

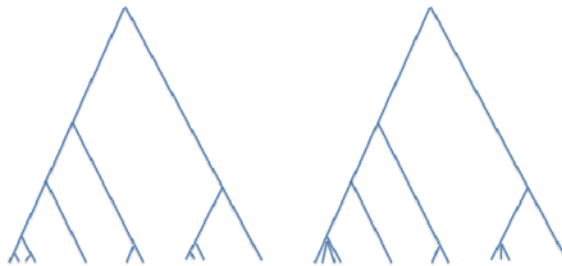
Approximate Local Trees and CoaHMMs

1.10.2010

Genomes are now routinely being determined for large sets of individuals within a population and methods to analyze these have been investigated. Such analyses can estimate the fundamental parameters of the population model describing the population or allow ancestral analysis: statements about historical path of the genomes and genes, which can reveal events of interests such as migrations and selective sweeps.

The ideal structure to describe the relationship among the genomes is the ancestral recombination graph (ARG) that was first described by Griffiths (1982) and Hudson (1983). However, for several reasons making precise statements about the ARG is unrealistic. One reason is that all knowledge obtained from the genomes are at best the local trees for all genomic regions, but local trees don't determine the ARG. A second reason is that obtaining the local trees is computationally very demanding. Thus investigating approximate but fast methods to local trees are of great interest. One first natural approximation is to discretize the continuous time in the local trees. This has been explored in the so-called coaHMMs (). Even if only the topologies were registered for the local trees, there is a very large number of these. However, it is expected from coalescent theory that these tree topologies contains sub-trees with very short branches that rarely changes as we move along the genomes and what really is of interest is the longer branches that relate these subgroups.

This project explores HMMs what only has a reduced state space, by collapsing these very closely related leaves into one state. However, it is not possible to know in advance which sets of sequences are closely related and should be collapsed.



The tree on the left is a standard phylogeny/local tree where all inner nodes connect on an ancestor and two descendants corresponding to a duplication events. The number of topologies with k leaves is $k!(k-1)!/2^{k-1}$, which is very large. To the right is a compressed version of the same tree, where two groups have been collapsed and their internal relationship has not been fully described, but they are assumed to be equidistant. This will essentially reduce a k -tree to a k' -tree, where the created groups are replacing their members as leaves. In this example an 11-tree is reduced to a 7-tree. For very large trees, the reduction would be more drastic.

Some choice will have to be made before it is possible to design an HMM that jumps between neighboring local trees:

1. How many leaves and collapsed sets are there? It seems reasonable to be set by a single parameter $(k-k')/k$. It could easily be investigated by simulation how good the approximation would be as a function of this.

2. Are the sets to be collapsed pre-given? It is pre-given from a pre-analysis of the data and the groups can't be changed, then the k-tree problem has been reduced to a k'-tree problem with some "error" within the groups.
3. Within the collapsed sets, are all sequences to have the same distance to the root of the collapsed set. It seems a nice assumption to make or one could choose a distance that fitted the size of the subset using coalescent theory.
4. When jumping between neighboring trees, can the collapsed sets be changed? For larger data analysis, it seems unreasonable to keep them totally fixed. But at the same time, they cannot be allowed to be dissolved/created freely because then any computational gain would be lost.

If a fast approximate tool is to be devised, one should make the choices that allows the fastest computations.

Project Plan.

References

Hein, Schierup and Wiuf (2005) Gene Genealogies, Variation and Evolution OUP

Julien Y. Dutheil, Ganesh Ganapathy, Asger Hobolth, Thomas Mailund, Marcy K. Uyenoyama and Mikkel H. Schierup (2010)
"Ancestral Population Genomics: The Coalescent Hidden Markov Model Approach" Genetics 183: 259–274