COATi: statistical pairwise alignment of protein coding sequences

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

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

* cartwright@asu.edu

Abstract

Sequence alignment is an essential method in bioinformatics and the basis of many analyses, including phylogenetic inference, ancestral sequence reconstruction, and gene annotation. Sequence artifacts and errors made in alignment reconstruction can impact downstream analyses leading to erroneous conclusions in comparative and functional genomic studies. For example, abiological frameshifts and early stop codons are common artifacts found in protein coding sequences that have been annotated in reference genomes. While such errors are eventually fixed in the reference genomes of model organisms, many genomes used by researchers contain theses artifacts, and researchers often discard large amounts of data in comparative genomic studies to prevent artifacts from impacting results. 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 allows 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). While sophisticated advancements have been made, the challenge of alignment inference has not been fully solved (Morrison 2015). The alignment of protein coding DNA sequences is one such challenge, and a common approach to this problem is to perform alignment inference in amino-acid space (e.g. 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, existing aligners typically force gaps to occur between codons, whereas in natural sequences, only about 40% of indels occur between codons (Taylor et al. 2004; Zhu 2022). This mismatch between aligner assumptions and biology can produce sub-optimal alignments and inflated estimates of sequence divergence (Fig. 1).

Genome quality impacts conclusions drawn from comparative genomic studies, and uncorrected errors in the alignment stage can lead to erroneous results in comparative and functional genomic studies (Schneider et al. 2009; Fletcher and Yang 2010; Hubisz et al. 2011). Genomes for model organisms often get refined over many iterations and achieve high quality with meticulously curated protein coding sequences. In contrast, genomes for non-model organisms might

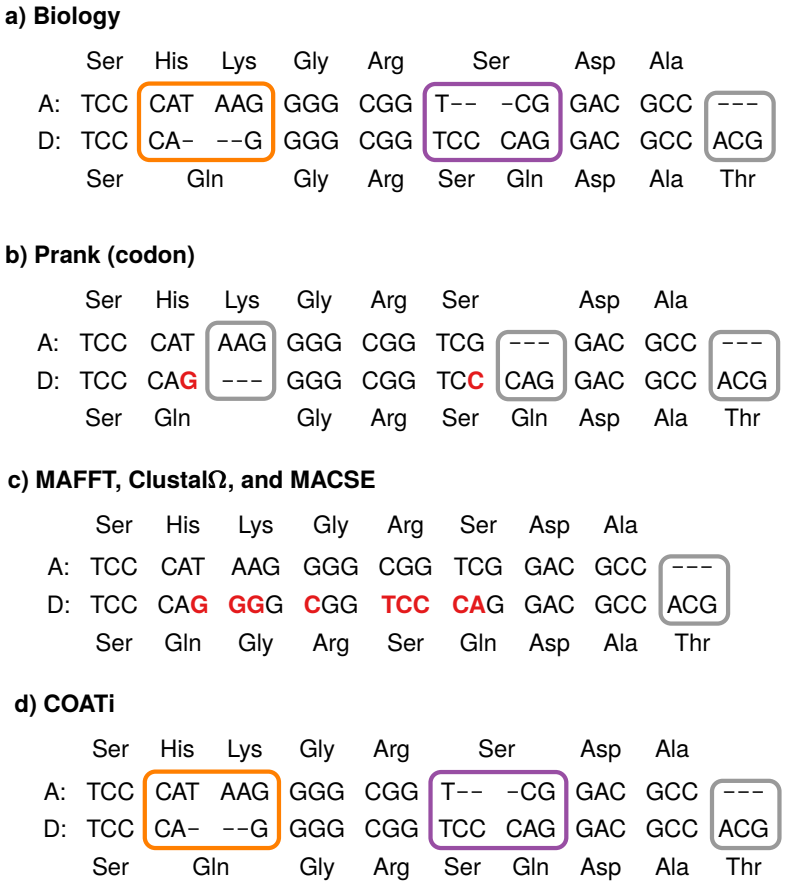

Figure 1: Standard algorithms produce suboptimal alignments. (a) shows a possible alignment of an ancestor sequence (A) and a descendant sequence (D). (b), (c), and (d) are the results of different aligners. Nucleotide mismatches are highlighted in red. Phase 0, phase 1, and phase 2 indels are shown in gray, purple, and orange, respectively. Additionally, the orange indel is type II (an amino-acid indel plus an amino-acid change) while the purple indel is type I (an amino-acid indel only). COATi is the only aligner able to retrieve the biological alignment in this example.

only receive partial curation and typically have lower quality sequences and annotations. These genomes often lack the amount of sequencing data needed to fix artifacts, including missing exons, erroneous mutations, and indels (Jackman et al. 2018). When comparative and functional genomics studies include data from non-model organisms, care must be taken to identify and manage such artifacts; however, current alignment 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.

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 and Holmes 2007). Pair-HMMs are computational machines with two output tapes. Pair-HMMs 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 $P(\mathrm{X}, \mathrm{Y})$ (Yoon 2009).

A limitation of pair-HMMs is that they 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 that can can model the generation 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 ancestral sequence, X, represented by $P(\mathrm{Y}|\mathrm{X})$. Furthermore, well-established algorithms for combining FSTs in different ways allow the design of complex models by combining simpler FSTs (Bradley and 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. COATi uses composition to derive a statistical alignment model from the combination of smaller FSTs, each representing a specific process.

COATi implements the pairwise alignment of a potentially lower-quality sequence against a high-quality sequence as a path through the Evolution FST (Fig. 2) (c.f. Holmes and Bruno, 2001). Here, COATi treats the high-quality (reference) sequence as the "ancestor" and the potentially lower-quality sequence as the "descendant". 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 (by weight) 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). However, the Evolution FST has a large state space to keep track of codon substitution rates when codons might be interspersed with indel events. This additional state space increases the computational complexity of the alignment algorithm.

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

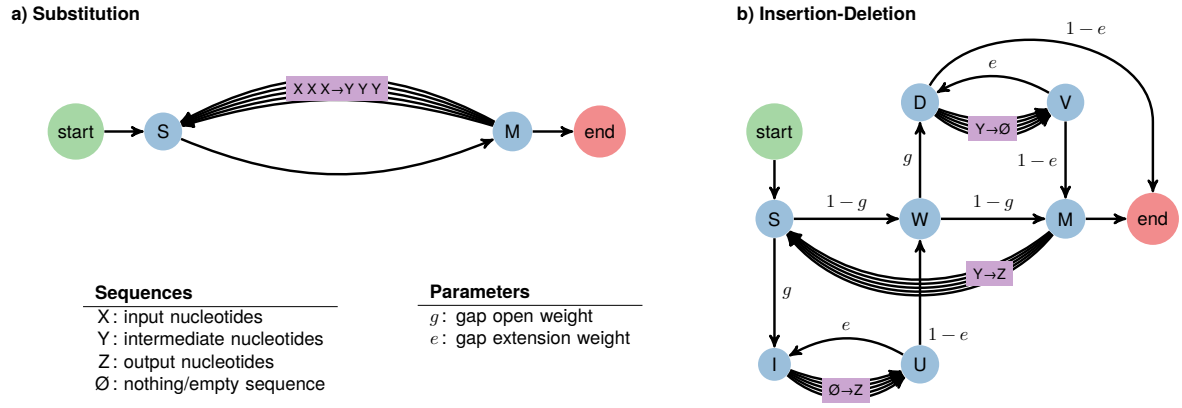

Figure 2: The Evolution FST is assembled by composing a substitution FST and an indel FST. Each node represents a state in an FST while arcs display possible transitions between states (and their weights). Unlabeled arcs have weights of 1. (a) The substitution FST encodes a 61×64 codon substitution model with 3904 arcs from M to S. These arcs consume three nucleotides from the input tape and emit three nucleotides to the output tape. The weight of each arc is a conditional probability derived from a codon substitution model. (b) The indel FST allows for insertions (I to U) and deletions (D to V). Insertion arcs are weighted according to the codon model's stationary distribution of nucleotides, and deletion arcs have a weight of 1. Contiguous insertions and deletions are always arranged for insertions to precede deletions to limit equivalent alignments.

To reduce the runtime complexity of COATi, we have also developed an approximation of the Evolution FST that can be implemented with standard dynamic programming techniques. This approximation uses a marginal substitution model where the output nucleotides are independent of one another and only depend on the input codon and position. This produces a $(61 \times 3) \times 4$ substitution model and eliminates the need to track dependencies between output nucleotides.

A marginal substitution model is calculated from a standard substitution model by calculating the marginal probabilities that each ancestral codon produces specific descendant nucleotides at each reading frame position. Specifically, let $P_{\text{cod}}\left(Y_0 \cdot Y_1 \cdot Y_2 | X_0 \cdot X_1 \cdot X_2\right)$ represent transition probabilities from a standard codon model, and

$$P_{\text{mar}}\left(Y_p = y \middle| X_0 \cdot X_1 \cdot X_2\right) = \sum_{Y_0 \cdot Y_1 \cdot Y_2} I(Y_p = y) P_{\text{cod}}\left(Y_0 \cdot Y_1 \cdot Y_2 | X_0 \cdot X_1 \cdot X_2\right)$$

represent the marginal transition probabilities, where $p \in \{0,1,2\}$ is the position of the descendant nucleotide relative to the ancestral reading frame and I is an indicator function defined $I(e) = \{1 \text{ if } e \text{ is true and } 0 \text{ otherwise}\}$. COATi contains marginal models for both 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 in a codon where the substitution 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 orthologs 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 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. **TODO: which version of coati are we using. Should we show both? Or show both in a supplement?**

After downloading, we removed 2232 sequences longer than 6000 nucleotides, identified 8369 sequence pairs that contained gaps identified by at least one aligner, and 5399 ungapped sequence pairs. We then randomly introduced gap patterns extracted from all five methods into the ungapped sequence pairs to generate the benchmark alignments. Alignment accuracy was measured using the distance metric d_{seq} (Blackburne and 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 (Li 1993). **TODO: briefly explain F1 score**

	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: COATi generates better alignments than other alignment algorithms. Results of COATi, PRANK, MAFFT, ClustalΩ, and MACSE aligning 5399 empirically simulated sequence pairs. Perfect alignments have $d_{seq} = 0$, best alignments have the 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.

TODO: Add one short paragraph about future work. E.g. including a three-mer based indel model, and weighting indel phases differently.

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 GitHub: https://github.com/jgarciamesa/coati-testing.

Acknowledgments

This research was funded by NSF award DBI-1929850.

Conflict of interest: none declared.

References

Abascal F, Zardoya R, and Telford MJ. 2010. Translatorx: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic acids research* 38:W7–W13.

Allauzen C, Riley M, Schalkwyk J, Skut W, and Mohri M. 2007. Openfst: A general and efficient weighted finite-state transducer library. In *International Conference on Implementation and Application of Automata*, pages 11–23. Springer.

Bininda-Emonds, Olaf. 2005. transalign: using amino acids to facilitate the multiple alignment of protein-coding dna sequences. *BMC bioinformatics* 6:1–6.

Blackburne BP and Whelan S. 2011. Measuring the distance between multiple sequence alignments. *Bioinformatics* 28:495–502. ISSN 1367-4803.

Bradley RK and Holmes I. 2007. Transducers: an emerging probabilistic framework for modeling indels on trees. *Bioinformatics* 23.

Fletcher W and Yang Z. 2010. The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. *Molecular biology and evolution* 27:2257–2267.

Holmes I and Bruno WJ. 2001. Evolutionary hmms: a bayesian approach to multiple alignment. *Bioinformatics* 17:803–820.

Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T, et al. 2002. The ensembl genome database project. *Nucleic acids research* 30:38–41.

Hubisz MJ, Lin MF, Kellis M, and Siepel A. 2011. Error and error mitigation in low-coverage genome assemblies. *PloS one* 6:e17,034.

Jackman SD, Coombe L, Chu J, Warren RL, Vandervalk BP, Yeo S, Xue Z, Mohamadi H, Bohlmann J, Jones SJ, et al. 2018. Tigmint: correcting assembly errors using linked reads from large molecules. *BMC bioinformatics* 19:1–10.

Katoh K, Misawa K, Kuma Ki, and Miyata T. 2002. Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic acids research* 30:3059–3066.

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

Li WH. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *Journal of molecular evolution* 36:96–99.

Löytynoja A. 2014. Phylogeny-aware alignment with prank. In *Multiple sequence alignment methods*, pages 155–170. Springer.

Morrison DA. 2015. Is sequence alignment an art or a science? *Systematic Botany* 40:14–26.

Muse SV and Gaut BS. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular biology and evolution* 11:715–724.

Ranwez V, Harispe S, Delsuc F, and Douzery EJ. 2011. Macse: Multiple alignment of coding sequences accounting for frameshifts and stop codons. *PloS one* 6:e22,594.

Rosenberg MS. 2009. *Sequence alignment: methods, models, concepts, and strategies*. Univ of California Press.

Schneider A, Souvorov A, Sabath N, Landan G, Gonnet GH, and Graur D. 2009. Estimates of positive darwinian selection are inflated by errors in sequencing, annotation, and alignment. *Genome biology and evolution* 1:114–118.

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, et al. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Molecular systems biology* 7:539.

Taylor MS, Ponting CP, and Copley RR. 2004. Occurrence and consequences of coding sequence insertions and deletions in mammalian genomes. *Genome research* 14:555–566.

Yoon BJ. 2009. Hidden markov models and their applications in biological sequence analysis. *Current genomics* 10:402–415.

Zhu Z. 2022. Profiling of indel phases in coding regions. Ph.D. thesis, Arizona State University.