Lecture 30. Genomics models, part 1. Tree alignment

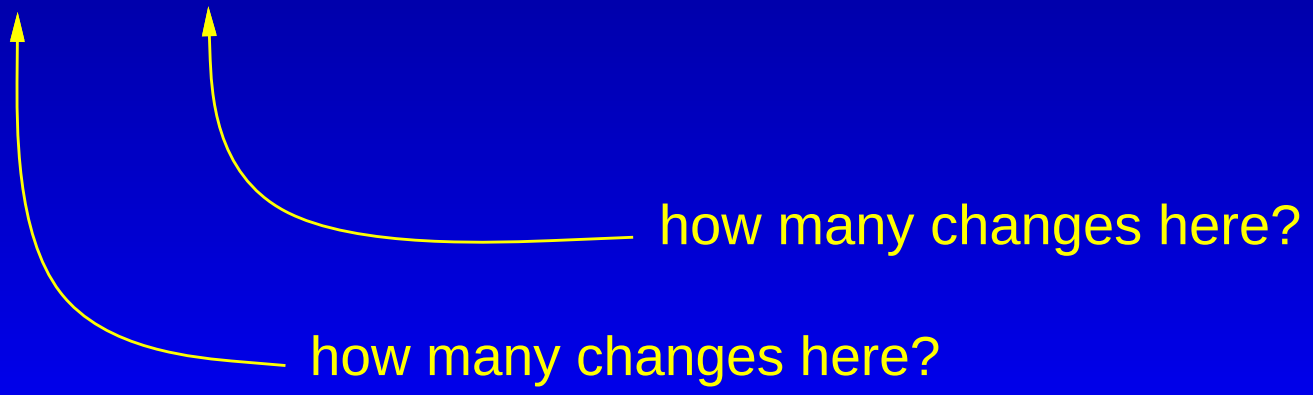
Joe Felsenstein

Department of Genome Sciences and Department of Biology

The difficulty with pairwise alignment – an example

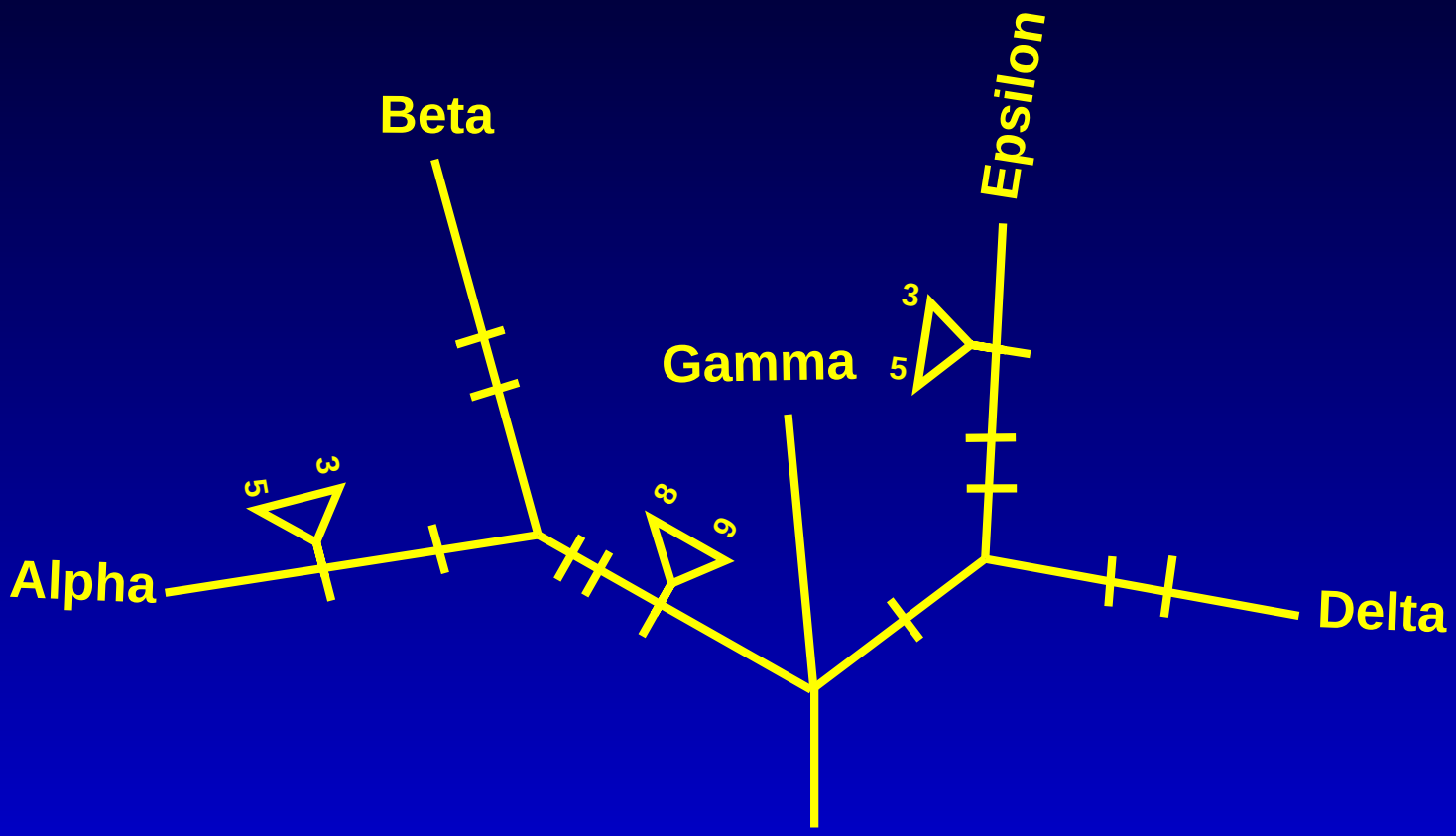
What would we count if there were more than two species, e.g.:

species	sequence
Alpha	A C C G A A T – – A T T A G G C T C
Beta	A C – – – A T – – A G T G G G A T C
Gamma	A A – – – A G G C A T T A G G A T C
Delta	G A – – – A G G C A T T A G C A T C
Epsilon	C A C G A A G G C A T T G G G C T C



Can we evolve those sequences with only 2 indels? No!

Here is a tree that has 3: (plus 10 substitutions)



(it is one of 6 trees tied for best, and there are also alternative placements of some of the changes shown above)

Sankoff's integration of phylogenies, alignment

Sankoff (in Sankoff, Morel, and Cedergren, 1973; see also Sankoff, 1975 and Sankoff and Rousseau, 1975) suggested that one should infer alignments and phylogenies simultaneously, reconstructing sequences at interior nodes, and scoring a tree by the sum of the alignment penalties along its branches.

Sankoff, Morel and Cedergren produced a tree for 5S RNA in this way, using the biggest computers available then.

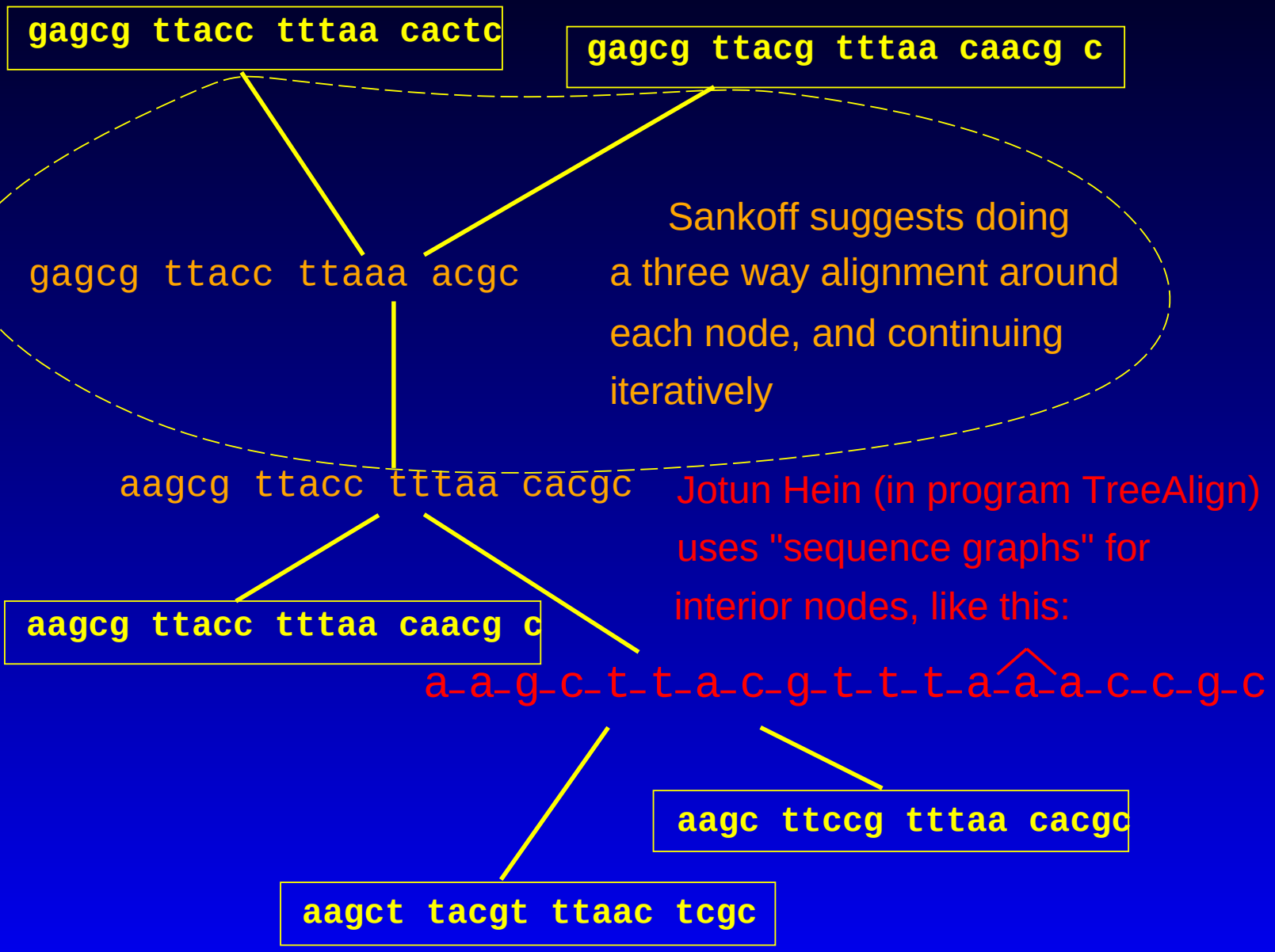
This is basically a parsimony approach. It is of course at least as hard as the no-insertion, no-deletion ordinary parsimony case (in fact it's a lot harder).

More recent developments

Jotun Hein (1990) implemented an approximate version, with some distance-based corrections to weight branch lengths, in his program `TREEALIGN`.

Sankoff's original scheme has been more recently implemented in Ward Wheeler's program `MALIGN`. Karl Nicholas's `GENEDOC` is a Windows program that will align along a given tree.

Sankoff, Morel, and Cedergren algorithm (1973):



Progressive Alignment

(Feng and Doolittle, 1987)

- Do all pairwise alignments
- Make a “guide tree" using the alignment scores as distances
- Align neighbors on the tree, merging alignments as one goes down (“once a gap always a gap")

Implemented in `ClustalV, ClustalW` (Higgins and Sharp, 1989) and the GCG package’s `PILEUP`. Early versions of these used UPGMA for trees, more recent ones use Neighbor-Joining.

So far the most successful family of alignment programs. It gave credibility to tree alignment approaches after they were mostly ignored while the multiple sequence alignment literature developed arbitrary methods that do not take relatedness of sequences into account.

Difficulty with Clustal progressive alignment

Problem: does not go back up the guide tree. Decisions made in different branches can be incompatible with each other and are all retained:

Maximum Likelihood Alignment

Bishop and Thompson (1986) pioneered use of a probabilistic model of 1-base insertion and deletion. Their model allowed transition probabilities to be calculated (approximately, for short divergence times).
The likelihood uses a dynamic programming algorithm to sum the total probability of all sequences of insertion, deletion, and base change that could lead to one sequence from another:

$$\mathsf{L} \;=\; \mathrm{Prob}(\mathsf{B} \,|\, \mathsf{A}, \mathsf{t}, \lambda, \mu)$$

Thus it *does not choose a single alignment* but *adds up over all possible alignments*. Once the parameters and divergence time have been estimated we can find the single alignment that contributes most to the likelihood.

Adding up the likelihood

is a dynamic programming algorithm of the usual sort:

$$\begin{aligned} S(A_m, B_n) &= S^1(A_m, B_n) + S^2(A_m, B_n) \\ &\quad + S^3(A_m, B_n) + S^4(A_m, B_n) \end{aligned}$$

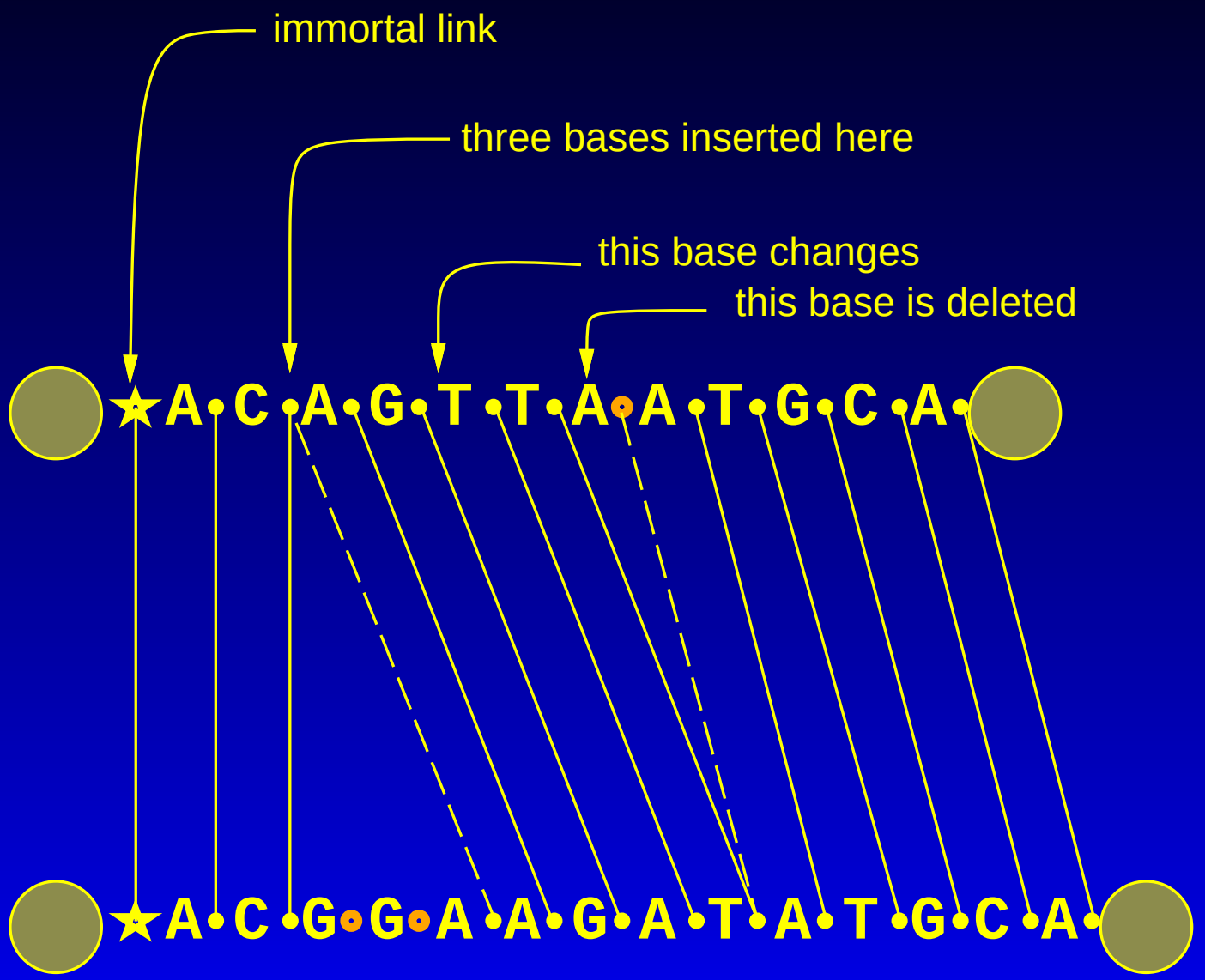
except that it *adds up* the probabilities of getting to each point from the three preceding points. So we are getting the *sum of the probabilities* of all ways of aligning the end of B_m with the end of A_n.

Thorne's model

(Thorne et. al., 1991; see also Allison and Yee, 1990)

- Links can be deleted (rate μ per unit time). The base to its left is deleted with it.
- The leftmost link is immortal.
- Links can be inserted, one at a time to the right of any existing link (rate λ per existing link). A nucleotide randomly drawn from the equilibrium distribution is inserted to the left of each new link.
- Bases can change (at rate 1 per unit time) according to one of the usual base change models.

Thorne's model, illustrated



Thorne's model: summing up the likelihood

Thorne et. al. (1991) use a dynamic-programming algorithm adding up over all alignments. A key feature is that these two alignments have different meanings:

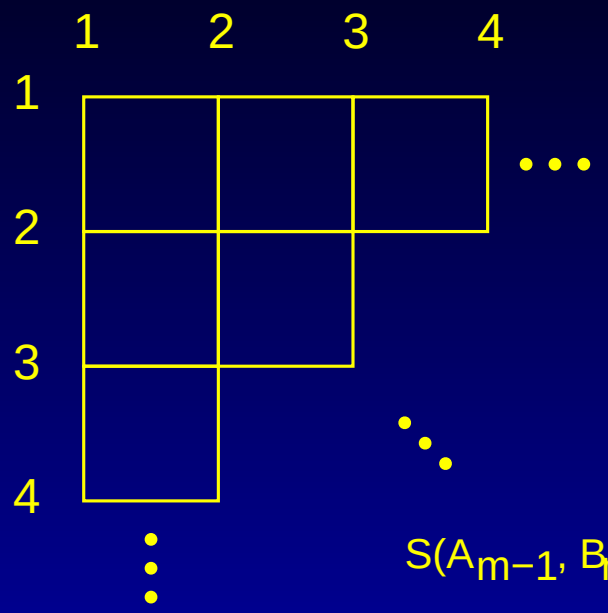
```
T - A G
T A - G
```

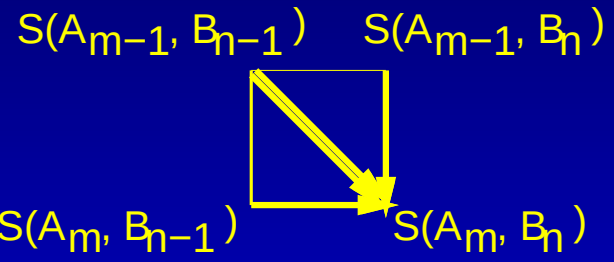
The above means the top A was deleted, then the bottom A inserted after the T.

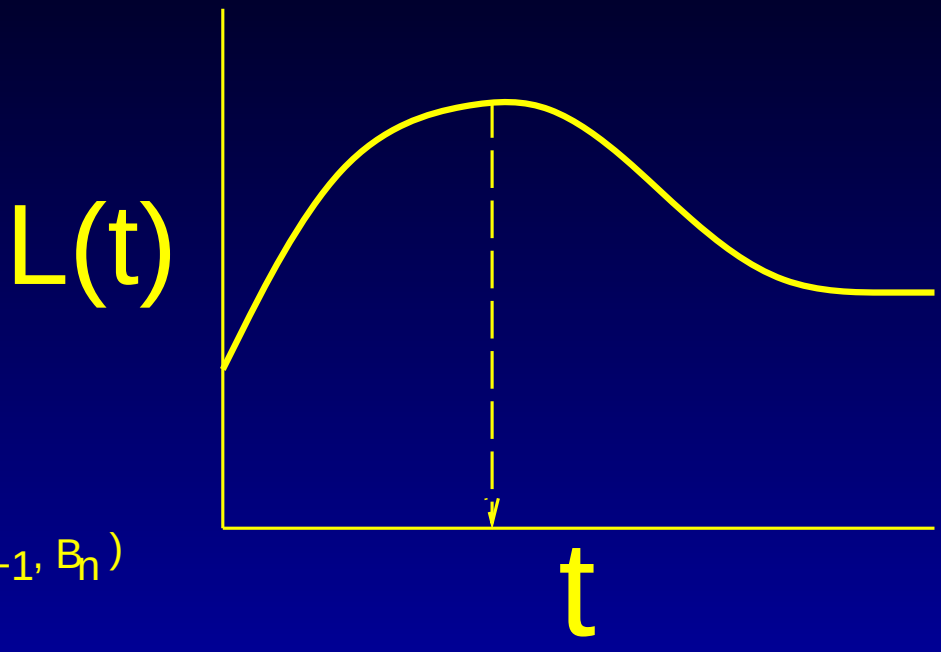
```
T A - G
T - A G
```

But this one means the bottom A was inserted after the top A, which was then deleted later.

Thorne's model: summing up the likelihood







More on Thorne's model

This model has transition probabilities that can be calculated (if $\lambda < \mu$). It can therefore deal with distant divergences.

A dynamic programming method adds up likelihood over all alignments. However it is not realistic in that only single-base insertions and deletions are permitted.

A likelihood ratio test is possible testing whether $t < \infty$. This is in effect a test of whether the sequences are related.

Only practical for few sequences at a time. Jeff Thorne's program `STATALIGN` for two sequences. Jotun Hein has algorithms that can handle small binary trees.

The future

- Markov Chain Monte Carlo (MCMC) methods will be the way maximum likelihood (or Bayesian) multiple sequence alignment will be done.
- Probably the best way to do it is to explicitly place indel events on the tree and move them around (and add and remove them). This way we can use more complex and realistic models such as ones that have indels of length more then 1.
- There is a serious issue of how to run these to adequately sample.
- These methods also give us a “posterior" distribution of different plausible alignments.

References

Agarwal, P., and D. J. States. 1996. A Bayesian evolutionary distance For Parametrically aligned sequences. *Journal of Computational Biology* **3:** 1-18. **[A Bayesian version of statistical alignment]**

Allison, L. and C. N. Yee. 1990. Minimum message length and the comparison of macromolecules. *Bulletin of Mathematical Biology* **52:** 431-453. [**Further development of maximum likelihood alignment]**

Allison, L. and C. S. Wallace. 1994. The posterior probability distribution of alignments and its application to parameter estimation of evolutionary trees and to optimization of multiple alignments. *Journal of Molecular Evolution* **34:** 418-430. **[More on their approach]**

Bishop, M. J. and E. A. Thompson. 1986. Maximum likelihood allignment of DNA sequences. *Journal of Molecular Biology* **190:** 159-165. **[The first paper on maximum likelihood alignment]**

Feng, D.-F., Doolittle, R. F. 1987. Progressive sequence aligment as prerequisite to correct phylogenetic trees. *Journal of Molecular Evolution* **25:** 351-360. **[Progressive alignment]**

References, continued

Hein, J. 1990. Unified approach to alignment and phylogenies. pp. 626-645 in *Methods in Enzymology, vol. 183*, ed. R. F. Doolittle. Academic Press, New York. **[Implementing Sankoff's method with some distance method corrections]**

Hein, J. 2000. Statistical alignment: computational properties, homology testing and goodness-of-fit. *Journal of Molecular Biology* **302:** 265-279. **[Implementation of TKF model alignment]**

Hein, J. 2001. An algorithm for statistical alignment of sequences related by a binary tree. *Pacific Symposium on Biocomputing* **6:** 179-190. **[Getting more real about using TKF with (smallish) trees]**

Higgins, D. G. and P. M. Sharp. 1989. Fast and sensitive multiple sequence alignments on a microcomputer. *Computer Applications in the Biological Sciences (CABIOS)* **5:** 151-153. **[Progressive alignment implemented (in `ClustalV`)]**

Jiang, T., E. L. Lawler, and L. Wang. 1994. Aligning sequences via an evolutionary tree: complexity and approximation. pp. 760-769 in *Proceedings of the Twenty-Sixth Annual ACM Symposium on the Theory of Computing., Montréal, Quebec, Canada, 23-25 May 1994.* ACM, New York. **[Complexity of tree alignment, approximation to it]**

References, continued

Metzler, D., R. Fleissner, A. Wakolbinger, and A. von Haeseler. 2001. Assessing variability by joint sampling of alignments and mutation rates. *Journal of Molecular Evolution* **53:** 660-669. **[Bayesian approach with TKF model and MCMC sampling of alignments and parameter values]**

Sankoff, D. D., C. Morel, and R. J. Cedergren. 1973. Evolution of 5S RNA and the nonrandomness of base replacement. *Nature New Biology* **245:** 232-234 **[The first tree alignment paper]**

Sankoff, D. D. 1975. Minimal mutation trees of sequences. *SIAM Journal of Applied Mathematics* **28:** 35-42. **[Algorithms for reconstruction of sequences at nodes of a tree]**

Sankoff, D. D., Rousseau, P. 1975. Locating the vertices of a Steiner tree in arbitrary space. *Mathematical Programming* **9:** 240-246. **[More on algorithms for reconstruction of sequences at nodes of a tree]**

Sankoff D. D. and R. J. Cedergren. 1983. Simultaneous comparison of three or more sequences related by a tree. pp. 253-263 in *Time warps, string edits, and macromolecules: the theory and practice of sequence comparison*, ed. D. Sankoff and J. B. Kruskal. Addison-Wesley, Reading, Massachusetts. **[Further development and review of tree alignment]**

References, continued

Metzler, D., R. Fleissner, A. Wakolbinger, and Schwikowski, B, and M. Vingron. 1997. The deferred path heuristic for the generalized tree alignment problem. *Journal of Computational Biology* **4:** 415-431. **[Improvement on Hein's sequence graph method]**

Steel, M. and J. Hein. 2001. Applying the TKF model to sequence evolution on a star tree. *Applied Mathematics Letters* **14:** 679-684. **[Faster calculations for TKF on a "star" tree]**

Thorne, J. L., H. Kishino, and J. Felsenstein. 1991. An evolutionary model for maximum likelihood alignment of DNA sequences. *Journal of Molecular Evolution* **33:** 114-124. **[Further development of maximum likelihood pairwise alignment]**

Thorne, J. L., H. Kishino and J. Felsenstein. 1992. Inching toward reality: an improved likelihood model of sequence evolution. *Journal of Molecular Evolution* **34:** 3-16. **[Even further development of maximum likelihood pairwise alignment (a crude model of larger-scale insertions and deletions)]**

Thorne, J. L. and H. Kishino. 1992. Freeing phylogenies from artifacts of alignment. *Molecular Biology and Evolution* **9:** 1148-1162. **[Show that trees and alignments reinforce each other so if do one first, biases the other]**

References, continued

Thorne, J. L. and G. A. Churchill. 1995. Estimation and reliability of molecular sequence alignments. *Biometrics* **51:** 100-113. **[EM method for updating parameters in a likelihood alignment]**

Metzler, D., R. Fleissner, A. Wakolbinger, and Wang, L. and T. Jiang. 1994. On the complexity of multiple sequence alignment. *Journal of Computational Biology* **1:** 337-348. **[Tree alignment is NP hard]**

Wheeler, W. 1996. Optimization alignment: the end of multiple sequence alignment in phylogenetics? *Cladistics* **12:** 1-9. **[Assumes that at each region which has gaps, interior nodes of tree must be same sequence as one of the tips. An approximation, related to a step in Jiang et al.'s approximate approach.]**

How it was done

This projection produced as a PDF, not a PowerPoint file, and viewed using the Full Screen mode (in the View menu of Adobe Acrobat Reader):

- using the `prosper` style in LaTeX,
- using Latex to make a `.dvi` file,
- using `dvips` to turn this into a Postscript file,
- using `ps2pdf` to mill it into a PDF file, and
- displaying the slides in Adobe Acrobat Reader.

Result: nice slides using freeware.