MSA Impact on Downstream Analysis

Minhyuk Park

Discussants: Payam, Maya, Yufeng, and Wanxian

Paper

Class of Multiple Sequence Alignment Algorithm Affects Genomic Analysis

by Benjamin P. Blackburne and Simon Whelan

Published in Molecular Biology and Evolution, Volume 30, Issue 3, March 2013, Pages 642-653

Outline

I. Motivation

II. Methods

    a. Dataset

    b. The Two Classes

III. Key Findings

    a. Phylogenetic Tree Estimation

    b. Detection of Adaptive Evolution

IV. Summary and Discussion

Motivation

MSA is used for

- Phylogeny Estimation

- Interspecies Gene Function Comparison

- Protein Structure Estimation

- Protein Classification/Annotation

Most are based on finding homologous pairs.

Motivation - Impact of MSA

MSA serves an input to various computational biology methods

- Accuracy of MSA may play a role on the accuracy of these methods

- Different MSA methods can produce very different alignments

Methods

To Compare Different MSAms, strict data sanitization is used

- Short, non-fragmented sequences

- contains human sequences

- Few Gaps

- No Excessively long branches on the alignment tree

- Random sampling from the dataset

Methods - Two Classes of MSA

Similarity Based MSAms

- ClustalW

- Muscle

- T-Coffee

Evolution Based MSAms

- Prank

- BAli-Phy

Methods - Similarity Based MSAms

- Progressive Approach such as ClustalW, Muscle

- Consistency Approach such as ProbAlign, T-Coffee, MAFFT

- Not aware of biological evolution data

Methods - Evolution Based MSAms

- Inferring alignment based on an explicit indel model

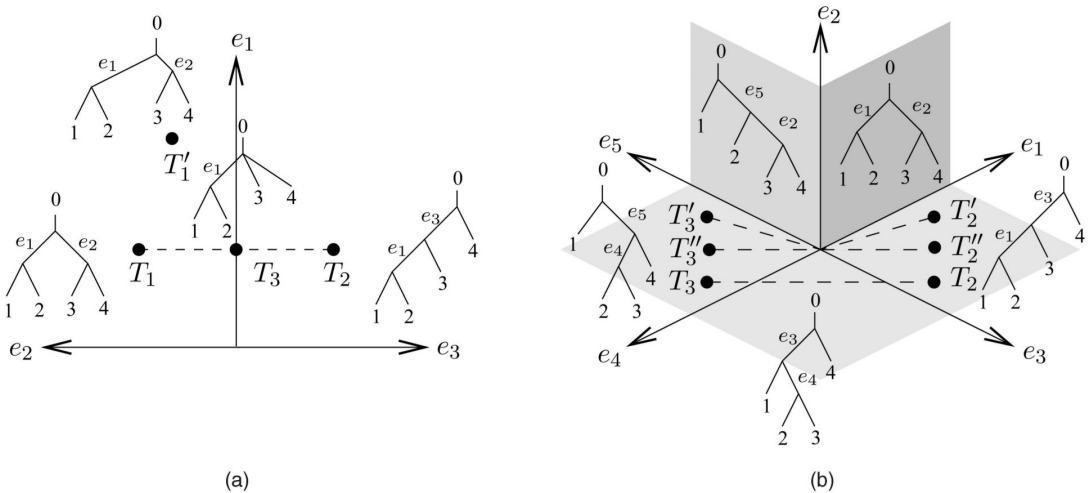
- Prank, BAli-Phy (both align at the codon level)

- Attempts to produce evolutionarily correct trees

Key Findings - Phylogenetic Tree Estimation

- Uses RAxML to estimate the tree
- Measured by Geodesic distances between the estimated trees
    - Shows the differences in tree topology and branch lengths
- PCA projection on the first two axis
- 1st axis describes 81.9% of the variation, separating evolution vs similarity

Geodesic Distance?



Geodesic Distance: Shortest path through the tree space

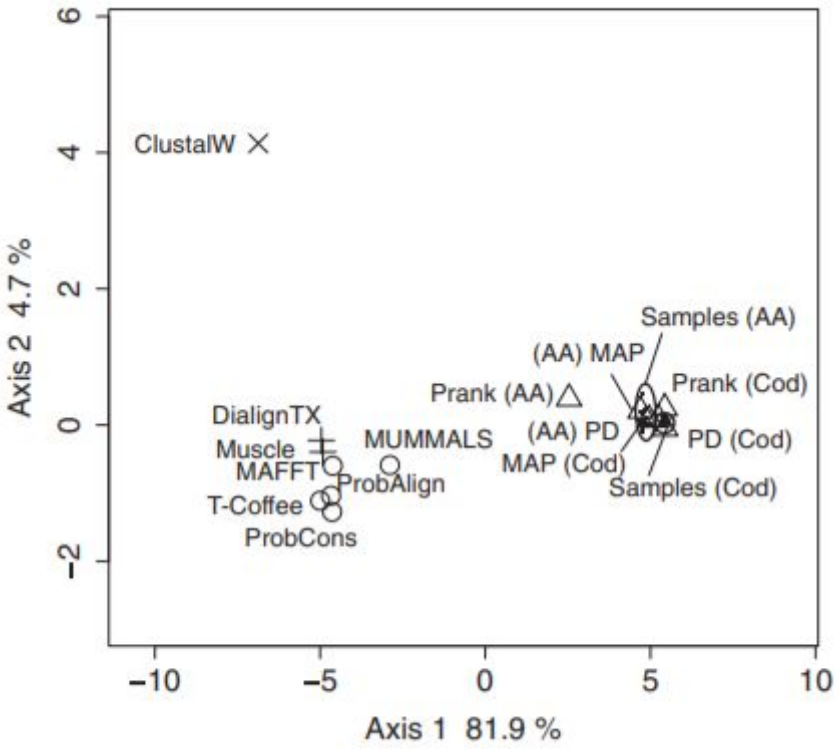
- divided into regions called orthants

- each region is a unique topology

- coordinates defined by the edge length for respectives axes

- orthants adjacent if it can be reached with one edge swap

Figure 2 from the Paper



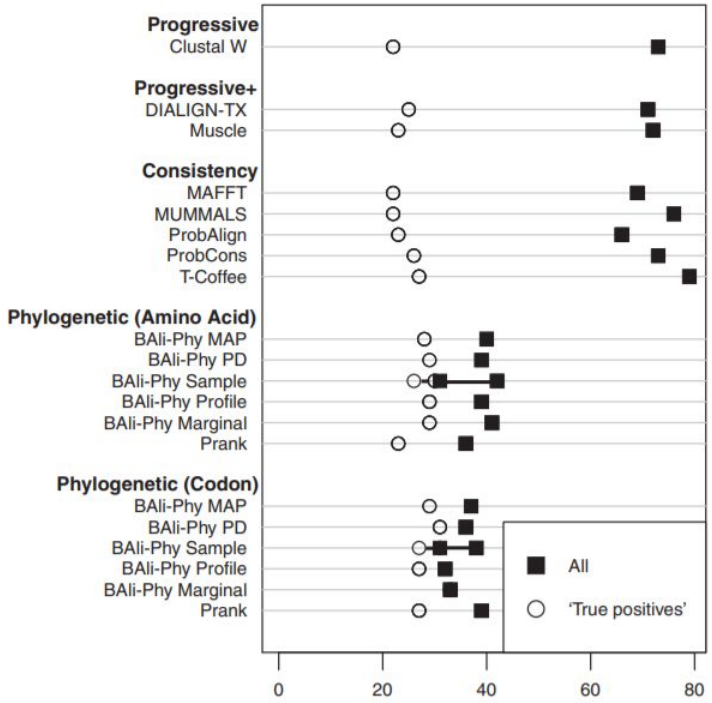
This shows the mean Geodesic distances for each inferred tree from RAxML, projected using PCA. Axis 1 accounts for 81.9% of the tree estimates. We can see that it separates the similarity based MSAms and the evolution based MSAms.

Key Findings - Detection of Adaptive Evolution

- Estimating the relative ratio of nonsynonymous vs synonymous substitutions in the codons

- Similarity MSAms infer more families to be under adaptation than evolution based MSAms do.

Figure 3 from the Paper



This shows on the y axis the total number of families out of 200 that were inferred to be under adaptive evolution. We can see here that all the evolution based alignment methods infer fewer numbers of families to be under adaptive evolution than the similarity based methods.

Question

Why do the classes Matter?

- maybe because evolution MSAms use dynamic scoring matrices and gap penalties based on the length of the branch in the guide tree, similarity MSAms use fixed scoring matrices and gap penalties

- evolution MSAms try to enforce an evolutionary meaningful history for gaps which tend to produce less dense MSAs compared to similarity MSAms

In Conclusion

The class of MSA (similarity vs evolution) affects downstream analysis in demonstrable ways.

- tree topology

- branch lengths

The paper does not dive into the details of why or what the impact of these different classes are or which methods are more biologically reliable, only that the classes impact downstream analysis.

Course Project

- Similar methodology of using different alignment methods piped to RAxML

- measuring the accuracy of GTR matrix obtained from RAxML

- Will be interesting to note how the different classes impact the GTR matrix output from RAxML

Notes

- Low correlation between mean alignment distance and RF distance (0.3)

    - alignment differences do not correlate to different tree topologies

- Study focuses on closely related sequences so on more divergent datasets the subgroups within each classes such as progressive vs consistent MSAms may have differences that are more pronounced

- Authors recommend using both classes of MSAms and carefully analyze the differences between the two outputs

FAQ

1. Why is Prank clustered into one corner?

A: Possible because Prank uses an overly simplistic model of indels that is better suited for simulated data

2. What is the figure saying when it projects trees onto a 2d euclidean space?

A: Distance is simply the number of edge transforms needed through the orthants