Final Technical Report:

Analysis of Molecular Data Using Statistical and Evolutionary Approaches

Kevin Atteson

and

Junhyong Kim
DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, make any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.
DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.
1 Overview of Accomplishments

In this report, we discuss research projects which were undertaken or completed during the author's tenure as a Sloan/DOE Postdoctoral Fellow in the Ecology and Evolutionary Biology Department of Yale University from November 1997 through June 1999. In particular, the following research projects, begun in the author's previous appointments, were completed during this period:

1. The computation of the probabilities of patterns occurring in DNA sequences (website developed and paper presented summer 1998).

2. The asymptotic redundancy of Bayes rules for Markov chains (journal article appeared September 1999).


In addition, the following research projects, begun in the author's current appointment, were completed during this period:

1. Convex hypothesis testing with applications to phylogenetics (paper presented summer 1999 and journal paper to be submitted).

2. Identifiability of phylogenetics with gamma distributed rates across sites.

Finally, the author contributed to the following research projects initiated by others.


2. Population genetics of unequal crossover (publication status to be determined).

1.1 Work on the Asymptotic Redundancy of Bayes rules for Markov Chains

During the author's tenure at Yale, he prepared the final draft of the work from his thesis on applications of statistical decision theory to data compression, in particular when Markov chains are used at the statistical model. This work was published in the IEEE Transactions on Information Theory[Att99a].
1.2 Computation of Pattern Probabilities

As the databanks of available DNA and protein sequence data grow, so does the number of groups searching for patterns in these sequences. This project, begun while the author was a post-doc at Penn, was aimed at providing tools for determining the significance of various patterns in DNA. In particular, algorithms and software were developed to determine the probability of a regular expression. These algorithms were both faster and more general than those which had been previously presented in the literature. During the author's tenure at Yale, the author's paper was accepted, the final draft was prepared and presented and published for the Seventh International Conference on Intelligent Systems in Molecular Biology [Att98]. In addition, a website was developed providing the software tools developed.

2 Consistency of Neighbor-Joining Phylogenetics

In 1978, Felsenstein demonstrated that the still popular parsimony method of phylogenetics would not produce the true tree even given arbitrarily large quantities of data for a simple statistical model. In short, parsimony is statistically inconsistent. In the last decade, there have been numerous results elucidating the consistency of various methods and the identifiability (essentially, the existence of consistent methods) for various models in this area. This paper represents the author's first contribution in this area, demonstrating that the class of methods including the most popular of the computationally efficient methods, neighbor-joining, is in fact statistically consistent. Furthermore, the author provides finite sample size performance bounds. This work was started during the author's post-doc at Penn but the final draft was submitted during the post-doc at Yale. This paper appeared in [Att99b].

3 Convex Hypothesis Testing with Applications to Phylogenetics

In 1994, Steel demonstrated that no method can possibly reconstruct phylogenetic trees when rates vary across sites. In particular, phylogenies are unidentifiable when rates vary arbitrarily across sites. While parsimony is inconsistent, it does have a certain robustness to variation of rates across sites due to the convex nature of its acceptance sets, as noted by the author's mentor, Junhyong Kim. In this work, the author demonstrated that parsimony is in fact robust to

1 Note that the term "regular expression" is sometimes used differently in bioinformatics than in formal language theory. Here we are using the term in the more general formal language theory sense.
a far larger set of models in which the rates can be dependent across sites. This work was discussed as part of a series of 2 talks by the author and his mentor at the 1999 Evolution meeting. In addition, this work is to be submitted for journal publication in the near future[Att00b].

4 Identifiability of Phylogenetics with Gamma Distributed Rates across Sites

While phylogenetic trees are unidentifiable when rates are distributed arbitrarily across sites, the author has been able to show that trees are identifiable when rates are distributed according to a gamma distribution with unknown parameters, at least for simple models of evolution. This case is of particular importance because this assumptions are used in practice by many researchers. This work is to be submitted for journal publication[Att00a].

5 Gene Expression Array Analysis via Singular Value Decomposition

The expression levels of genes represents a substantial portion of the internal state of living cells. The genome era makes it possible to measure the expression level of every gene in the genome, providing a unique window into the fundamental mechanisms underlying all life. However, analysis of this data requires the development of techniques in addition to the standard “clustering” now routinely performed. In this work, the author assisted his mentor and Yale EEB graduate student Scott Rifkin in developing methods for exploring such data using the singular value decomposition. This work has been accepted for journal publication[SRKar].

6 Population Genetics of Unequal Crossover

In joint work with Yale EEB graduate student Max Shplak, the author derived the stationary distribution of a particular stochastic model of mutation by unequal crossover, with possible applications to early evolution. The model can be viewed as an infinite dimensional nonlinear dynamic system, the analysis of which can be completed in very few cases. The publication of this analysis awaits completion of some technical details.
References


[Att00a] Kevin Atteson. Identifiability of phylogenetic trees when rates are gamma distributed across sites. manuscript, 2000.

[Att00b] Kevin Atteson. Relaxation of the i.i.d. assumption in convex hypothesis tests. manuscript, 2000.