

Detecting Correlated Events in a Pathway by Phylogenetic Analysis

1.4.09

Motivation and Background. Bacteria are noted for the metabolic diversity and the degree to which they exploit many niches. This is done by fast evolution of their metabolic capabilities. The genomes of many closely related bacteria are now available. This offers the possibility of tracing metabolic evolution on a phylogeny relating the genomes. Using simple (independent loss/gain or reactions) or complex (incorporating dependencies among reactions) stochastic models of metabolic evolution, it is possible to test if there are coordination between the evolution of different reactions. We will only use simple models in this investigation. The basic models were developed to detect coordinated evolution between a pair of traits for instance a morphological trait and a nucleotide position in a series of recent papers (Huelsenbeck et al.,2003; Nielsen,2002). These models can be adapted to the set of traits found consisting of all possible reactions and possible metabolic capabilities of a bacteria. The main question is how to incorporate the network structure on the set of reactions. Just viewing all reactions as independent would be uninteresting or not relate well to the biological problem.

Which questions could be asked by simple extensions of correlation analysis? Is there a tendency for neighbour reactions to be coupled? Is the appearance of a reaction increase the likelihood of the appearance of a reaction neighboring and downhill to it? That would be a reasonable expectation.

Similar dependency models have been analysed for sequences by Pedersen, Jensen, Hobolth, Siepel, Haussler and others. The main novelty in the present model would be that the sequential structure has been changed into a graphical structure.



To the left is illustrated the evolution of a pair on one branch of a phylogeny. 00 is the starting configuration and 11 the ending configuration of the pair. Simple techniques allow the calculation of the distribution of much time was spent in the four possible configurations [(0,0), (0,1), (1,0), (1,1)]. This can be compared with distribution of the being in these states but not conditioning on start and ending configuration. On a single branch this is not very powerful, but on a large phylogeny the situation is different. One possible evolutionary trajectory is that 0,0 (two thin lines) evolved into 0,1 that evolved into 1,1 (two thick lines)

To the right a simple example shows a metabolic universe of 6 reactions and 4 are present (solid lines) in the metabolism now. Two edges are neighbors if the output of one is the input node of the other. This gives us 15 possible pairs of reactions. Any pair can be tested for convergence. There are 6 neighboring pairs.

It would be natural to expect that there would be positive selection for the appearance of a reaction, if there was a substrate for it, created by another reaction. In the little metabolism above, the presence of “b” would be favoured if “a” was present. Ideally, a model for the evolution of a complete network should be explored, but for the present purpose, we will not explore the correlation of neighbor reactions with taking the “contagious dependence” were correlations are created between non-neighboring reactions via a path of correlations between neighbors. Ignoring “contagious dependence” will seriously reduce the hardness of the problem and make much larger analysis feasible. We will call this the “independence heuristic” - IH. IH will make reaction appear multiple times in the overall probability expression, as they can be neighbors via multiple reactions.



To the left is shown the presence (filled black circle) and absence (empty circle) configurations observed at the leaves of a tree. The question is: Is there is a correlation in the evolution of the two. If they signified reaction “a” and “b” above, then one could expect so. Clearly any test of correlation would benefit from using a large number of observed metabolisms, ie a large phylogeny.

If three reactions all were neighbors through metabolite M, it would be of interest to investigate models for all three reactions simultaneously to investigate if there were correlations to be explained beyond pairwise correlations.

