Genetics/MBT 541

Spring, 2001

lecture 17

Comparative genomic methods

Joe Felsenstein

Department of Genetics

email: joe@genetics
Gene families – Reconciled trees

Goodman et. al. (1979) suggested finding the combination of phylogeny and gene tree that minimized a parsimony score that was the sum of the number of nucleotide substitutions, the number of gene duplications and the number of losses (called by them “gene expression events”).

Page (1994) and Page and Charleston (1994) developed algorithms to compute the score for a given gene tree and a given (assumed known) phylogeny (see also Mirkin et. al., 1995 and Zhang, 1997).
species tree

gene tree

reconciled trees
Probabilistic models of gene families

Lindsey Dubb, *in. prep.*) has developed a probabilistic model approach to gene families.

Given a model of duplication and loss of copies of a gene family, modelled by a simply birth-and-death process with rates $\lambda$ and $\mu$, we can evaluate the likelihood for a phylogeny, summing over all histories ($H$) of duplication and loss that would lead to the observed gene tree ($G$).

$$L = \sum_H \text{Prob} (H | \lambda, \mu) \text{Prob} (G | H)$$

This involves summing over all possible numbers of “doomed” lineages that could exist at each interior node of the tree.

If we do not know the gene tree precisely, we can also sum over all histories the probability of getting the observed data via that history:

$$L = \sum_H \sum_G \text{Prob} (H | \lambda, \mu) \text{Prob} (G | H) \text{Prob} (D | G)$$

This last must be done by Markov Chain Monte Carlo (MCMC) methods.
Uses of likelihood approaches

Note that neither of the preceding approaches infers a single history $H$ – their objective is to get likelihoods for the parameters $\lambda$ and $\mu$.

If one does not infer the gene tree, what good is this?

- One can get estimates of rates of duplication and loss.
- One can use likelihood ratio tests of whether these rates are the same in different gene families.
- One can test assertions that the duplication rate is lower (or the loss rate higher) when there are many copies of a gene family (to do this one needs a model with rate of duplication dependent on copy number, as an alternative).

There is a danger in inferring gene tree $G$ and then treating this inference as if it were data. One’s statistical methods then do not take into account the error in inferring $G$. 
To sum the probability of getting the observed gene tree, we need to add up over all possible (unobservable) histories of "doomed" lineages (here an example is shown with dark dashes involving one extra duplication and three extra losses)
Inversions

Watterson et. al. (1978) posed the problem of how to transform one sequence of markers into another with the minimum number of inversions ("reversal distance"). Caprara (1999 showed this is NP-hard.
Hannenhalli and Pevzner (1995, 1999) found that when the problem is recast as having known orientation of the genes, a polynomial-time algorithm can be constructed for computing the distance.
The polynomial algorithm is, in worst-case, $O(n^5)$ which is rather slow. It can be used to make a distance method, but this is not corrected for reversals. It would be better to analyze multiple genomes by a parsimony method. This would require development of an algorithm for reconstructing ancestral gene orders, for a minimum of three genomes. This has not yet been done in a polynomial-time algorithm.
The Hannenhalli-Pevzner algorithm makes use of a *breakpoint graph* which represents each signed marker by a pair of points (one at its beginning and one at its end). Dashed curves are drawn between originally adjacent points that are now separate:

\[
\begin{align*}
-6 & \quad -5 & \quad +11 & \quad -2 & \quad -1 & \quad +7 & \quad -12 & \quad +3 & \quad +4 & \quad -10 & \quad -9 & \quad -8 & \quad +13 \\
\end{align*}
\]

gives the following as the breakpoint graph
Breakpoint approximations

Blanchette, Bourque and Sankoff (1997) suggested approximating the criteria (for several kinds of events) by counting the number of times breakpoints arise or are eliminated. Minimizing this is not quite the same thing as minimizing the number of events, but it is a lot easier to handle.

Sankoff and Blanchette (1998) propose finding the tree (and the ancestral genomes in it) that minimize the number of breakpoints summed over all branches in the tree. This can be done (approximately) by finding the ancestral genome at each point in the tree that minimizes this criterion given the three neighbors:
This is a “median problem” which is NP complete, but is doable for moderate size genomes because it can be reduced to the Travelling Salesman Problem (TSP).

They construct a complete graph the weights of whose edges are the number of times that pair of markers is adjacent in the set of neighboring genomes. The TSP problem is solved for this graph.
Probabilistic models

Sankoff and Goldstein (1989) were the first to propose a probabilistic model (of random inversions). This was highly intractable – it appears that Markov Chain Monte Carlo methods are the only practical way to evaluate likelihoods and search among phylogenies.

Note that one may not be inferring the sequence of events, but rather the species tree that creates the conditions for a variety of possible events.

Sankoff and Blanchette (1999) have approximated the appearance and disappearance of breakpoints as a multi-state Jukes-Cantor model and based an approximate phylogenetic analysis on that.
References


Hannenhalli, S., C. Chappey, E. Koonin, and P. Pevzner.


