Measuring distances between multiple sequence alignments

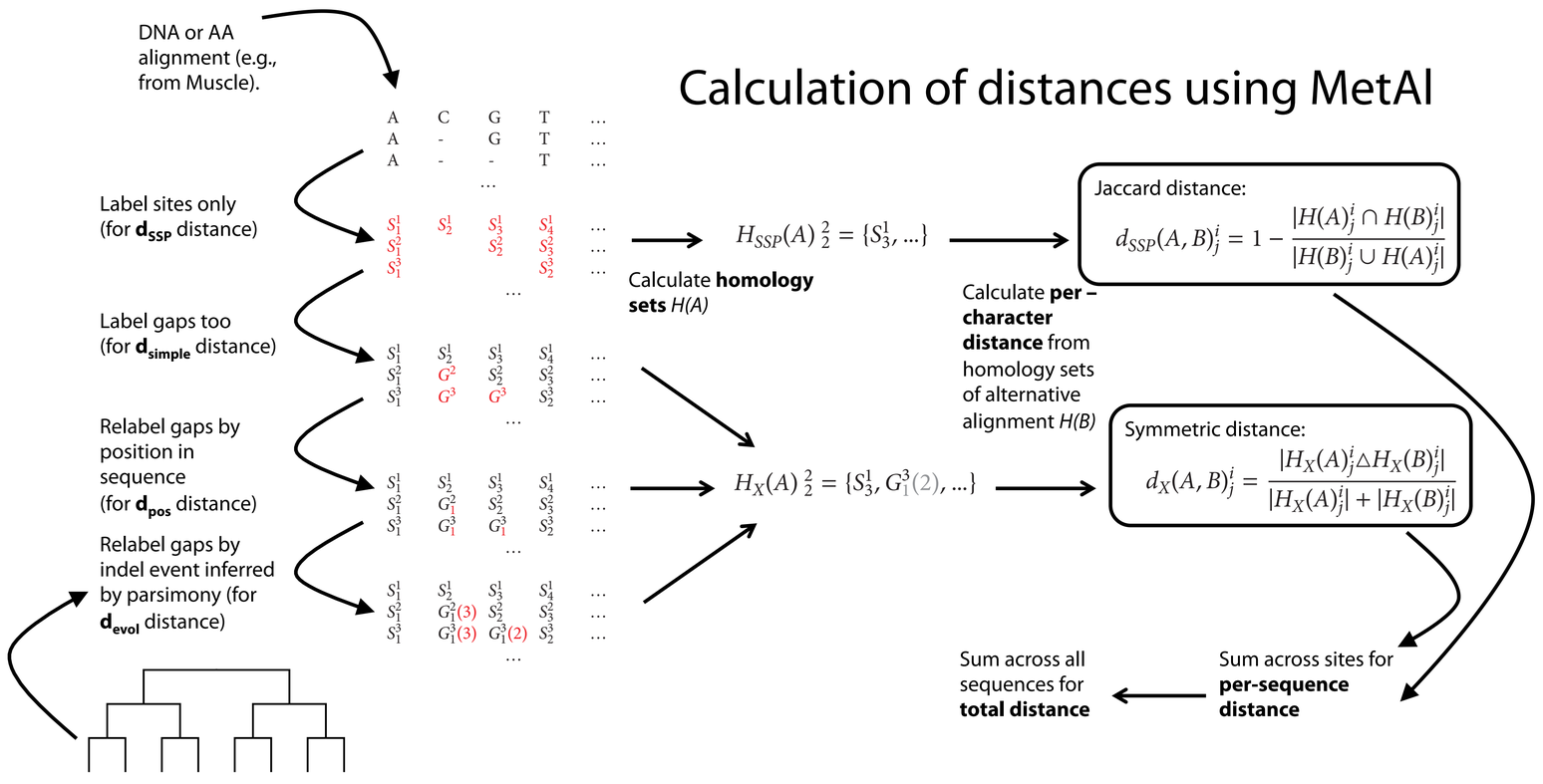
Ben Blackburne and Simon Whelan

Faculty of Life Sciences, University of Manchester

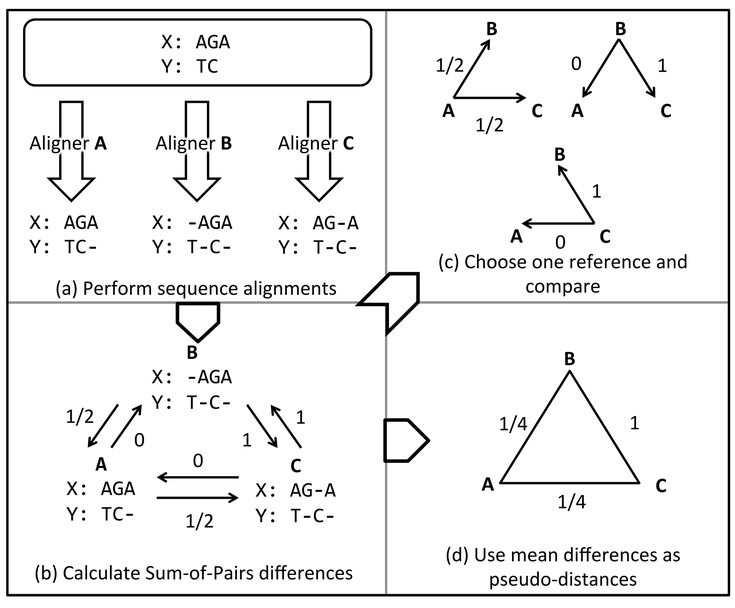
Multiple sequence alignment (MSA) is a core method in bioinformatics. Studies in molecular evolution frequently start with one or more MSAs and their accuracy may affect the scientific conclusions of a study. Many MSA methods have been developed, each using its own objective function and heuristic to produce a MSA. Different methods of inferring MSAs produce different results in all but the most trivial cases, but there are no reliable methods available to compare these results. We present four formal metric functions for measuring the distance between alignments, implemented in the program "MetAl".

How do the metrics work?

We have devised and implemented **four metrics** in the tool **MetAl**. Metric d_{SSP} is a correction to correctly symmetrise the sum-of-pairs similarity score into a proper metric. It treats character-character matches explicitly, and character-gap matches implicitly. Metrics d_{simple}, d_{pos} and d_{evol} treat both character-character and character-gap matches explicitly, and differ by treating gaps as identical states (d_{simple}), differing by position in the sequence (d_{pos}), or differing by the inferred indel event on a phylogeny (d_{evol}). Each distance can be calculated per-character, per-sequence or summed over the whole alignment pair:



What's wrong with current scores?



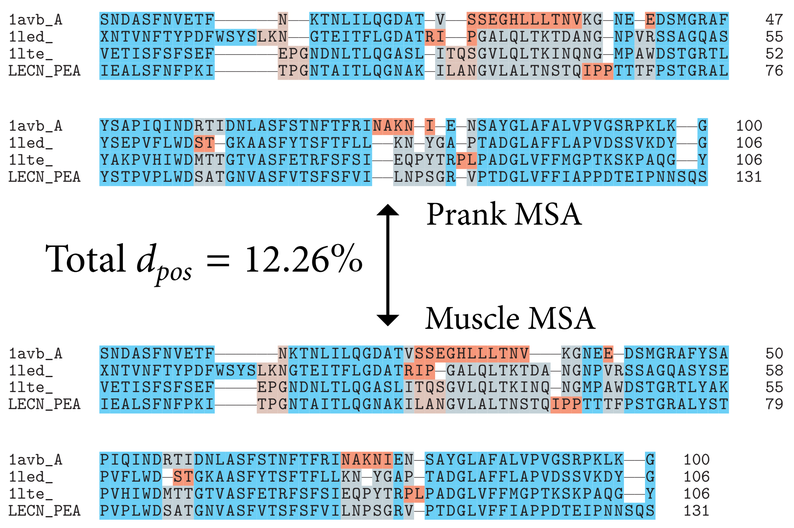
Two toy sequences are aligned three different ways by three different methods. How could we compare them using the sum-of-pairs **(SP)** count of True Postives?

- With one alignment as a reference our results strongly depend on the reference chosen (box (c), above).
- Averaging two SP scores into a pseudo-distance produces an impossible triangle (box (d)).

Solution: A distance that is non-negative, zero iff alignments are identical, symmetric, and obeys the triangle inequality.

Example: Measuring the distance between two alternative MSAs

Using MetAl, we can compare two alignments produced by Prank and Muscle. We choose to use the d_{pos} metric, and annotate segments of the alignments, below, with per-character distances by colour (close to distant):

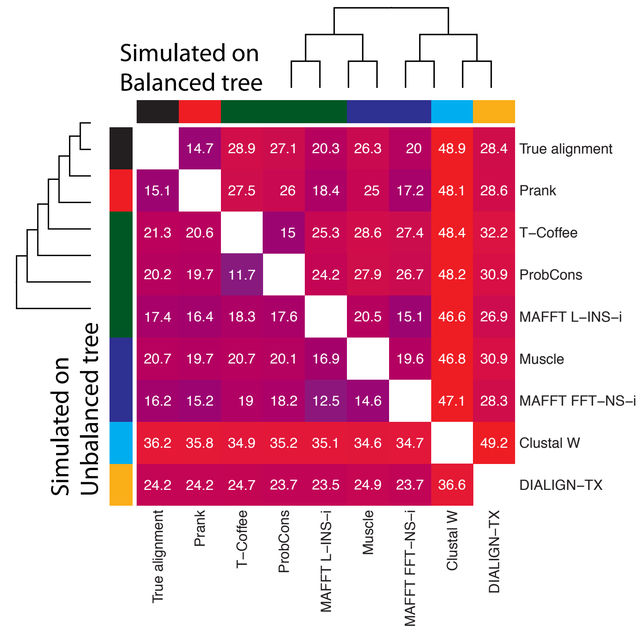


What do the distances mean?

We normalise and express total distances as a percentage. A d_{simple}, d_{pos}, or d_{evol} of x% means **each non-gap character differs by** x% **of its homology pairings** (on average) between the two alternative MSAs. For d_{SSP} the distance is the percentage of pairings between observable characters that do not occur in both alignments.

Application: Comparison of MSA Programs

Using MetAl distances we can compare results between different MSA algorithms across large numbers of alignments. The **heatmap, below**, gives the mean percentage distance (d_{evol}) between the **true MSA** and MSAs produced by a variety of phylogenetic, consistency, progressive (with extra refinements), and greedy-matching methods. The data are **simulated using Indelible** on balanced (upper-right) and unbalanced (lower-left) trees of total length 1.3.
Note that distances between different MSA algorithms are of a similar magnitude to distances from the "truth".



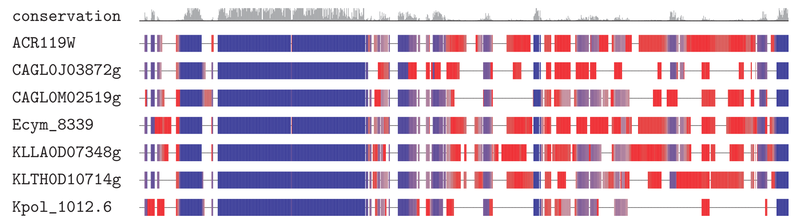
Four metrics – What's the Difference?

- d_{SSP} Symmetrised Sum-of-Pairs. Ignores gaps.
- d_{simple} Includes gaps without information on position.
- d_{pos} Counts two gap locations as identical if they are in the same sequence position.
- d_{evol} Counts two gaps are identical if they are in the same position and correspond to the same inferred indel event.

Note that $d_{evol} \geq d_{pos} \geq d_{simple}$

Application: Alignment Distance Fingerprinting

MetAl can produce JSON-formatted output annotating the per-character distance between two alignments. This output can be trivially reformatted as input for visualisation tools, for instance, TEXshade:



The fingerprint graphic above shows how two alignments of a yeast gene family differ between Muscle and ProbCons. The Muscle alignment is annontated with the distance between the two, from close to distant.

Availability

MetAl is available from:
`http://kumiho.smith.man.ac.uk/whelan/software/metal/`
Download this poster at:
`http://logspace.co.uk/smbe2012.pdf`
We thank the BBSRC for support.

References

[1] B. P. Blackburne, S. Whelan, Bioinformatics **28**, 495 (2012) doi:10.1093/bioinformatics/btr701