FATHMM: Frameshift Aware Translated Hidden Markov Models

Genevieve Krause,[1*] (genevieve.krause@umontana.edu),
Travis Wheeler[1]

[1]University of Montana, Missoula.

Project Repository: https://github.com/TravisWheelerLab/hmmer/tree/frameshift

Project Goals

Frameshifts can occur in protein coding regions of DNA sequences due to natural processes such as pseudogenization, in which a gene that has fallen out of use acquires mutations in the normal course of replication and repair. They can also result from errors made during DNA sequencing, particularly when using newer long-read sequencers [1]. Improving the annotation of frameshifted sequences is therefore an important step in improving the annotation of both highly decayed pseudogenes [2], and microbial metagenomics dataset that increasingly rely on long read sequencers for assembly [3]. In pursuit of this goal, we have created FATHMM, a sequence similarity search tool that produces accurate translated alignments between protein profile hidden Markov models and DNA sequences containing frameshifts.

Abstract

High-quality annotation of DNA typically relies on sequence alignment to produce evidence of evolutionary relationships between sequences. Annotation of protein coding DNA can be achieved through translated alignment, in which the DNA is aligned to known proteins by translating codons into amino acids. In DNA that contains errors, the correct open reading frame can be obscured by frameshift-inducing insertions or deletions, preventing accurate translation (Fig 1). By explicitly modeling frameshifts within codons, FATHMM allows translated alignments to accommodate these errors without changing the translation of the subsequent codons. This is achieved through the use of a frameshift aware hidden Markov model (Fig 2), supported by dynamic programing algorithms that allow for variable length codons. The result is significant improvement in the sensitivity and specificity of translated alignments for frameshifted DNA (Fig 3), with only a moderate increase in overall alignment run time (Tbl 1).

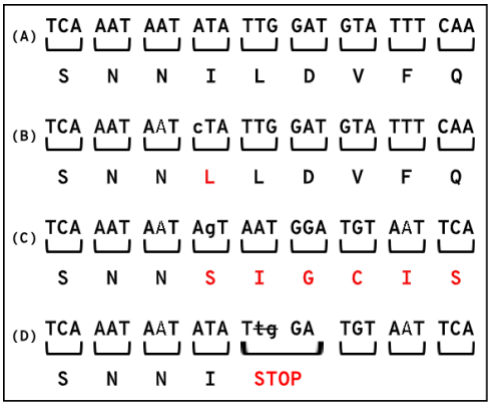

Figure 1. Indels break open reading frames.

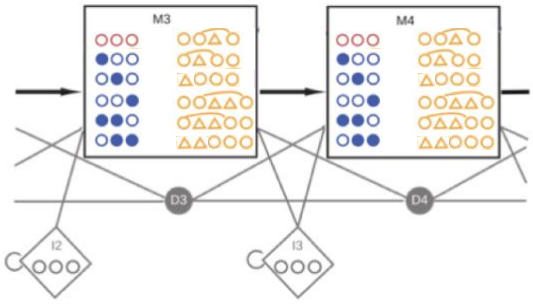

Figure 2. Portion of Frameshift-aware profile hidden Markov model.

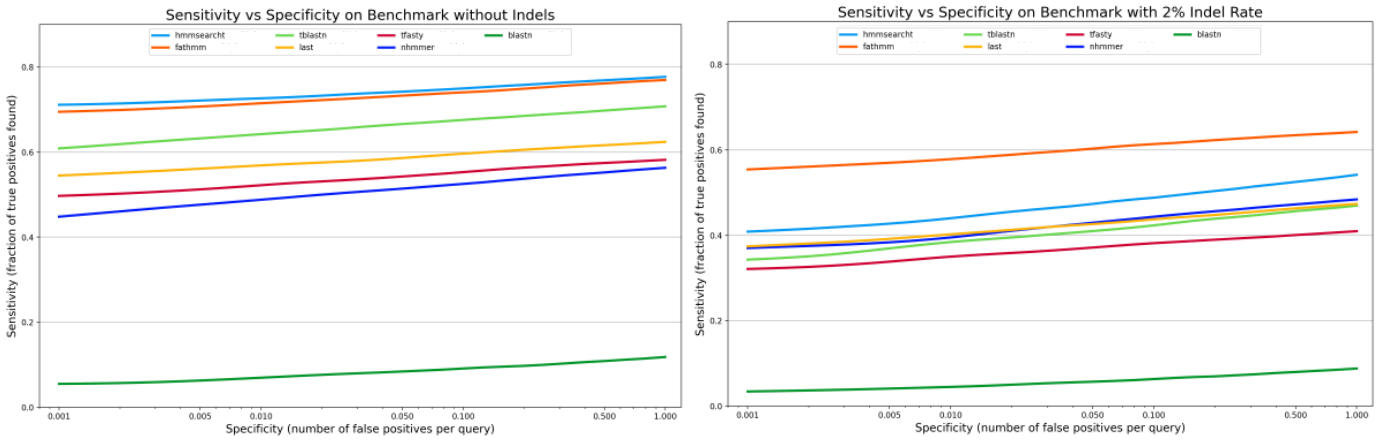

Figure 3. Sensitivity as a function of false-positive annotation for various translated search tools, on benchmarks without (left) and with (right) frameshift-inducing insertions and deletions (indels). FATHMM (orange) is much more sensitive when indels are present, but not prone to error when indels are not present.

Benchmark Runtime (Hours)							
Indel Rate	FATHMM	LAST	TFASTY	hmmsearcht	tBLASTn	nhmmer	BLASTn
0%	68.90	7.43	20.96	11.91	5.95	29.33	5.33
2%	66.65	6.41	17.09	8.89	3.62	26.16	9.02

Table 1. Runtimes of various tools tested in the frameshift benchmark from Fig 3.

References

1. Watson, M., & Warr, A. (2019). Errors in long-read assemblies can critically affect protein prediction, *37*(February), 124–126.

2. Cheetham, S. W., Faulkner, G. J., & Dinger, M. E. (2020). Overcoming challenges and dogmas to understand the functions of pseudogenes. *Nature Reviews Genetics*, *21*(3), 191–201. https://doi.org/10.1038/s41576-019-0196-1

3. Bharti, R., & Grimm, D. G. (2019). Current challenges and best-practice protocols for microbiome analysis, *00*(October), 1–16. https://doi.org/10.1093/bib/bbz155

Funding statement.
This work was supported by NIH grant R15 GM123487 and DOE grant DE-SC0021216.