Modeling artifacts in sequence alignment

Juan J Garcia Mesa

Reed Cartwright
Banu Ozkan
Jay Taylor
Ted Pavlic

Biodesign Institute
Arizona State University

Ira A. Fulton Schools of Engineering
Arizona State University

Background - Artifacts in Genomic Data

Uncorrected errors in genomic datasets can lead to inaccurate results in functional and comparative genomic studies [8].

- Early stop codons
- Frameshifts
- Within-codon indels

	Lys	Ala	Leu	Leu
H:	AAG	GGC	CTC	TTG
G:	AAG	---	GTC	TTG
	Lys	-	Val	Leu

	Lys	Ala	Leu	Leu
H:	AAG	GGC	CTC	TTG
G:	AAG	G--	-TC	TTG
	Lys	Val	-	Leu

	Pro	Pro	Lys	Leu
H:	CCC	CCC	AAG	CTG
G:	CCC	CCG	---	CTG
	Pro	Pro	-	Leu

	Pro	Pro	Lys	Leu
H:	CCC	CCC	AAG	CTG
G:	CCC	CC-	--G	CTG
	Pro	Pro	-	Leu

Background - Sequence Alignment

- Hypothesis of which characters are related by common descent [2].
- Sequence alignment is a fundamental task that precedes many genomic analyses [7].
- Often seen as an *ad hoc* problem [6].

Background - Shortcomings of Current Aligners

- Often based on AA translations

Background - Shortcomings of Current Aligners

- Often based on AA translations
- Codon models

Background - Shortcomings of Current Aligners

- Often based on AA translations
- Codon models

Trouble processing stop codons

Background - Shortcomings of Current Aligners

- Often based on AA translations
- Codon models

  Trouble processing stop codons
- No aligner combines codon models with frameshifts

Background - Shortcomings of Current Aligners

- Often based on AA translations

- Codon models

  Trouble processing stop codons

- No aligner combines codon models with frameshifts
- Combination of AA model with frameshifts

  Lacks statistical model, slow

Aims - COdon-aware Alignment Transducer

COATi will be a **statistical aligner** implementing **codon substitution models** that allows **gaps at any position** and is **robust to artifacts**.

- Aim 1: statistical pairwise alignment.
- Aim 2: artifacts in genomic datasets.
- Aim 3: estimation of parameters.

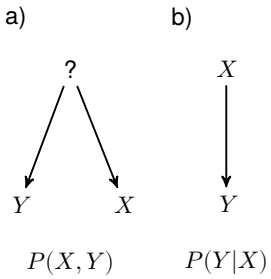
Aim 1 - Finite State Transducers

Pair hidden Markov models

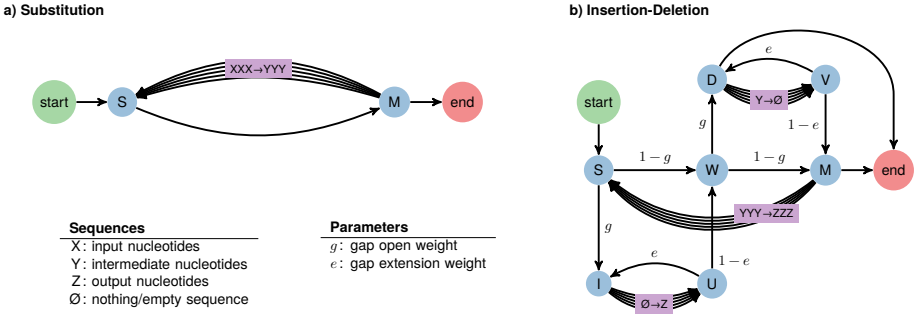
- Primary technique in statistical pairwise alignment.
- Generates two sequences from an unknown common ancestor (a).
- Alignment represents $P(X, Y)$.

FSTs

- An FST generates a descendant sequence given an ancestral one (b).
- Alignment represents $P(Y|X)$.
- Algorithms for combining FSTs (e.g. composition).



Aim 1 - Evolution FST



- Composing (a) and (b) results in the evolution FST.
- Nodes represent states in an FST, arcs display possible transitions.
- Combines a codon substitution model with indels that can occur at any position.

Aim 1 - Substitution Model

Codon substitution with instantaneous substitution rate matrix Q:

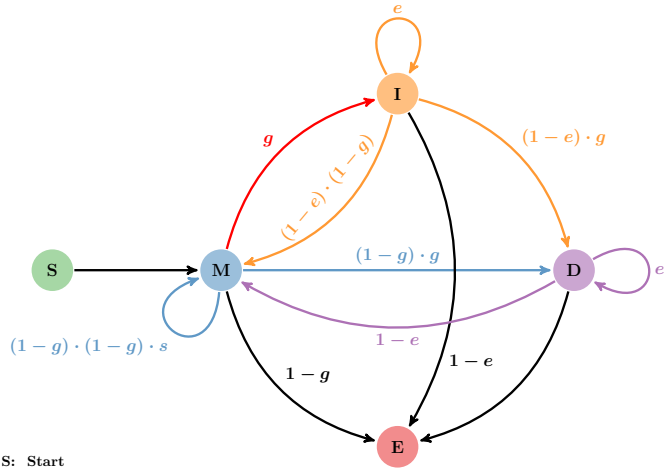
$$Q_{ij} = \begin{cases} \mu_{ij} & \text{if } i \text{ and } j \text{ are synonymous} \\ \omega \cdot \mu_{ij} & \text{if } i \text{ and } j \text{ are nonsynonymous} \end{cases}$$

$$Q_{ii} = -\sum_{j:j \neq i} Q_{ij}$$

- μ_{ij}: mutation rate of codon i to j.
- ω: coefficient of selection.
- Supports a variety of models

Aim 1 - Dynamic Programming

- FST pairwise alignment is performed via **composition** (expensive).
- Solution: reducing the evolution FST to three states and aligning via dynamic programming.



Aim 2 - Artifacts

- Artifacts are common in genomic data sets, especially in non-model organisms.
- Current practices involve discarding data.
- COATi will align a sequence from a non-model organism against a high-quality sequence as a path through an FST.

Aim 2 - Marginal Substitution Model

- Substitution models assume sequences are accurate.
- Not the case for non-reference genomes.
- Marginal substitution model is robust to erroneous nucleotides.

$$P'_{cod_1,nuc,pos} = \sum_{cod_2} \begin{cases} P(cod_2 \,|\, cod_1) & \text{if } cod_2[pos] = nuc \\ 0 & \text{otherwise} \end{cases}$$

Aim 2 - Marginal Substitution Model

- Substitution models assume sequences are accurate.
- Not the case for non-reference genomes.
- Marginal substitution model is robust to erroneous nucleotides.

$$P(nuc = A, pos = 1|ACT) = \sum_{cod} \begin{cases} P(cod \,|\, ACT) \text{ if } cod[1] = A \\ 0 \text{ otherwise} \end{cases}$$

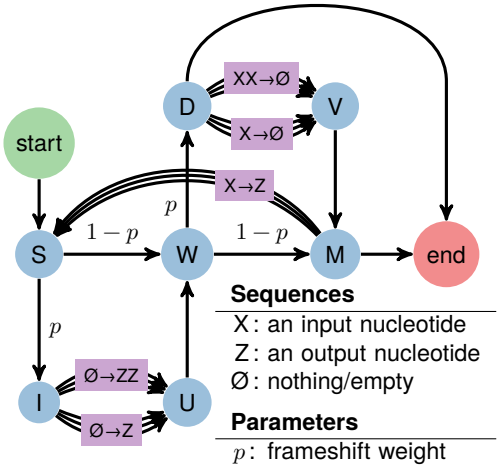
Aim 2 - Ambiguous Data

- Ambiguous nucleotides are common in low-quality data.
- Marginal model will handle all 15 cases for descendant sequence.
- Typically replaced by average over all possibilities.
- Alternative approaches to handle ambiguous data.

IUPAC nucleotide code	Base
A	Adenine
C	Cytosine
G	Guanine
T (or U)	Thymine (or Uracil)
R	A or G
Y	C or T
S	G or C
W	A or T
K	G or T
M	A or C
B	C or G or T
D	A or G or T
H	A or C or T
V	A or C or G
N	any base
. or -	gap

Aim 2 - Model Frameshifts

- FST that specifically models frameshifts (lengths 1-2).
- Longer frameshifts are modeled by setting the indel FST to length multiple of 3 and composing it with the frameshift FST.



Aim 2 - Biological Frameshifts

- Frameshifts in coding sequences are expected to be artifacts due to purifying selection.
- In some cases frameshifts are believed to be biological.
- To my knowledge, this particular issue has not been addressed.

Aim 3 - Expectation-Maximization

- Ability to infer parameters estimates from data.
- EM [3] is a classic iterative method for deriving estimates of parameters in statistical models with latent variables.
- E-step: infer information about latent variables.
- M-step: improve parameter estimates.

Aim 3 - Model Parameters

Substitution model

- GTR[10] underlying model for MG94: 6 nucleotide transition $\boldsymbol{\sigma}$.
- 4 nucleotide or 64 codon frequencies: $\boldsymbol{\pi}$.
- Coefficient of selection $\boldsymbol{\omega}$.

Indel model

Standard model

- Gap opening: $\boldsymbol{g}$.
- Gap extension: $\boldsymbol{e}$.

Extended model

- Insertion opening: $\boldsymbol{i_o}$.
- Insertion extension: $\boldsymbol{i_e}$.
- Deletion opening: $\boldsymbol{d_o}$.
- Deletion extension: $\boldsymbol{d_e}$.

Preliminary Data

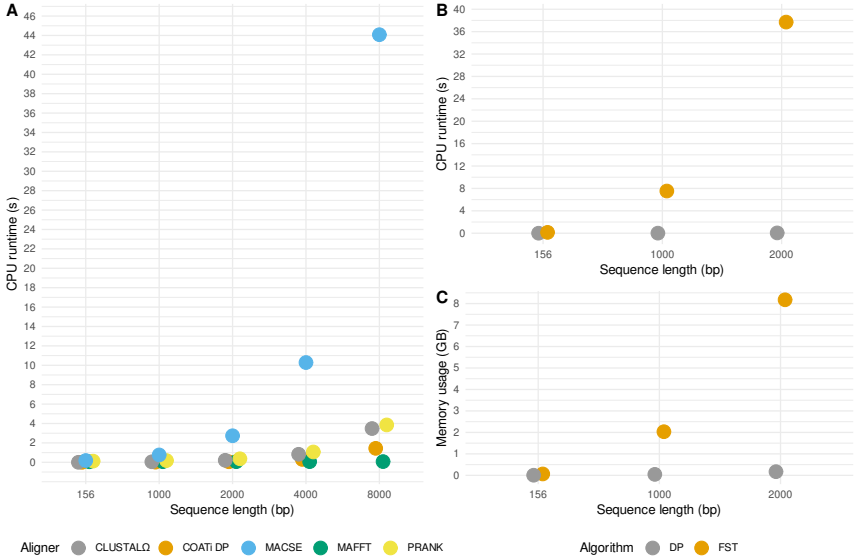
- Downloaded 4000 human-gorilla homologous pairs from ENSEMBL [4].
- Aligned and extracted gap patterns from 1660 alignments.
- Introduced gaps into remaining 2340 alignments to simulate 'true' data set.
- Remove gaps and compare accuracy of aligners retrieving 'true' data.
- Metrics
  - Distance metric
  - Number of perfect alignments
  - Accuracy of selection

Preliminary Data

	COATi	PRANK	MAFFT	CLUSTALΩ	MACSE
Avg alignment error (d_{seq})	0.00060	0.01086	0.00671	0.01300	0.00611
Perfect alignments	1300	86	1282	634	1059
Best alignments	1756	188	1463	666	1129
Imperfect alignments	437	1651	455	1109	678
Accuracy of positive selection	97.3%	87.3%	85.8%	69.1%	81.5%
Accuracy of negative selection	99.8%	98.9%	98.7%	97.3%	98.5%

- COATi performs best on all metrics.
- AA-based aligners (CLUSTALΩ, MACSE) have difficulties retrieving positive selection.
- PRANK (no frameshifts) together with CLUSTALΩ have the highest alignment error.
- MAFFT (DNA model) and MACSE (AA-based and frameshifts) have lower alignment error but also have difficulties with positive selection.

Preliminary Data - Memory and Runtime



Future work

Extend COATi pairwise to multiple sequence alignment.

- Initial alignment without an input phylogenetic tree.
- Iterative refinement by sampling alignment space.

Questions

References I

1. 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).

2. Cartwright, R. A. Problems and solutions for estimating indel rates and length distributions. *Molecular biology and evolution* **26,** 473–480 (2009).

3. Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* **39,** 1–38. ISSN: 00359246. %7Bhttp://www.jstor.org/stable/2984875%7D (2022) (1977).

4. Hubbard, T. *et al.* The Ensembl genome database project. *Nucleic acids research* **30,** 38–41 (2002).

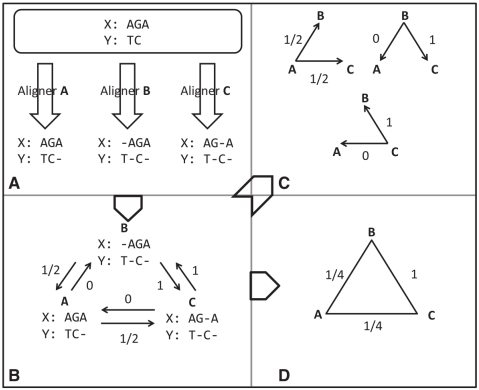
References II

5. Löytynoja, A. in *Multiple sequence alignment methods* 155–170 (Springer, 2014).
6. Morrison, D. A. *Multiple Sequence Alignment is not a Solved Problem.* 2018. arXiv: 1808.07717 [q-bio.PE].
7. Rosenberg, M. S. *Sequence alignment: methods, models, concepts, and strategies.* (Univ of California Press, 2009).
8. 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).
9. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular systems biology* **7,** 539 (2011).
10. Tavaré, S. *et al.* Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on mathematics in the life sciences* **17,** 57–86 (1986).

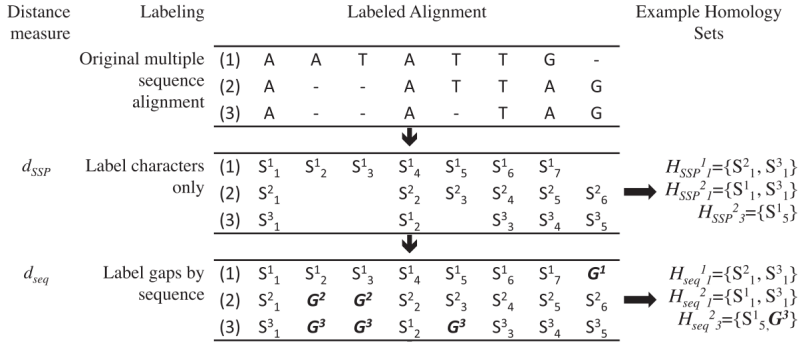
Distance Metric

Non-negative function, $d(x, y)$, metric conditions:

- $d(x, y) = 0$ iff $x = y$
- $d(x, y) = d(y, x)$ symmetry
- $d(x, z) \leq d(x, y) + d(y, z)$ triangle inequality



Distance Metric



$$d(A,B) = \frac{1}{c}\sum_i \sum_j d(A,B)^i_j = \frac{1}{c}\sum_i \sum_j \frac{|H(A)^i_j \triangle H(B)^i_j|}{|H(A)^i_j| + |H(B)^i_j|}$$

A	A	T	A	T	T	G	-	A	A	T	A	T	T	-	G
A	-	-	A	T	T	A	G	A	A	T	-	-	T	A	G
A	-	-	A	-	T	A	G	A	A	-	-	-	T	A	G

$d_{seq}(A,B)_1^1 = 0$
$d_{seq}(A,B)_1^2 = 0$
$d_{seq}(A,B)_1^3 = 0$

$d_{seq}(A,B)_2^1 = \frac{2}{4} = \frac{1}{2}$
$d_{seq}(A,B)_2^2 = \frac{1}{4}$
$d_{seq}(A,B)_2^3 = \frac{1}{4}$

Personal Background

- B.S. Computer Science, University of Barcelona, 2012-2014 (transfer)
- B.A. Interdisciplinary Studies (Sustainability & Computational Mathematical Sciences), ASU, 2017
- Assistant Software Engineer, 2017-2018
- Biological Design PhD, ASU, 2018-

Accuracy of Selection

d_N and d_S ($d_N/d_S = \omega$) are used to estimate the selection a given protein or DNA section is experiencing.

- d_N: number of non-synonymous changes over non-synonymous sites.
- d_S: number of synonymous changes over synonymous sites.
- $\omega \approx 1$: neutral selection.
- $\omega > 1$: positive selection.
- $\omega < 1$: purifying selection.

Accuracy of Selection

F_1 score to test correct inference of selection.

$$
\begin{aligned}
F_1 &= \left(\frac{2}{recall^{-1} + precision^{-1}} \right) \\
&= 2 \cdot \frac{precision \cdot recall}{precision + recall} \\
&= \frac{2 \cdot TP}{2 \cdot TP + FP + FN}
\end{aligned}
$$

Where precision$= \frac{TP}{TP+FP}$ and recall$= \frac{TP}{TP+FN}$.