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

Sequence alignment is an essential method in bioinformatics and the basis of many analyses including phylogenetic inference, ancestral sequence reconstruction, and gene annotation. Errors made in alignment reconstruction can impact downstream analyses leading to erroneous conclusions in comparative and functional genomics. Common artifacts in unpublished reference genomes include abiological frameshifts and early stop codons. While for model organisms this is eventually fixed, for many species this is not the case and current curation efforts require discarding large amounts of data. To address this need, we present COATi, a statistical codon-aware pairwise aligner that supports complex insertion-deletion models and can handle artifacts present in genomic data. COATi will allow users to reduce the amount of discarded data while generating more accurate sequence alignments.

Introduction

Sequence alignment is a fundamental task in bioinformatics and a cornerstone step in comparative and functional genomic analysis (Rosenberg 2009). Modern sequence analysis began with the heuristic homology algorithms of Needleman and Wunsch in 1970 (Smith, Waterman, *et al.* 1981) and, while the methods developed since then have improved, alignment inference is not a solved problem (Morrison 2015).

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 a statistical model. In addition, while modeling indels to appear within codons is rare, this is often the case (Zhu & Cartwright, personal communication, 2019). Considering gaps to only appear between codons can result in missing the optimal alignment and inflate estimates of sequence divergence (Fig. 1).

Uncorrected errors in the alignment stage can lead to erroneous results in comparative and functional genomic studies (Schneider *et al.* 2009). Current methods are ill-equipped to handle common artifacts in genomic data, requiring costly curation practices that discard significant amounts of information. To address this problem, we present COATi, short for COdon-aware Alignment Transducer, a pairwise statistical aligner that incorporates codon substitution models and is robust to artifacts present in modern genomic data.

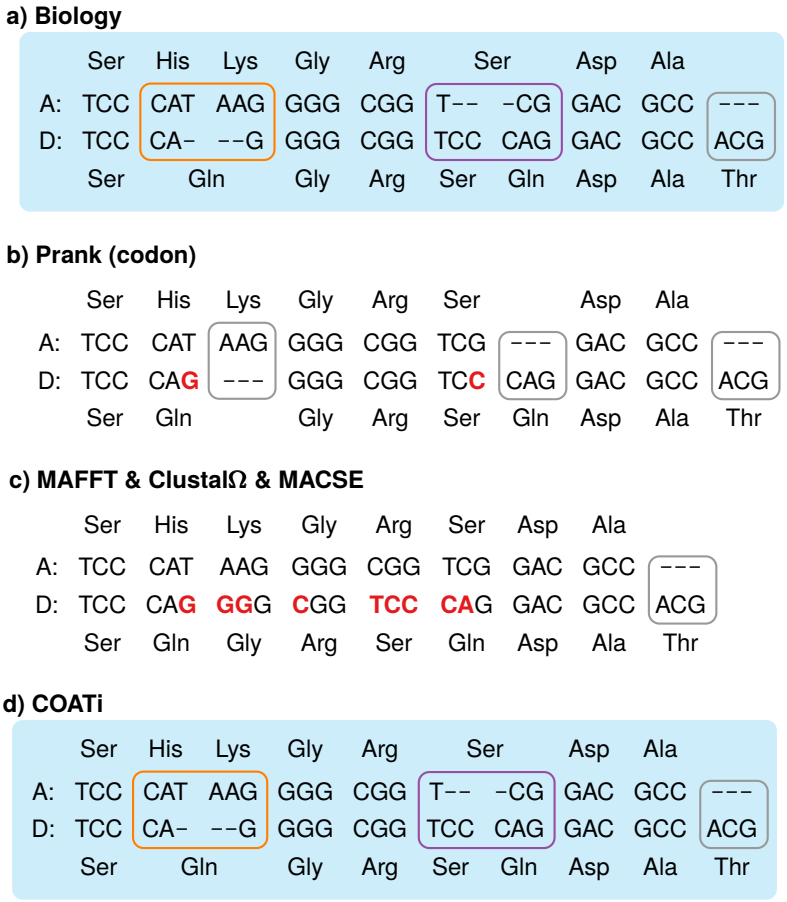

Figure 1: Standard algorithms produce suboptimal alignments. (a) shows a possible alignment of an ancestor (A) sequence and a descendant (D) sequence. (b), (c), and (d) are the results of different aligners. Nucleotide mismatches are highlighted in red, phase zero, one, and two indels are shown in gray, purple, and orange, respectively. Phase one indel (purple) is also of type I (synonymous) while phase two indel (orange) is of type II (nonsynonymous). COATi is able to retrieve the biological alignment (highlighted in blue).

Materials and Methods

Statistical alignment is typically performed using pairwise hidden Markov models (pair-HMMs), which have the ability to rigorously model molecular sequence evolution (Bradley & Holmes 2007). Pair-HMMs are computational machines with two output tapes that contain a finite number of states typically labeled match, insert, and delete that emit symbols (nucleotides or amino acids) to one or both tapes. Each tape represents a sequence and a path through a pair-HMM is a possible pairwise alignment. Conceptually, these machines generate two sequences (X and Y) from an unknown ancestor and can calculate the probability that two sequences are related, represented by $P(X,Y)$ (Yoon 2009).

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 similar benefits to pair-HMMs with the additional feature to generate a descendant sequence given an ancestral one. FSTs consume symbols from an input tape and emit symbols to an output tape. Properly weighted, an FST can calculate the probability that a descendant sequence Y evolved from an ancestor sequence X, represented by $P(Y|X)$. Furthermore, well-established algorithms for combining FSTs in different ways allow the design of complex models by combining simpler FSTs (Bradley & Holmes 2007). 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 model implemented in

COATi is designed by composing smaller FSTs, each representing a specific process.

Genomes for model organisms are often of high quality after being refined over many iterations and having their coding sequences meticulously curated. On the contrary, non-model organisms typically have lower-quality genomes that have been only partially curated. Low-quality genomes often lack the amount of sequencing data needed to fix artifacts, including missing exons, erroneous mutations, and indels (Jackman *et al.* 2018). FSTs and their powerful methods provide a well-suited framework to statistically align a sequence from a non-model organism against a sequence from a model organism.

COATi implements the pairwise alignment of a low-quality sequence (descendant) against a high-quality sequence (ancestor) as a path through the Evolution FST (Fig. 2), based on existing transducers (e.g. Holmes & Bruno 2001). This FST is the result of composing a substitution FST that encodes a codon model (Fig. 2-a) and an indel FST that models insertions and deletions, including frameshifts (Fig. 2-b). A key innovation of this FST with respect to others is the combination of a codon substitution model with a nucleotide-based geometric indel model that allows gaps to occur at any position.

Composing both sequences with the Evolution FST results in the transducer of all possible alignments. Any path through this FST represents a pairwise alignment, while the shortest path corresponds to the best alignment. All FST operations in COATi, including model development, composition, search for the shortest path, and other optimization algorithms, are performed using the C++ openFST library (Allauzen *et al.* 2007).

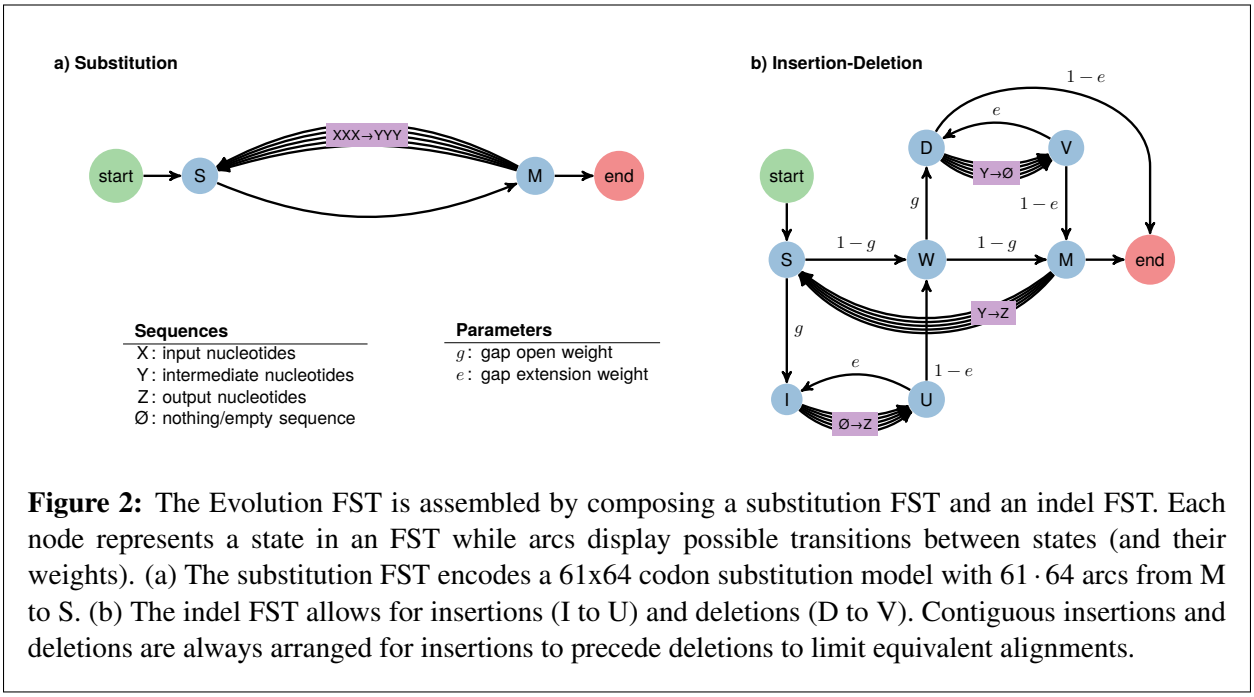

Figure 2: The Evolution FST is assembled by composing a substitution FST and an indel FST. Each node represents a state in an FST while arcs display possible transitions between states (and their weights). (a) The substitution FST encodes a 61x64 codon substitution model with $61 \cdot 64$ arcs from 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.

Codon substitution models are uncommon in sequence aligners, despite their extensive use in phylogenetics. COATi implements the Muse and Gaut (1994) codon model (codon-triplet-mg) and the Empirical Codon Model (Kosiol *et al.* 2007) (codon-triplet-ecm). It also lets the user provide a codon substitution matrix. The default FST model (codon-triplet-mg) does not allow substitutions

from stop codons, although it supports mutations to (early) stop codons under the assumption that these are artifacts common in low-quality data.

COATi also features a marginal substitution model with 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 Muse and Gaut or the Empirical Codon Model, resulting in the marginal models codon-marginal-mg (default model) and codon-marginal-ecm. These models emphasize the position where the substitution in a codon occurs, help restrict the effects of low-quality data in the descendant sequence, and allow more than one substitution per codon. In combination with the indel model, alignment using the marginal model is implemented using dynamic programming.

Results and Discussion

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.

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 benchmark 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).

	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.

COATi was significantly more accurate (lower d_{seq}) at inferring simulated alignments compared to other methods; all p-values were less than $2.2 \cdot 10 - 16$ according to the one-tailed

Wilcoxon signed rank test. In addition, COATi produced more perfect alignments, less imperfect alignments, and more accurately retrieved events of positive selection (Table 1). It obtained better results compared to a wide variety of alignment strategies. ClustalΩ, performing a common approach of aligning via amino acid translations, obtained the highest average alignment error and had difficulties retrieving positive selection. MACSE, which allows frameshifts, is also based on an amino acid model and obtained similar results to the DNA-based MAFFT. PRANK, using a codon model, had a similar average alignment error to MACSE and MAFFT but had issues recovering the simulated alignments.

Despite human and gorilla sequences having a relatively short evolutionary distance, COATi showed a biologically significant improvement over other methods, with an average alignment error nine-fold smaller than the next best method. 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.

Availability

The source code for COATi, along with documentation, is freely available on GitHub: https://github.com/CartwrightLab/coati and is implemented in C++. Code to replicate the analysis can be found on GitHug: https://github.com/jgarciamesa/coati-testing.

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. Jackman, S. D. *et al.* Tigmint: correcting assembly errors using linked reads from large molecules. *BMC bioinformatics* **19,** 1–10 (2018).

9. 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).

10. Kosiol, C., Holmes, I. & Goldman, N. An empirical codon model for protein sequence evolution. *Molecular biology and evolution* **24,** 1464–1479 (2007).

11. Li, W.-H. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *Journal of molecular evolution* **36,** 96–99 (1993).

12. Löytynoja, A. in *Multiple sequence alignment methods* 155–170 (Springer, 2014).

13. Morrison, D. A. Is sequence alignment an art or a science? *Systematic Botany* **40,** 14–26 (2015).

14. 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).

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

16. 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).

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

18. Smith, T. F., Waterman, M. S., *et al.* Identification of common molecular subsequences. *Journal of molecular biology* **147,** 195–197 (1981).

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