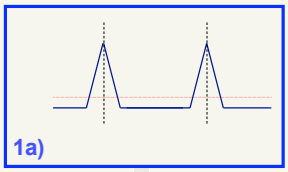
A Simple Algorithm To Automatically Detect Recombinants in DNA/RNA Sequence Alignments

Iain Milne & Frank Wright

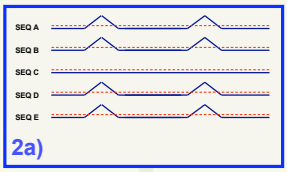
Biomathematics & Statistics Scotland
Scottish Crop Research Institute
Invergowrie
Dundee, Scotland

Introduction - Phylogenetic analysis can be complicated by the presence of recombinant sequences within multiple sequence alignments. This is because the phylogenetic tree model assumes that sequences are diverging independently of each other. However, each recombinant sequence is a mosaic of sub-sequences that have diverged independently. One solution is therefore to identify the recombinant sequences, derive non-recombinant sub-sequences from them, and then estimate a phylogenetic tree showing the relationships among the original non-recombinant sequences and the derived sub-sequences. We have developed a four-step automatic algorithm that produces output that is easy to interpret by biologists, and have implemented it in our new **TOPALi** Java program (Milne et al, 2004).

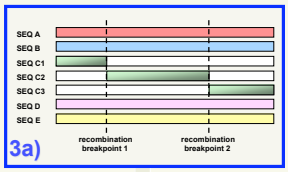
After an essential initial pruning step to remove duplicate or near-duplicate sequences from the alignment, the automatic algorithm runs as follows:



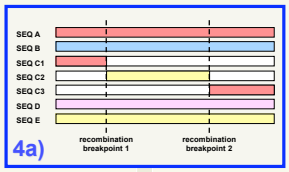
Estimate positions of recombination breakpoints using existing methods and test statistical significance



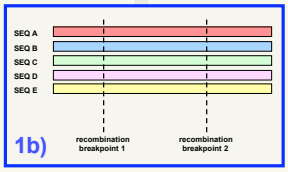
Infer the recombinants using our *leave-one-out-at-a-time* method. Note the loss of signal when C is removed.



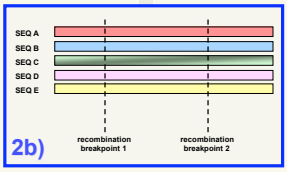
Split recombinant into sub-sequences, substituting these for the original sequence, adding gaps as required



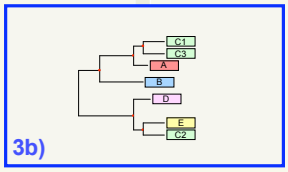
Assign each of C's sub-sequences to nearest non-recombinant sequence based on pairwise similarity



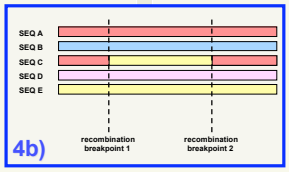
Map positions of significant recombinant breakpoints onto the multiple alignment



Identify the estimated recombinant sequence (Sequence C) in the multiple alignment



Estimate a Bayesian *reconstituted* phylogenetic tree using non-recombinants and C's sub-sequences



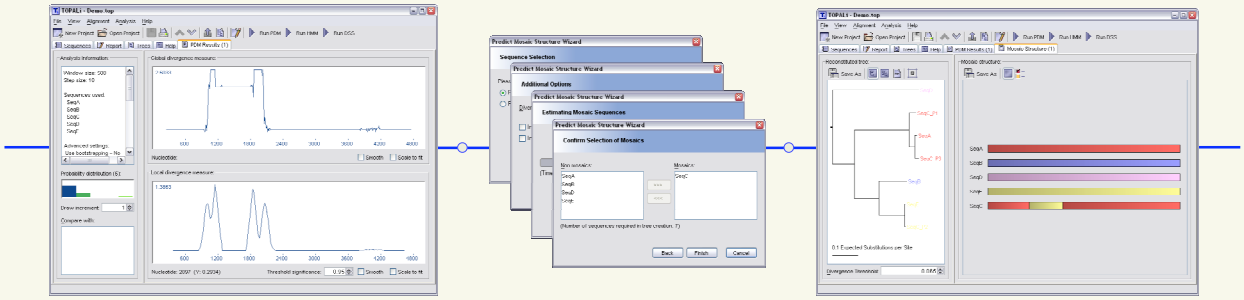
Produce schematic alignment, showing that Sequence C is a mosaic of A-types and E-types

Step 1 - Recombination Breakpoint Estimation. Most research has concentrated on this step, combined with a subsequent manual comparison of phylogenetic trees for the regions each side of the breakpoint. Over fifteen methods have been produced, including three by our group (DSS, HMM, PDM; available in the **TOPALi** program).

Step 2 - Recombinant Sequence Inference. Very little work has been done on this. Weiller (1998) applied a leave-one-out-at-a-time approach to graphically explore alignments. We have also used this approach and have developed a statistical test based on parametric bootstrapping. The significance threshold is produced by simulating alignments using an F84+Gamma model. Recombinant sequences are inferred if their removal does not result in a significant reduction in the observed recombinant signal.

Step 3 – Reconstituted Tree Estimation. To reconstruct the evolutionary history of the non-recombinant sequences and derived sub-sequences, we used a character-based phylogenetic approach rather than a distance-based one. **TOPALi** uses the Bayesian approach for this step, which performs very well but is computationally intensive for large numbers of sequences.

Step 4 – Produce Schematic Alignment. This final step was carried out by generating a distance matrix from the reconstituted tree and assigning colour to the derived sub-sequences according to their pairwise similarity to full-length non-recombinant sequences.



TOPALi running the steps from above on a simulated dataset of length 5000bp. The two recombination breakpoints can be seen in the lower left-hand graph (detected here via TOPALi's Probabilistic Divergence Method). Once the recombinant sequence has been identified (middle image), the reconstituted tree and final schematic alignment are displayed together. Notice how TOPALi moves the recombinant sequence to the bottom of the diagram.

Future Work - Identifying recombinant sequences is a difficult task and this is a first attempt at automating the entire process. The algorithm works well with alignments containing small numbers of recombinants. We are currently investigating its behaviour with a range of simulated and real datasets.

References

Milne I, Wright F, Rowe G, Marshall DF, Husmeier D & McGuire G (2004) **TOPALi: software for automatic identification of recombinant sequences within DNA multiple alignments**, *Bioinformatics* 20(11), pp 1806-7

Weiller GF (1998) **Phylogenetic Profiles: a graphical method for detecting genetic recombinations in homologous sequences**, *Molecular Biology and Evolution* 15, pp 326-335

This work was funded by the BBSRC/EPSRC Bioinformatics Initiative (Ref: BIO10494) and the Scottish Executive (SEERAD)

Biomathematics & Statistics Scotland

http://www.bioss.ac.uk/software.html