Poster C-4

A phylogenetic reconstruction method based on genetic algorithms and hierarchical clustering

Authors:
Ryosuke Watanabe (*Computer Science Department, ITESM Campus Estado de México*)
Edgar E. Vallejo (*Computer Science Department, ITESM Campus Estado de México*)
Enrique Morett (*Department of Cellular Engineering and Biocatalysis, IBT UNAM*)

Short Abstract: We present a method for phylogenetic reconstruction based on genetic algorithms and hierarchical clustering. We formulate the fitness function as a combination of UPMGA and MP scores. We use taxas from several lineages to evaluate the proposed method. Experimental results indicate that our model is capable of producing accurate phylogenies.

Long Abstract:

Over the years, bioinformatics has been a central methodology in molecular evolution research. Particularly, phylogenetic reconstruction algorithms have been instrumental in the understanding of evolution at the molecular level.

There is an increasing interest in applying genetic algorithms to computationally intensive bioinformatics problems. This holds much promise. For example, previous work on phylogenetic reconstruction using genetic algorithms has proven to produce competitive results.

A fact that complexifies phylogenetic reconstruction is that the optimality criterion itself may be a misleading representation of the evolutionary history. In general, there is no sound and complete method capable of producing the optimal phylogenetic. For example, Distance Matrix Methods have obvious limitations for constructing molecular phylogenies when the rate of gene substitution is not constant. Another example is Maximum Parsimony, which produce the correct tree only when there are no backward and no parallel substitutions at each nucleotide site and the number of nucleotides examined is very large. We would like to posit here that different perspectives of phylogenetic reconstruction can be combined to produce more accurate phylogenetic reconstruction.

Previous work on the application of evolutionary algorithms to phylogenetic reconstruction have almost exclusively focused on the time complexity of the phylogenetic reconstruction problem. The genetic algorithm is used to systematically search the space of phylogenetic trees in an attempt to find the optimal tree. We believe that evolutionary algorithms provide not only an efficient method for searching large solution spaces, but also provide the opportunity to explore the space of optimality criteria. We propose here to exploit this opportunity by defining an hybrid fitness function consisting of a
linear combination of the unweighted pair-group method with arithmetic mean (UPMGA) and the Maximum Parsimony scores. We believe that this approach holds the potential for efficiently establishing a good accurate meassurement to characterize the evolutionary

history of species.

We designed a genetic algorithm with ad hoc operators to evolve a population of phylogenetic trees. The chromosome of the GA consists of a hierarchical clustering representation of the leaves of the tree. Pairwise distance comparisons are conducted on this sequence to reconstruct the original tree. In addition, we designed a collection of specialized genetic operators that are appropriate for the proposed representation. These genetic operators are designed in such a way that their application to parent trees guarantee to produce offspring trees that fall into the solution space.

We evaluated the proposed method using protein sequences from the most representative lineages (8-132 taxa). The results were
compared to those produced by several methods provided by the PHYLIP package.

The grouping of sequences produced by the different methods is similar as they created a clear separation between sequences from different species. Particularly, the tree produced by our method is similar to that produced by the Maximum-Likelihood method, which is considered one of the most accurate phylogenetic reconstruction
Overall, experimental results showed that this model is capable of producing accurate trees in a reasonable time. The results were consistent even when either a moderate or a large number of taxa was considered. The design of accurate and efficient phylogenetic reconstruction algorithms would provide an alternative framework for understanding the properties of evolutionary process of the molecular level.

Felsenstein, J. 1995. PHYLIP (phylogeny inference package. Version
3.61. Department of Genetics, University of Washington.

Lewis, P. O. 1998. A genetic algorithm for maximum-likelihood phylogeny inference using nucleotide sequence data. Mol Biol Evol. 15(3):277-83.

Lemmon A. R., Milinkovitch M. C.
The metapopulation genetic algorithm: An efficient solution for the problem of large phylogeny estimation.
PNAS 99(16):10516-21

Matsuda, H. 1996. Protein phylogenetic inference using maximum-likelihood and genetic algorithms. In L. Hunter and T.E. Klein, eds. Proceedings of the Pacific Symposium on Biocomputing '96. World
Scientific.

Nei, M. and S. Kumar 2000.
Molecular Evolution and Phylogenetics.
Oxford University Press.