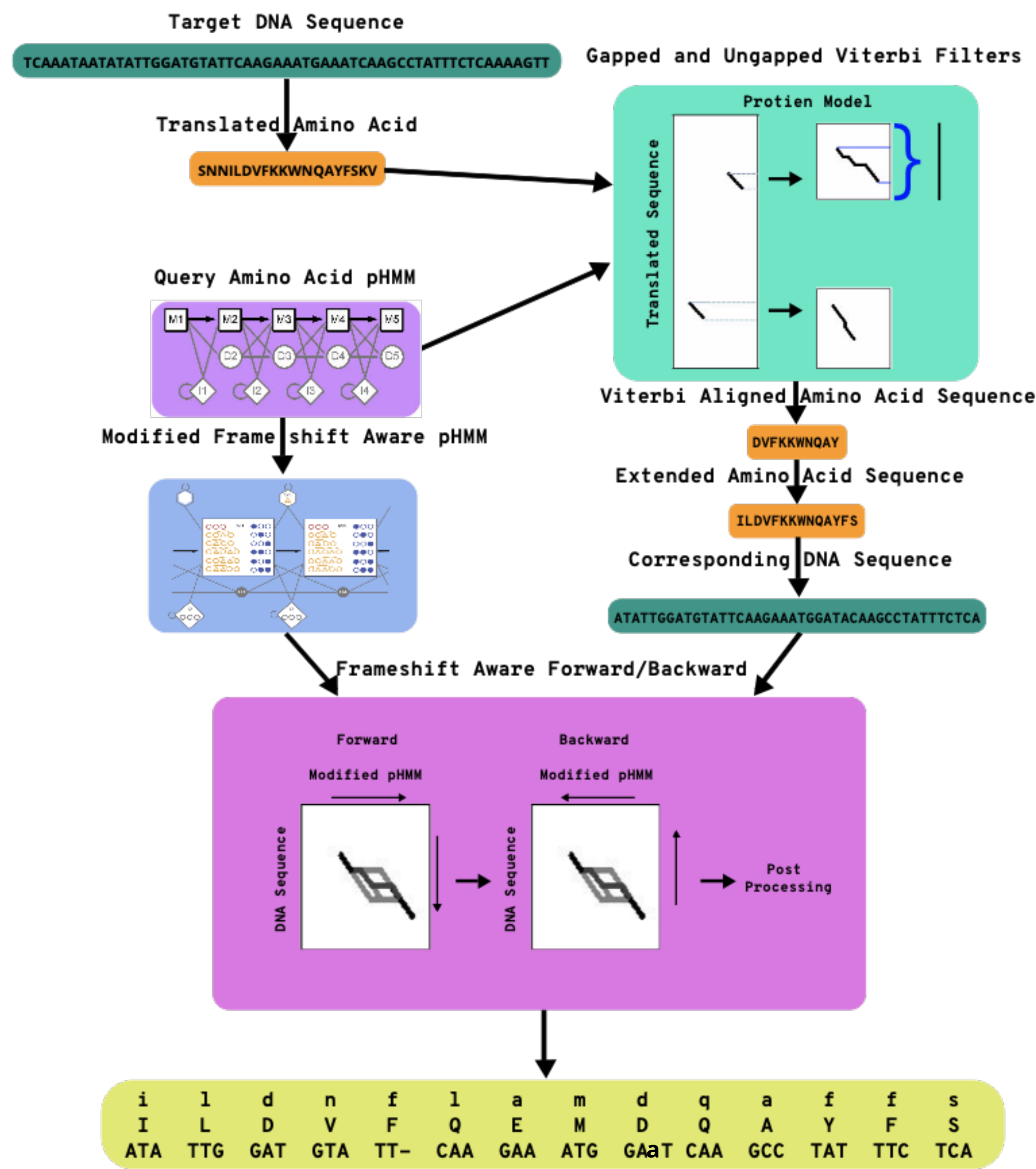
## Frameshift Pipeline

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 comparisons. 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.

Figure 5

CobN 255	1	t	k	l	e	l	v	q	*	k	l	e	l	s	e	i	e	G	k	v	f	v	r	a	l	g	f	a	s	v	283	
MAGTRE 181	TTT	ACA	ACA	CTA	GAA	GTA	GTT	AAA	GAA	AGA	TAT	GAG	TAT	CAT	AA	AAT	GAA	GAT	AAG	ATA	TTT	ATA	AAA	GCT	ATT	GG	TAT	ATT	AAG	ACC	268	
CobN 284	k	k	n	v	k	s	g	a	y	a	r	a	l	f	r	a	v	a	n	r	l	n	f	v	v	k	v	l	l	r	w	313
MAGTRE269	AAA	CCT	GTG	TTA	ATT	AAA	TCT	ATC	TAT	AAC	AAC	AAG	ATG	GCA	TTT	CTA	---	ATC	ATA	GTT	AAT	TAT	ATT	ATCG	TCG	TTA	TTG	ATA	AAT	CAA	353	
CobN 314	l	e	l	k	l	k	a	r	a	l	v	f	v	l	a	n	y	P	v	s	d	s	r	i	g	n	g	v	g	l	343	
MAGTRE 354	ATT	GGA	TTA	AAG	-T-	CTA	ACC	TCA	CAA	ATG	GTT	TAC	ATC	TTA	GCT	AAT	TAT	CTCT	CTG	GAT	AAT	TCA	AAG	ATT	GAT	GAT	AAT	GTT	GAT	TCA	442	
CobN 344	d	t	i	a	s	v	v	n	i	h	a	l	l	v	r	q	k	l	l	t	r	v	e	l	v	n	e	l	l	f	373	
MAGTRE 443	AAT	ACA	TCG	AAA	AGA	ATA	GTT	ACC	AT-	CAA	CAG	TAT	GTT	GAA	CAA	ATG	ATA	CTT	TTG	ATG	ATT	AGG	TCT	TTA	ATA	GAG	GTA	TTA	ATT	GAT	531	
CobN 374	G	v	t	n	a	s	n	l	n	r	v	v	r	v	s	v	c	l	r	a	a	e	v	*	*	*	s	d	-	-	400	
MAGTRE 532	GGT	ATA	ACA	AAT	AAA	ACC	AAC	TTA	AAC	AGA	TCA	ATT	TGT	TAT	GAT	ATA	ATG	AGC	TTA	AAC	TTG	ATTA	ATA	TTA	ATA	AAT	TAA	TTG	GAT	ACC	AAA	623
CobN 481	a	y	k	k	l	r	l	w	G	k	a	e	v	d	P	f	v	v	n	e	f	s	s	l	s	v	l	k	l	a	k	430
MAGTRE 624	TAT	AAA	AAA	CTG	AGA	ATC	CTG	AGA	ATC	ATG	AGT	CCA	GAT	ATA	TTT	ATA	ATCT	TCA	GGA	TGT	TAT	AAC	TTA	CAA	CTA	TTA	GAC	TTG	GAA	CAA	714	
CobN 431	s	y	v	v	q	P	t	r	G	y	G	l	v	n	P	s	y	h	s	a	f	v	v	P	c	k	f	y	-	-	460	
MAGTRE 715	ATC	ATG	GTA	TTA	ATT	TAA	TCA	ATT	AGA	GAC	TAT	GGC	TTG	ATA	AAT	CCA	AGT	ATA	TAT	TAC	CTA	CAT	CCG	TTG	TAT	CCT	ACG	GGT	TATT	TAT	805	
CobN 461	w	l	s	y	v	f	f	*	*	e	r	v	s	a	n	a	l	l	i	n	v	G	k	h	G	s	l	e	w	l	488	
MAGTRE 806	TAT	TTA	AGT	TAT	TTG	TTT	ATT	TTA	AAA	AAA	AGA	ACT	AAT	TCT	AGGCT	ACT	TTA	ATA	TTA	AAC	ATT	AAT	AAA	TAT	AAT	AAT	CAA	AGA	ATA	AAT	897	
CobN 489	P	G	k	a	n	v	l	s	r	s	c	y	P	e	q	l	v	g	l	G	l	p	n	l	y	l	y	i	v	n	d	518
MAGTRE 898	GAC	ACA	CTG	ATC	TGT	AGC	TCTAA	TTG	AGA	T-	TAT	TCA	GAA	CAA	ATT	GCA	AGA	TCT	ACA	ACA	AAT	GCT	TAC	CTC	CGT	ATT	GTT	AAT	AAT	987		
CobN 519	P	G	e	g	t	q	a	k	r	r	i	s	v	i	v	a	s	l	l	l	l	l	l	l	l	l	l	l	l	a	d	548
MAGTRE 988	CTT	GAT	ATAG	AAT	AAA	CAT	ATT	AGG	AGA	ATGAA	TAT	TAT	TCA	ATA	ATC	ATT	GAC	CAG	GAT	ATA	CTA	CTAG	TTG	CTG	ATC	ATC	CGA	CAA	TCT	TTG	1081	
CobN 549	c	r	v	g	s	l	s	l	r	q	r	v	e	s	s	k	y	y	c	n	l	l	g	l	q	f	f	r	s	g	578	
MAGTRE 1082	ATT	GAA	AAA	AGG	GAG	ATG	CTAA	CTT	AAT	TGC	TAT	CCC	AAA	TGA	GGG	AAT	GAA	TAT	AAC	AAC	ATG	TTG	AAT	TCA	--A	CAA	TTC	AAA	TGT	AAA	1170	
CobN 579	l	h	v	y	g	h	v	p	g	s	l	l	l	l	l	l	l	l	l	l	l	l	l	l	l	l	l	l	l	604		
MAGTRE 1171	L	H	Y	U	C	E	L	S	L	P	E	I	Y	I	Y	I	Y	I	Y	I	Y	I	Y	I	Y	I	Y	I	Y	I	Q	1260
CobN 605	k	v	a	a	i	c	k	r	i	k	a	w	s	r	r	s	-	-	-	-	-	-	-	-	-	-	-	-	-	-	620	
MAGTRE 1261	G	F	V	N	L	S	L	N	I	N	Q	Q	H	R	S	H	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1308	
	GGT	TTT	GTT	AAT	TTA	AGT	TTA	AAT	ATC	AAC	CAA	CAA	CAT	GGA	TCA	CAT	-	-	-	-	-	-	-	-	-	-	-	-	-	-		

Found by Frameshift Aware Implementation

Found by phmmert

Codon with insertion

Codon with deletion