COATi: statistical pairwise alignment of protein coding sequences

Juan J. Garcia Mesa[1,2], Ziqi Zhu[1,3], Reed A. Cartwright[1,3]

1 Biodesign Institute, Arizona State University
2 Ira A. Schools of Engineering, Arizona State University
3 School of Life Sciences, Arizona State University

* corresponding@author.mail

Abstract

Summary: COATi is a statistical codon-aware pairwise aligner that supports complex insertion-deletion models and is able to handle artifacts present in genomic data.
Availability: The source code for COATi, along with documentation, is freely available on GitHub: https://github.com/CartwrightLab/coati and is implemented in C++.
Supplementary information:

1 Introduction

Sequence alignment is a cornerstone step in bioinformatics (Rosenberg 2009). Uncorrected errors in sequence alignment can lead to erroneous results in functional and comparative genomic studies (Schneider *et al.* 2009). Given that errors and artifacts are common in molecular data, this requires costly curation practices that discard large amounts of information. In addition, a common strategy is to perform alignment inference in the amino acid space (Bininda-Emonds, Olaf 2005; Abascal *et al.* 2010). While this approach is an improvement over DNA models, it discards information, underperforms compared to alignment at the codon level, and fails in the presence of artifacts such as frameshifts and early stop codons. Although some aligners incorporate codon substitution models, they do not support frameshifts or lack as statistical model.

To address this problem, we present COATi, short for COdon-Aware Alignment Transducer, a statistical pairwise aligner that incorporates codon substitution models and is robust to artifacts present in genomic data.

2 Description

Statistical alignment is typically performed using pairwise hidden Markov models (pair-HMMs). Pair-HMMs are computational machines that have the ability to rigorously model molecular sequence evolution and can calculate the probability that two sequences are related, represented $P(X,Y)$ (Yoon 2009). However, a limitation of pair-HMMs is the ability to only model the evolution of two related sequences from an unknown ancestor.

Finite-state transducers (FSTs) have the same benefits as pair-HMMS with the additional ability to generate a descendant sequence given an ancestral one. Properly weighted, an FST can calculate the probability that a sequence Y (descendant) evolved from sequence X (ancestor), represented $P(Y|X)$. Furthermore, the existence of well-established algorithms for combining FSTs in different ways (Bradley & Holmes 2007) allows the design of complex models by combining

simpler FSTs. A powerful and versatile algorithm for comparative sequence analysis is composition, which consists of sending the output of one FST into the input of a second FST. The FST model implemented in COATi is designed by composing smaller FSTs, each representing a specific process.

Pairwise alignment in COATi is implemented via the Evolution FST (Fig. 1), based on existing transducers (e.g. Holmes & Bruno 2001). The Evolution FST is formed by composing a substitution FST that encodes a 64x64 codon model (Fig. 1-a) and an indel FST that models insertions and deletions, including frameshifts (Fig. 1-b). The substitution models available are Muse and Gaut (Muse & Gaut 1994) (MG94) and the Empirical Codon Model (Kosiol *et al.* 2007) (ECM). The innovation of the Evolution FST with respect to other transducers is the combination of a codon substitution model that allows stop codons with gaps that can occur at any position of any length.

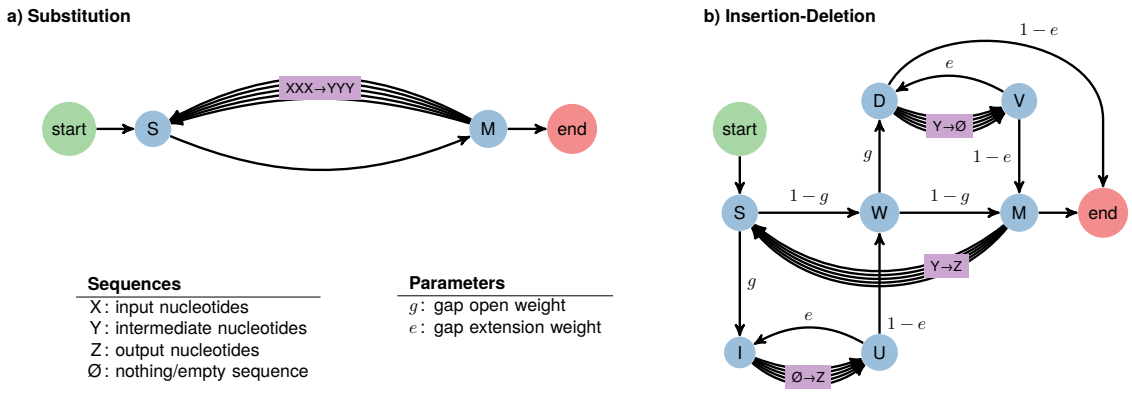

Figure 1: The Evolution FST is assembled by composing a substitution FST and an indel FST. Each node represents a state in an FST while arcs display possible transitions between states (and their weights). (a) The substitution FST encodes a 61x64 codon substitution model with 61*64 arcs from M to S. (b) The indel FST allows for insertions (I to U) and deletions (D to V). Contiguous insertions and deletions are always arranged for insertions to precede deletions to limit equivalent alignments.

The alignment FST is the result from composing both sequences with the Evolution FST. Any path through the alignment FST represents a pairwise alignment, while the shortest path corresponds to the best alignment. All FST operations including model development, composition, search for the shortest path, and other optimization algorithms are performed using the C++ openFST library (Allauzen *et al.* 2007).

COATi also features a marginal substitution model with substitution probability matrix

$$P'_{ijp} = \sum_{cod} \begin{cases} P(i\,|\,cod) & \text{if } cod_p = j \\ 0 & \text{otherwise} \end{cases}$$

Where P'_{ijp} represents the probability that codon i from the ancestor sequence changes to nucleotide j of the descendant sequence at position $p \in \{0,1,2\}$ of the reading frame. P is the

substitution probability matrix MG94 or ECM. This model emphasizes the position where the substitution in a codon occurs and helps restrict the effects of low quality data in the descendant sequence. In combination with the indel FST alignment with the marginal model can be implemented using dynamic programming, which results in a significant speed up with similar accuracy.

3 Methods

Using 16000 human genes and their gorilla homologous pairs from the ENSEMBL database (Hubbard *et al.* 2002) we simulated a data set of pairwise alignments with empirical gap patterns. We used the data set to evaluate the accuracy of popular cutting edge aligners ClustalΩ v1.2.4 (Sievers *et al.* 2011), MACSE v2.06 (Ranwez *et al.* 2011), MAFFT v7.407 (Katoh *et al.* 2002), and PRANK v.170427 (Löytynoja 2014) together with COATi fst and marginal models.

After downloading, sequences longer than 6000 nucleotides were filtered out (2232), and 8369 alignments contained gaps identified by at least one aligner. We randomly introduced gap patterns extracted from all five methods into the 5399 initially ungapped sequence pairs to generate the true alignments. Alignment accuracy was measured using the distance metric d_{seq} (Blackburne & Whelan 2011) between simulated and inferred alignments. In addition, accuracy of positive and negative selection was calculated using the F_1 score by estimating k_s and k_a statistics (W.-H. Li 1993).

4 Results

Both models in COATi were significantly ($p < 2.2x10^{-16}$) more accurate at inferring simulated alignments compared to other aligners according to the one-tailed Wilcoxon signed ranked test, with an average alignment error (d_{seq}) value of $1x10^{-3}$. In addition, COATi produced more perfect alignments ($d_{seq} = 0$), less imperfect alignments, and more accurately retrieved events of positive selection (fst model 91.9%, marginal model 90.8%) (Supplementary Table 1). Compared to COATi, MACSE (allows frameshifts) and MAFFT (using DNA) had an average alignment error an order of magnitude larger than COATi and a lower accuracy retrieving events of positive selection at a rate of 81.5% and 85.8% respectively. PRANK (using codons) and ClustalΩ (using amino acid translations) had an average alignment error two orders of magnitude larger than COATi and a 87.3% and 69.1% accuracy retrieving events of positive selection, respectively. The accuracy retrieving events of negative selection was similar across all five aligners (97.7% ± 1.67%).

	COATi	**PRANK**	**MAFFT**	**ClustalΩ**	**MACSE**
Avg alignment error (d_{seq})	0.00101	0.01010	0.00982	0.01582	0.00932
Perfect alignments	2452	22	2175	1150	1580
Best alignments	3624	155	2763	1609	2081
Imperfect alignments	1136	3566	1413	2438	2008
F1 score of positive selection	90.8%	80.5%	73.5%	61.3%	70.6%
F1 score of negative selection	99.1%	98.0%	97.2%	96.0%	97.4%

Table 1: Accuracy of COATi, PRANK, MAFFT, ClustalΩ, and MACSE, on 5399 simulated sequence pairs. Perfect alignments have ($d_{seq} = 0$), best alignments have lowest d_{seq}, and imperfect alignments have $d_{seq} > 0$ when at least one aligner found a perfect alignment.

5 Discussion

COATi is an FST-based application that can calculate the optimal alignment between a pair of sequences in the presence of artifacts using a statistical model. It will allow researchers to analyze more data with higher accuracy and facilitate the study of important biological processes that shape genomic data.

Acknowledgments

Funding

This research was funded by an NSF-IIBR grant (grant number).

Conflict of interest: none declared.

References

1. Abascal, F., Zardoya, R. & Telford, M. J. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic acids research* **38,** W7–W13 (2010).
2. Allauzen, C., Riley, M., Schalkwyk, J., Skut, W. & Mohri, M. *OpenFst: A general and efficient weighted finite-state transducer library* in *International Conference on Implementation and Application of Automata* (2007), 11–23.
3. Bininda-Emonds, Olaf. transAlign: using amino acids to facilitate the multiple alignment of protein-coding DNA sequences. *BMC bioinformatics* **6,** 1–6 (2005).
4. Blackburne, B. P. & Whelan, S. Measuring the distance between multiple sequence alignments. *Bioinformatics* **28,** 495–502. ISSN: 1367-4803. eprint: https://academic.oup.com/bioinformatics/article-pdf/28/4/495/563214/btr701.pdf. https://doi.org/10.1093/bioinformatics/btr701 (Dec. 2011).
5. Bradley, R. K. & Holmes, I. Transducers: an emerging probabilistic framework for modeling indels on trees. *Bioinformatics* **23** (2007).
6. Holmes, I. & Bruno, W. J. Evolutionary HMMs: a Bayesian approach to multiple alignment. *Bioinformatics* **17,** 803–820 (2001).
7. Hubbard, T. *et al.* The Ensembl genome database project. *Nucleic acids research* **30,** 38–41 (2002).
8. Katoh, K., Misawa, K., Kuma, K.-i. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic acids research* **30,** 3059–3066 (2002).
9. Kosiol, C., Holmes, I. & Goldman, N. An empirical codon model for protein sequence evolution. *Molecular biology and evolution* **24,** 1464–1479 (2007).
10. Li, W.-H. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *Journal of molecular evolution* **36,** 96–99 (1993).
11. Löytynoja, A. in *Multiple sequence alignment methods* 155–170 (Springer, 2014).

12. Muse, S. V. & Gaut, B. S. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular biology and evolution* **11,** 715–724 (1994).

13. Ranwez, V., Harispe, S., Delsuc, F. & Douzery, E. J. MACSE: Multiple Alignment of Coding SEquences accounting for frameshifts and stop codons. *PloS one* **6,** e22594 (2011).

14. Rosenberg, M. S. *Sequence alignment: methods, models, concepts, and strategies* (Univ of California Press, 2009).

15. Schneider, A. *et al.* Estimates of positive Darwinian selection are inflated by errors in sequencing, annotation, and alignment. *Genome biology and evolution* **1,** 114–118 (2009).

16. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular systems biology* **7,** 539 (2011).

17. Yoon, B.-J. Hidden Markov models and their applications in biological sequence analysis. *Current genomics* **10,** 402–415 (2009).