

Temporal Multiple Statistical Alignment

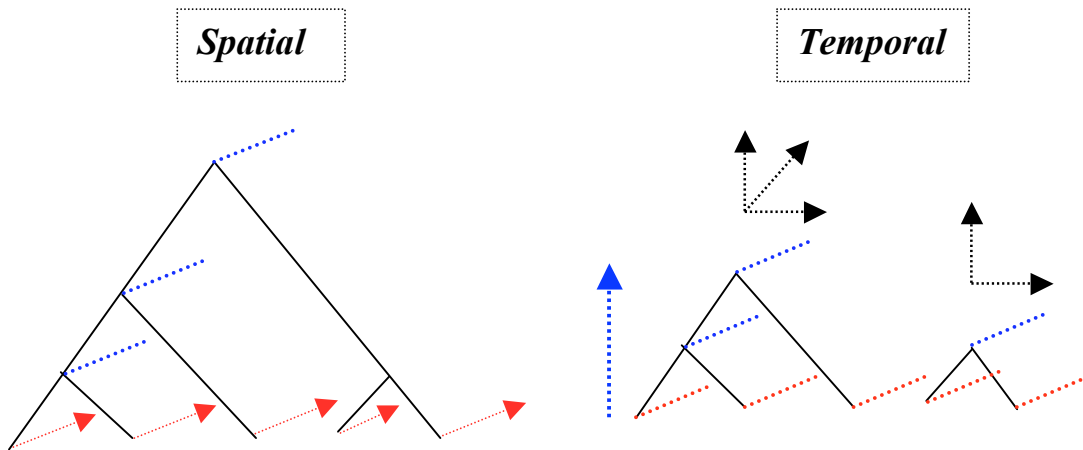
1.6.07

A series of algorithms evolutionary biology both involves time (*temporal*) and sequence (*spatial*) and will have different formulations. Four examples already described in the literature are i. Likelihood calculations on a pedigree, where Elston and Stewart (1971) is temporal and Lander and Green (1987) is spatial. ii. Simulations under the coalescent with recombination, where Hudson (1983) is temporal and Wiuf and Hein (1999) is spatial. iii. Finding minimal recombination histories, where Song and Hein (2005) is spatial and Lyngsø, Song and Hein (2006) is temporal. iv. Score based multiple alignment was originally formulated spatially (Sankoff, 1975), but in a generalisation including affine gap penalty Knudsen (2003) switched to a temporal formulation.

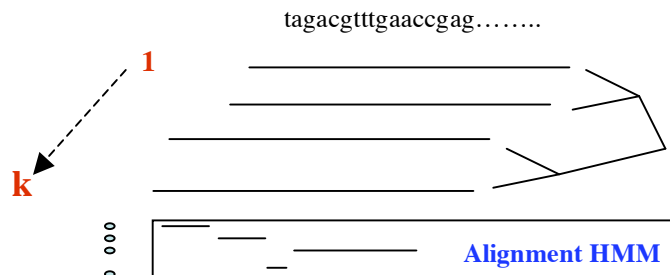
Multiple Statistical Alignment Algorithm is traditionally spatial (Hein, Jensen and Pedersen, 2003), but here we propose a temporal version. The commonality of all the algorithms is the observation as illustrated in this figure:



On the left we observe 6 neighbor positions (and thus 5 neighbor pairs) for one unit of time and events can happen between neighbors. On the right we observe 1 set of neighbor positions for five units of time and events can happen in each time unit. When assuming independence, two situations are stochastically equivalent.

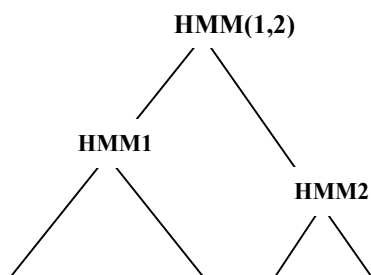


On the left a multiple alignment (or likelihood) is constructed by adding more and more for all sequences at the same time. On the right a multiple sequence alignment is constructed by starting at the bottom and merging (aligning) sequences and alignments until all sequences has entered the phylogeny.



Here 4 sequences (unaligned) are given and are related by a known phylogeny. A Hidden Markov Model (HMM) can be constructed (for details see Lunter et al., 2005), where each state is a possible alignment column. The probability of data (the sequences) can be obtained using the so called forward algorithm summing over all possible alignments. A single alignment can be obtained by the Viterby algorithm.

Clearly the two algorithms achieve the same thing, but they would be very different in how certain computational accelerations like corner cutting were used. Corner cutting ignores entries that are unlikely to be relevant to the solution of the problem. In the pairwise algorithm it will often be entries that are far away from the line connecting (0,0) and (11,12), (l_i is the length of sequence i). In the spatial algorithm corner cutting would have to be applied to the complete k -dimensional matrix, while the temporal formulation would proceed through a series of matrices of increasing dimension and corner cutting could be applied to any of these. This could provide major space and time reductions in favour of a temporal version. Corner cutting can be done in many ways and being efficient in this is crucial for getting a good algorithm. Some approaches can be found in Hein et al. (2000).



On the left, 2 sequences have been aligned using HMM1 and 2 sequences have been aligned using HMM2. How is HMM(1,2) constructed that can combine HMM1 and HMM2? The structure of HMM1 and HMM2 if using the TKF91 model is simple to describe and HMM(1,2) will inherit much of the simple structure. The main difference will be in the input to HMM(1,2) – it is not strings, but matrices (i) and each edge of the input matrices will be a column in an alignment and not a nucleotide (ii). Both (i) and (ii) can be solved in a straightforward manner.

2 month pilot project:

Week 1: Read references below and discuss them with supervisors. Expand this project into 1500 words with more detail.

Week 2: Investigate the pair-wise statistical alignment algorithm: Run existing software and alternative formulations of basic algorithm.

Week 3-4: Make HMM based pairwise aligner that can also handle generalisations of strings, such as higher dimensional alignment matrices. Integrate these into Tree-aligner.

Week 5-6: Investigate different corner cutting approaches and perform simple data analysis.

Week 7-8: Report writing is encouraged to be done continuously during the period, but should be a major activity in this period. The goal should be to be able to do statistical alignment of 4 sequences up to 500 nucleotides long. One could assume that tree and process parameters were always given to avoid a numerical optimisation problem.

References

- Elston, R.C. and Stewart, J. (1971) "A general model for the genetic analysis of pedigree data" *Human Heredity* 21: 523–543.
- Hein, J., C.Wiuf, B.Knudsen, Møller, M., and G.Wibling (2000): *Statistical Alignment: Computational Properties, Homology Testing and Goodness-of-Fit*. (*J. Molecular Biology* 302.265-279)
- J.Hein, J.Jensen and C.Storm (2003) "Algorithms for Multiple Statistical Alignment" (*PNAS* 100(25):14960-14965.)
- Bjarne Knudsen, B (2003): *Optimal Multiple Parsimony Alignment with Affine Gap Cost Using a Phylogenetic Tree*. *WABI*: 433-446
- Lander, E.S. and Green, P. (1987) "Construction of multi-locus genetic linkage maps in humans" *PNAS* 84: 2363–2367
- Lunter, Miklos, Drummond, & Hein (2005) "Alignment, Statistics and Evolution" (in "Statistical Methods in Molecular Evolution" ed. Rasmus Nielsen.)
- Lyngsø, Song and Hein (2005) "Minimal Recombination Histories by Branch and Bound" *WABI* 3692: 239–250.
- Sankoff, D. (1975). Minimal mutation trees of sequences. *SIAM J. Appl. Math.*, 28(35 - 42).
- Song and Hein (2005) "Constructing minimal ancestral recombination graphs" (*J. Compu.Biol.*12.2.147-162)
- Wiuf, C and J.J.Hein (1999): *The Coalescent with Recombination as a point process moving along sequences*. *Theoretical Population Biology* 55.248-259