

# COATi: statistical pairwise alignment of protein coding sequences

Juan J. Garcia Mesa<sup>1,2</sup>, Ziqi Zhu<sup>1,3</sup>, Reed A. Cartwright<sup>1,3,\*</sup>

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 these 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.

## 1 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 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

**a) Biology**

	Ser	His	Lys	Gly	Arg	Ser	Asp	Ala	
A:	TCC	CAT	AAG	GGG	CGG	T-- -CG	GAC	GCC	---
D:	TCC	CA-	--G	GGG	CGG	TCC CAG	GAC	GCC	ACG
	Ser		Gln	Gly	Arg	Ser	Gln	Asp	Ala
									Thr

**b) Prank (codon)**

	Ser	His	Lys	Gly	Arg	Ser		Asp	Ala
A:	TCC	CAT	AAG	GGG	CGG	TCG	---	GAC	GCC
D:	TCC	CAG	---	GGG	CGG	TCC	CAG	GAC	GCC
	Ser	Gln		Gly	Arg	Ser	Gln	Asp	Ala
									Thr

**c) MAFFT, ClustalΩ, and MACSE**

	Ser	His	Lys	Gly	Arg	Ser	Asp	Ala	
A:	TCC	CAT	AAG	GGG	CGG	TCG	GAC	GCC	---
D:	TCC	CAG	GGG	CGG	TCC	CAG	GAC	GCC	ACG
	Ser	Gln	Gly	Arg	Ser	Gln	Asp	Ala	Thr

**d) COATi**

	Ser	His	Lys	Gly	Arg	Ser	Asp	Ala	
A:	TCC	CAT	AAG	GGG	CGG	T-- -CG	GAC	GCC	---
D:	TCC	CA-	--G	GGG	CGG	TCC CAG	GAC	GCC	ACG
	Ser		Gln	Gly	Arg	Ser	Gln	Asp	Ala
									Thr

**Figure 1:** Standard algorithms produce suboptimal alignments. (a) shows the true 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  $P(X, Y)$  (Yoon 2009).

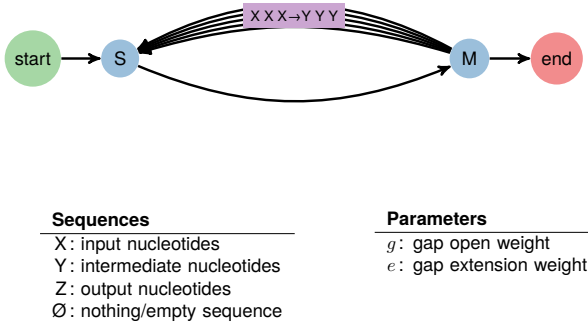
A limitation of pair-HMMs is that they only model the evolution of two related sequences from an unknown ancestor. Finite-state transducers (FSTs) have similar benefits to pair-HMMs with the additional feature that can model the generation of a descendant sequence given an ancestral one. FSTs consume 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  $P(Y|X)$ . 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.

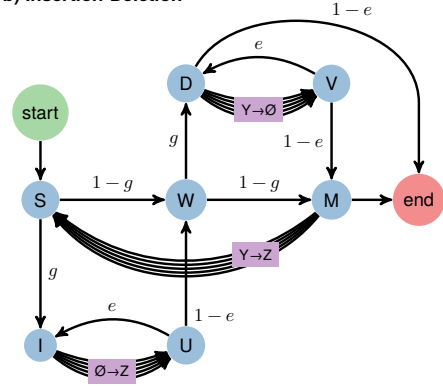
Composing both sequences with the Evolution FST results in the transducer of all possible alignments. Any path through this FST represents a pairwise alignment, while the shortest path (by weight) corresponds to the best alignment. All FST operations in COATi, including model development, composition, 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 support mutations to (early) stop codons under the assumption that these are artifacts common in low-quality data.

a) Substitution



b) Insertion-Deletion



**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  $61 \times 61$  codon substitution model with 3721 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. On top of the indel FST, we add a base-calling-error FST (Supplemental Materials Figure 5) to model sequencing errors. 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  $P_{\text{cod}}(Y_0 \cdot Y_1 \cdot Y_2 | X_0 \cdot X_1 \cdot X_2)$  represent transition probabilities from a standard codon model, and

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

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 (1994) 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 COATi and a suite of popular aligners spanning multiple alignment methods: ClustalΩ v1.2.4 (Sievers et al. 2011), MACSE v2.06 (Ranwez et al. 2011), MAFFT v7.505 (Katoh and Standley 2013), and PRANK v.150803 (Löytynoja 2014).

After downloading, we removed 2232 sequence-pairs longer than 6000 nucleotides and then aligned the remaining pairs with all five methods. At least one aligner added gaps to 6048 sequence pairs, and no aligner added gaps to 7719 sequence pairs. We then randomly introduced gap patterns extracted from all five methods into the ungapped sequence pairs to generate the benchmark alignments. The accuracy of inferred alignments compared to the benchmarks was measured using Blackburne and Whelan’s (2011) distance metric,  $d_{seq}$ . Additionally, because different alignments may be evolutionary equivalent, we scored alignments using the codon-marginal-mg model, and any alignment that had the same score as the benchmark alignment was considered equivalent to the benchmark alignment. The accuracy of identifying positive and negative selection was calculated using the  $F_1$  score by estimating  $k_s$  and  $k_a$  statistics (Li 1993) (Supplementary Methods). The  $F_1$  score evaluates the accuracy of a model by assigning equal importance to precision and recall, and ranges between 0 and 1, with a score of 1 representing a perfect result.

	COATi	MAFFT	PRANK*	MACSE	ClustalΩ
Method	Trip-MG	DNA	Codon	DNA+AA	AA
Avg alignment error ( $d_{seq}$ )	0.00214	0.01392	0.02001	0.01351	0.02691
Best alignments	5152	4833	4748	3754	2595
Perfect alignments	5722	5408	4706	2860	2937
Imperfect alignments	1066	1380	2082	3928	3851
F1 score for positive selection	98.2%	86.1%	88.4%	81.2%	71.0%
F1 score for negative selection	99.8%	98.6%	98.8%	98.3%	97.0%

\*PRANK produced 50 empty alignments, calculations are based on 7869 alignments.

**Table 1:** COATi generates better alignments than other alignment algorithms. Results of COATi, PRANK, MAFFT, ClustalΩ, and MACSE aligning 7719 empirically simulated sequence pairs. Best alignments have the lowest  $d_{seq}$  (including ties), perfect alignments have the same score as the true alignment, and imperfect alignments have a different score than the true alignment when at least one method found a perfect alignment.

COATi, using the codon-triplet-mg model, obtained better results compared to a wide variety of alignment strategies. It was significantly more accurate (lower  $d_{seq}$ ) at inferring the empirically simulated alignments compared to other methods; all p-values were less than  $2.1 \cdot 10^{-79}$  according to the one-tailed, paired Wilcoxon signed-rank tests (Supplementary Materials Figure 1). In addition, COATi produced more perfect alignments, less imperfect alignments, and more accurately inferred events of positive and negative selection (Table 1).

ClustalΩ generated alignments via amino acid translations and obtained the highest average alignment error while having difficulties retrieving positive selection. MACSE used a DNA-AA hybrid model, allowing frameshifts, and obtained similar results to MAFFT using a DNA model.

PRANK, using a codon model, had an average alignment error between MACSE/MAFFT and ClustalΩ but was unable to generate alignments for some sequence pairs.

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. Using COATi will allow researchers to analyze more data with higher accuracy and facilitate the study of important biological processes that shape genomic data.

Future work include extending the indel FST to combine a 3-mer gap model with a frameshift parameter and weighing each indel phase differently to reflect known selection on indel phases (Zhu 2022). We also plan on comparing the marginal and triplet models to evaluate the implications of the marginalization.

## Availability

The source code for COATi, along with documentation, is freely available on GitHub: <https://github.com/CartwrightLab/coati> and is implemented in C++. Additional information, code, and workflows 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. Translatorex: 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 O. 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 and Standley DM. 2013. Mafft multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution* 30:772–780.

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.