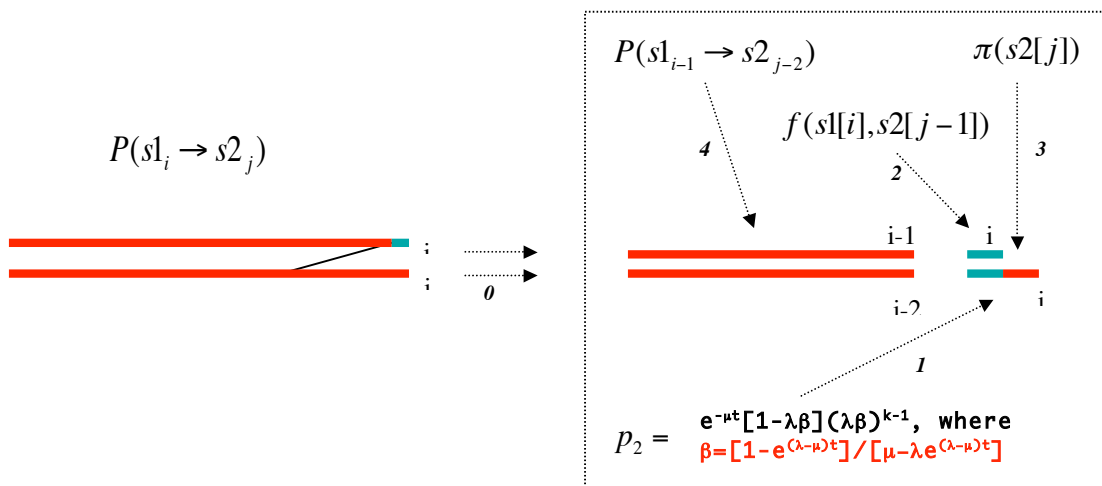


Statistical Alignment MCMC workbench

30.6.07

Since about year 2000 there has been a shift in sequence analysis methods from optimisation based methods to methods based on underlying stochastic model of evolution of sequences. This is a very positive trend that allows for much better and reliable analysis of sequences and genomes. However, statistical alignment is still a computationally very hard problem and it is important to devise efficient algorithms. It seems the way ahead is the use of MCMC algorithms and the aim of this project is to assess the efficiency of different MCMC statistical alignment methods on simple data sets to assess their computational properties. All MCMC methods for integrating over all possible alignments should be reviewed and implemented. There are probably 4-6 methods. This does not involve any investigation of different phylogenies. The reliability of the methods can be compared to a golden standard of dynamical programming algorithms programmed by Gerton Lunter and Rahul Satija.

i. Pairwise statistical alignment. The first model for statistical alignment was proposed for in 1986 by Bishop and Thompson. In 1991 the model was proposed that provides the basis for most statistical alignment today by Thorne, Kishino and Felsenstein. Although statistical alignment for a long time had the reputation of being very difficult, it can actually be solved by algorithms that are very similar to the traditional optimisation alignment. In the first illustration below it is shown how the probability of one sequence evolving into another can be decomposed by tracing what happens to the last nucleotide in the first sequence. Formulated as such, it will be cubic in running time, but this can be reduced to quadratic by a simple trick.



The probability of a given sequence evolves into another sequence can be decomposed using standard dynamic programming techniques. The probability that a prefix of length i of the first sequence evolves into a prefix of length j of the second sequence, $P(s1_i \rightarrow s2_j)$, can be decomposed into a series of cases by exhaustive listing of what happens to the i 'th nucleotide of $s1$. It can survive into $s2$ and have between 0 (including itself) and j descendants. To right only one case – it survives and has 2 (including itself) descendants. This case will be the product of four factors, three of which can be readily calculated. p_2 is the probability that a nucleotide survives and has an extra descendant. $f(s1[i], s2[j-1])$ is the probability that the actual nucleotide (for instance A) $s1[i]$ evolves into $f(s1[i], s2[j-1])$. $\pi(s2[j])$ is the probability of the inserted nucleotide according to the equilibrium distribution of the substitution process.

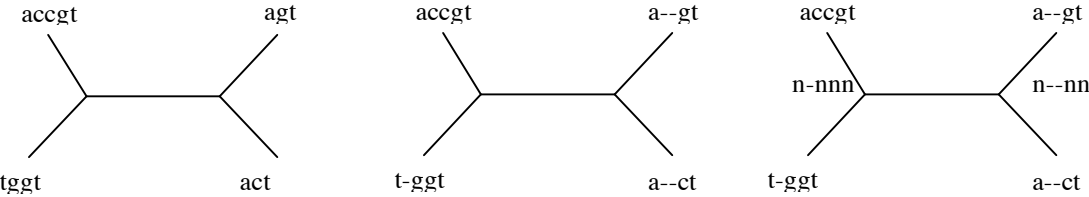
Following the 1991 article there have been a few papers that extended the basic model in obvious directions. Thorne, Kishino and Felsenstein in 1992 extended the model to have longer insertion-deletions by having long unbreakable segments as units instead of single nucleotides. These unbreakable segments clearly were somewhat artificial, and a more natural model was formulated by Miklos, Lunter and Holmes in 2003, that unfortunately was computationally much slower. The next natural steps are heterogeneity in the evolutionary process along the sequence and context dependency, but these extensions will most likely lead to much more complicated algorithms.

Recursion: $P(s1_i \rightarrow s2_j) = p'_0 P(s1_{i-1} \rightarrow s2_j) + \sum_{1 \leq k \leq j} P(s1_{i-1} \rightarrow s2_{j-k}) (p_k f(s1[i], s2[j-k+1]) \pi_{s2[j-k+2:j]} + p'_k \pi_{s2[j-k+1:j]})$

Initial condition: $P(s1_0 \rightarrow s2_j) = p''_j \pi_{s2[1:j]}$

Here the full recursion is written out summing over all possible fates of the i'th nucleotide in sequence 1. If it survives (the p_k) there will be between 1 and j descendants. If it doesn't (the p'_k) between 0 and j descendants. The initial condition states that the second sequence has been created from a prefix of length zero of the first sequence (the p''_j) and then filled with the correct nucleotides.

ii. *Multiple statistical alignment.* The extension from pairwise to multiple statistical alignment is not as straightforward as for optimisation alignment and the first such algorithm was devised by Mike Steel in 1999 (Steel and Hein, 2001). But again the resulting algorithms become very similar to the optimisation algorithms and with similar computational complexity. Although statistical alignment doesn't in itself give an alignment, it assigns probabilities to all alignments and this can be very useful for the interpretation of the evolutionary history of the sequences. It also leads to a series of subproblems (see illustration below): Do we only align the input sequences? Or also the ancestral sequences? Do we make statements about what the ancestral sequences are?



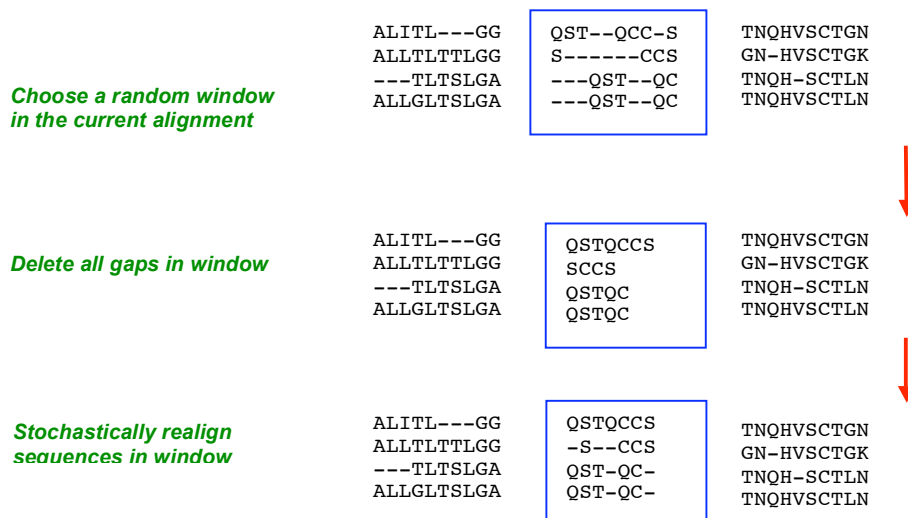
Three versions of the alignment problem: To the left sequences at the leaves are given unaligned. In the middle the sequences at the leaves are aligned/ To the right they are aligned and the alignment of the ancestral sequences are also known. The problem to the left is the most relevant for a researcher as the alignments are unobservable and cannot be expected to be given.

iii. *MCMC Approaches.* The computational burden of the multiple statistical alignment algorithms become completely prohibitive for more than 5-6 sequences and it is then natural to use MCMC algorithms. The first MCMC for statistical multiple alignment was proposed in Holmes and Bruno (2001), but it has precursors for pairwise alignment in Thorne and Churchill (1995) and Lawrence and Liu (1999). There are two aspects to multiple statistical alignment: the phylogeny and the alignment. We will focus on the alignment as the phylogenetic problem is well studied. In the illustration below two approaches are shown that resamples the alignment globally by traversing the phylogenetic tree and decomposes-reassembles the alignment according to a stochastic procedure.



Two groups have used the Gibbs sampler. Holmes and Bruno (2001) applied it to a rooted binary tree with associated alignment and each step would align the root sequences of two binary trees creating a pairwise alignment from which one can sample a proposed ancestral root. This operation is moved around in the tree until convergence. Jensen and Hein (2005) aligned around nodes (in contrast to Holmes-Bruno edge) need trip alignment. Jensen and Hein moved around the tree visiting nodes and re-sampling triple alignments with one internal node. The edge approach is based on a fast operation, but take many steps to converge, while the reverse is true for the node approach.

Below is shown an alternative approach where a subsection of the alignment is resembled. Clear this could be combined with approach above.



A random window is chosen and the alignment information erased. This sequence segments in this window is then stochastically realigned. A new window is chosen etc. This creates a walk in alignment space that in the limit visits all alignments proportionally to their probability.

iv. *The main problems to be addressed.* How is the performance of different methods assessed?

Project Plan

- i. Read key literature and expand this project description to 2000 words.
- ii. Simulate sequences from TKF91. Sequence number vary from 2, 3, 4, 5, 7, 10, 15, 20. Average sequence lengths 20, 50, 100, 200, 500. The phylogeny relating the sequences will be close to a balanced bifurcating tree with parameters (s, λ, μ) on each branch $(.1, .0095, .01)$, $(.2, .0190, .02)$, $(.5, .0475, .05)$ and $(1.0, .0950, .1)$. This gives 160 simulation setups. Each should be repeated 50 times. All experiments can be done with the true parameters given – ie no parameter estimation, which allows focus on how well these algorithms traverse the set of alignments.
- iii. For up to 5 sequences it is possible to compare likelihood to a program that performs the full dynamic programming.

References

- Fleissner R, Metzler D, von Haeseler A. Simultaneous statistical multiple alignment and phylogeny reconstruction. *Syst Biol.* 2005 Aug;54(4):548-61.
- 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.J.Hein (2001): A generalisation of the Thorne-Kishino-Felsenstein model of Statistical Alignment to k sequences related by a binary tree. (*Pac.Symp.Biocompu.* 2001 p179-190 (eds RB Altman et al.)
- J.Hein, J.Jensen and C.Storm (2003) "Algorithms for Multiple Statistical Alignment" (*PNAS* 100(25):14960-14965.)
- Holmes and Bruno (2001) "Evolutionary HMMs: a Bayesian approach to multiple alignment" *Bioinformatics* Vol. 17. 9. 803-820
- Jensen, J.L. & Hein, J. (2005) "A Gibbs sampler for statistical multiple alignment." *Statistica Sinica*, 15.4.889-908.
- Lunter, Miklos, Drummond, Jensen & Hein (2003) "Bayesian Phylogenetic Inference under a statistical insertion-deletion model" (WABI03, Hungary. Lecture Notes in Bioinformatics vol.2812. P228-244)
- Lunter, Song, Miklos & Hein (2003) "A one-state recursion for multiple statistical alignment" (*J. Comp. Biol.*, 10(6):869-889.)
- Lunter, Miklos, Drummond, & Hein (2005) "Alignment, Statistics and Evolution" (in "Statistical Methods in Molecular Evolution" ed. Rasmus Nielsen,)
- Gerton Lunter, Istvan Miklos, Alexei J Drummond, Jens L Jensen and Jotun Hein (2005) "Bayesian Coestimation of Phylogeny and Sequence Alignment" *BMC Bioinformatics* 6(1):83.
- Suchard and Redeling (2006) "BALi-Phy: simultaneous Bayesian inference of alignment and phylogeny" *Bioinformatics.* 22(16):2047-8
- Redeling and Suchard (2005) "Joint Bayesian estimation of alignment and phylogeny" *Sys.Biol.* 54(3):401-18.
- Steel, M. & J.J.Hein (2001): A generalisation of the Thorne-Kishino-Felsenstein model of Statistical Alignment to k sequences related by a star tree. (*Letters in Applied Mathematics* vol. 14.679-684)
- Thorne JL, Kishino H, Felsenstein J. An evolutionary model for maximum likelihood alignment of DNA sequences. *J Mol Evol.* 1991 Aug;33(2):114-24. Erratum in: *J Mol Evol* 1992 Jan;34(1):91.
- Thorne JL, Kishino H, Felsenstein J. Inching toward reality: an improved likelihood model of sequence evolution. *J Mol Evol.* 1992 Jan;34(1):3-16.
- Thorne JL, Churchill GA. Estimation and reliability of molecular sequence alignments. *Biometrics.* 1995 Mar;51(1):100-13.