Poster C-32

Clustering to minimize genetic overlap between clades

Authors:
Mukund Narasimhan (*Microsoft Corp.*)
Nebojsa Jojic (*Microsoft Corp.*)

Short Abstract: Phylogenetic tree construction methods which rely on multiple alignment as a foundation are restricted to gene/gene-fragments/small proteins. We show that parsimony satisfies a property called submodularity, and hence we can use recently discovered algorithms for submodular function minimization to construct maximum parsimony phylogenetic trees.

Long Abstract:
Traditional approaches to constructing evolutionary trees use multiple alignment as a foundation, and hence are unsuitable in the presence of large scale evolutionary events like recombination and horizontal gene transfers. This restricts the applicability of these techniques to gene/gene-fragments and (small) proteins. Attempting to determine evolutionary history
from small fragments of the genome yields results that
depend on the part of the genome chosen for this
phylogenetic analysis. As a result, different parts of the genome could lead to different evolutionary trees. Using the entire genome seems to be the best way of resolving these inconsistencies.
There has been some recent work which tries to construct distance metrics derived from the complete genome, and then apply distance based techniques like Neighbour-Joining to construct phylogenetic trees.
In this work, we present a novel new technique of constructing phylogenetic trees based on measures of similarity of multiple (i.e., more than 2)
sequences at a time. We show that many commonly used measures satisfy a property known as submodularity, and hence we can used recently discovered polynomial time algorithms for submodular minimization to construct phylogenetic trees based on parsimony instead of distances.
We also show that this method can be used for classification of HIV strains which yields better results than standard algorithms.
We present results on synthetic data and HIV data.

Our approach to constructing phylogenetic trees is based on minimizing the {\em genetic overlap between subtrees/clades}. Phylogenetic analysis often targets only the select genes of a species, under the assumption that the inferred evolutionary tree
would be very close to the true evolutionary tree of the species. However, this assumption does not always hold. For example, in [Holmes et al., PLoS Biology, 2005], it is shown that the phylogenetic analysis of the Human Influenza A H3N2 virus computed from various different coding regions of the genome results in different evolutionary trees. Still, in general,

using more of the information for
phylogenetic analysis poses several problems for traditional phylogeny algorithms. Both the maximum likelihood and maximum parsimony approaches to phylogeny require a sequence alignment, a
step complicated by evolution itself, esp. the events such as gene duplications, genome rearrangements, and reversals. To alleviate this problem, it is possible to define simple distance
measures which are easily computable without a single global alignment. For example, [Qi et al., Journal of Molecular Evolution, 2004] uses a distance metric based on the frequency of strings of length k, and
[Li et al., Sequence Distance and Phylogeny 2002], uses a distance metric based on an approximation of Kolmogorov complexity. Usually, once the distance measure is defined, a standard algorithm like Neighbour-Joining [Saitou et al., Molecular Biol.
Evolution, 1987] is used to search for an optimal phylogenetic tree. In addition to local
minima issues, describing clusters of genetic diversity strictly by pairwise sequence distances has its inherent representational limitations. For example, reasortment of genetic material in
influenza, or recombination in HIV, may create clades of viruses that can easily exchange large chunks of the genome, thus dramatically increasing pairwise distances. These clades are better described by a set of patterns that can be combined, than
by pairwise distances within the cluster and the distance to the viruses in other clades.

In this paper we describe an approach to extracting clusters of genetic diversity based on inter-cluster, rather than inter-sequence distance. The distance between two clusters is defined as the number of genetic components (e.g., short subsequences from the sequences in the cluster) that are shared
between the two clusters. This cluster distance is inherently global, in the sense that it cannot be computed from pairwise sequence distances only. However, we show that finding the optimal
split of the data into two clusters is equivalent to optimizing a submodular function, lending itself to polynomial optimization, e.g., by Queyranne's algorithm [Queyranne, Math. Programming, 1998].

As in previous clustering approaches to phylogeny, the split hierarchy can be thought of as a phylogenetic tree. To show this, we generated 16k random trees with different parameters, sampled sequences from them and reconstructed the
trees using our algorithm based on cluster rather than
sequence similarity, and compared it with recent distance-based techniques described in [Li et al., Bioinformatics, 2001] and [Qi et al., Journal of Molecular Evolution, 2004].

While this experiment only shows that our approach is more robust to tree imbalance (non-uniform rates of evolution), defining inter-cluster distance as the number of short subsequences shared, also has more obvious justification in biology. Many molecular
processes require preservation of shorter or longer motifs, and our measure of cluster similarity is less sensitive to complex evolutionary events such as duplication, re-arrangement and (within cluster) recombination.

To illustrate the utility of our technique on real biological data, we analyzed a species that undergoes complex and fast evolution - HIV 1. Application of our clustering algorithm to sequences from the Los Alamos HIV database leads to unsupervised reconstruction of major HIV clades without the need for alignment. Furthermore, as the cellular arm of the immune system reacts only to short \emph{linear} 9-11 amionacid patterns (epitopes) in viral proteins, the cluster centers are likely to provide a good summary of HIV diversity for cellular vaccine design.

The key contribution of this paper is the introduction of a sequence clustering technique which
the shared genetic information between the clusters. Even though this measure of cluster distance is more global in character, and cannot be derived from pairwise sequence distances alone, we show
that a split into two clusters can be performed optimally in polynomial time under some reasonable assumptions. We further show that the technique can be used to create phylogenies without alignment, outperforming recent techniques for alignment-free
phylogenetic analysis. Furthermore, we show that our technique is a reasonable alternative, and for some applications possibly a better fit, than the state of the art phylogenetic analysis of HIV
sequences, which depends on high quality -- and usually semi-manual -- sequence alignment.