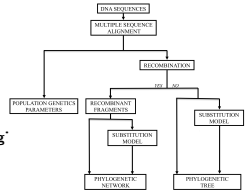
Artifacts in Sequence Alignment

Sequence Alignment

Sequence Alignment

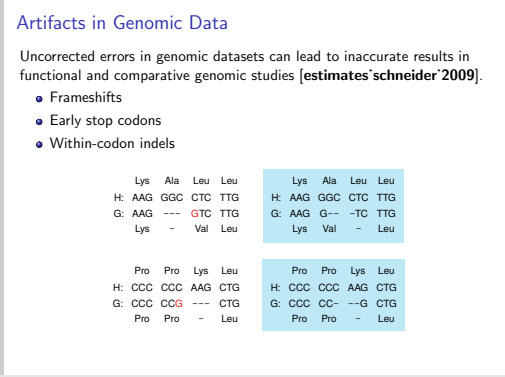
- Hypothesis of which characters are related by common descent [**problems˙cartwright˙2009**].
- Sequence alignment is a fundamental task that precedes many genomic analyses [**sequence˙alignment˙rosenberg˙**
- Often seen as an *ad hoc* problem [**morrison˙MSA˙2018**].



... such as phylogenetic inference, measurement of selection, gene annotation (among others).

Artifacts in Sequence Alignment

Artifacts in Genomic Data



1. Current coding sequence aligners don't model correctly within codon gaps and frameshifts because generally alns are done based on amino acid translations.
2. In data set used prelim results (4000 homologous human-gorilla pairs
3. Avg across aligners: 1.07 frameshifts/locus, 0.28 stop codons/locus
4. (in reality, 2340 were gapless, 2.59 frameshifts per locus)

Artifacts in Sequence Alignment

Shortcomings of Current Aligners

Shortcomings of Current Aligners

- Often based on AA translations

1. AA translations, which loses information
2. 61x61 substitution models, removing stop codons.
3. PRANK replace stop cods with 'NNN' & BAli-Phy fails completely
4. COATi & MACSE an order of magnitude less stop codons

Artifacts in Sequence Alignment

Shortcomings of Current Aligners

Shortcomings of Current Aligners

- Often based on AA translations
- Codon models

1. AA translations, which loses information
2. 61x61 substitution models, removing stop codons.
3. PRANK replace stop cods with 'NNN' & BAli-Phy fails completely
4. COATi & MACSE an order of magnitude less stop codons

Artifacts in Sequence Alignment

Shortcomings of Current Aligners

Shortcomings of Current Aligners

- Often based on AA translations
- Codon models
  - Trouble processing stop codons

1. AA translations, which loses information
2. 61x61 substitution models, removing stop codons.
3. PRANK replace stop cods with 'NNN' & BAli-Phy fails completely
4. COATi & MACSE an order of magnitude less stop codons

Artifacts in Sequence Alignment

Shortcomings of Current Aligners

Shortcomings of Current Aligners

- Often based on AA translations
- Codon models
  - Trouble processing stop codons
- No aligner combines codon models with frameshifts

1. AA translations, which loses information
2. 61x61 substitution models, removing stop codons.
3. PRANK replace stop cods with 'NNN' & BAli-Phy fails completely
4. COATi & MACSE an order of magnitude less stop codons

Artifacts in Sequence Alignment

Shortcomings of Current Aligners

Shortcomings of Current Aligners

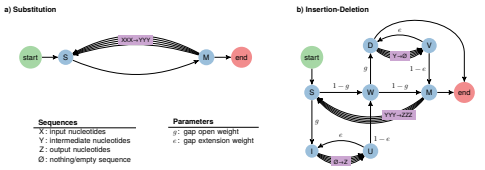
- Often based on AA translations
- Codon models
  - Trouble processing stop codons
- No aligner combines codon models with frameshifts
- Combination of AA model with frameshifts
  - Lacks statistical model, slow

1. AA translations, which loses information
2. 61x61 substitution models, removing stop codons.
3. PRANK replace stop cods with 'NNN' & BAli-Phy fails completely
4. COATi & MACSE an order of magnitude less stop codons

Artifacts in Sequence Alignment

Aim 1 - Evolution FST

Aim 1 - Evolution FST



- Composing (a) and (b) results in the evolution FST.
- Each state node represents a state in an FST.
- Arcs display possible transitions and their weights.

1. (a) encodes a 64x64 codon substitution matrix
2. (b) models insertions and deletions

Artifacts in Sequence Alignment

Aim 1 - Substitution Model

1. Muse & Gaut 1994, describe
2. Empirical Codon Model, describe

Aim 1 - Substitution Model

Codon substitution with instantaneous substitution rate matrix Q:

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

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

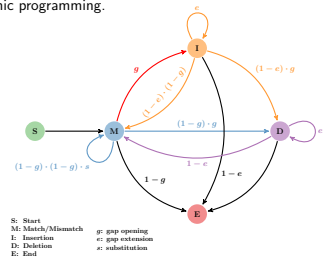
- μ_{ij}: mutation rate of codon i to j.
- ω: coefficient of selection.
- Supports a variety of models (e.g. MG94[**muse˙gaut˙1994**], ECM[**kosiol˙ECM˙2007**]).

Artifacts in Sequence Alignment

Aim 1 - Dynamic Programming

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.



1. which is a powerful operation that allows complex FSTs to be build from smaller parts. However, composition can be prohibitive when aligning sequence of a few thousand nucleotides and up. To solve this issue, the FST can be reduced to three states and solved via dynamic programming.

Artifacts in Sequence Alignment

Aim 2 - Marginal Substitution Model

1. From 64x64 (MG94/ECM) to 64x4x3

Aim 2 - Marginal Substitution Model

- Substitution models assume sequences are accurate.
- Not the case for non-reference genomes.
- Solution: pairwise align a high-quality against a low-quality sequence.
- Marginal substitution model is robust to erroneous nucleotides.

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

Artifacts in Sequence Alignment

Aim 2 - Ambiguous Data

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 nucleoide 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

1. Best nucleotide. Maybe as a function of t (weighted avg; smaller t =¿ more weight to best, larger t =¿ more avg)?

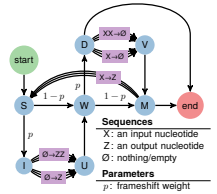
Artifacts in Sequence Alignment

Aim 2 - Model Frameshifts

1. Compare approaches

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 with the frameshift FST.



Artifacts in Sequence Alignment

Aim 2 - Biological Frameshifts

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.

1. Cases: pairs of compensatory indels, frameshifts in *Saccharomyces cerevisiae* (collaborators; "un-studied").
2. Not addressed the problem yet; ideas: convert to DNA after frameshift, "correct" reading frame after frameshift.

Artifacts in Sequence Alignment

Preliminary Data

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.
- Aligners that don't allow frameshifts (PRANK, CLUSTALΩ) have the highest alignment error.
- Tools that model frameshifts (MAFFT, MACSE) lower aln error but also have difficulties with positive selection.

1. 1st row: distance metric, the lower the better.
2. ClustalΩ has difficulty "recovering/maintaining" positive selection given its AA-based alignment.

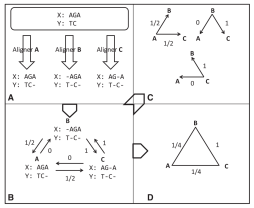
Artifacts in Sequence Alignment

Distance Metric

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

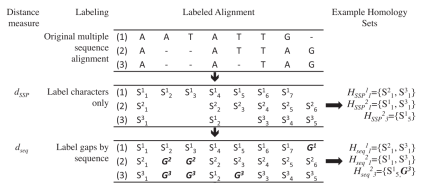


1. For a function to be a valid metric has to meet all three criteria (explain briefly). Sum of pairs, a common score used to evaluate differences between MSA based on the number of matches on each column, has been proven to not satisfy all 3 conditions. An example (fig).

Artifacts in Sequence Alignment

Distance Metric

Distance Metric



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

1. Homology set: characters on the same column, i.e. nucleotides said to be homologous.
2. Hamming distance: number of different positions or minimum number of substitutions required to change one set into the other.

Artifacts in Sequence Alignment

Accuracy of 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= $\frac{TP}{TP+FN}$.

1. F_1: weighted average of precision and recall. More informative score /statistic than accuracy.
2. Precision: ratio of correctly predicted positive observations to total predicted pos obs. Recall: (sensitivity) ratio of correctly predicted pos obs to all obs true positives.
3. sensitivity=TP/(TP+FN); specificity=TN/(TN+FP); precision=TP/(TP+FP); accuracy=(TP+TN)/(TP+FN+TN+FP)