DIMACS Technical Report 95-48
October 1995

Proceedings of Phylogeny Workshop

held at Princeton University

February 6 - 8, 1995

Host: Simon Tavaré
University of Southern California

DIMACS is a cooperative project of Rutgers University, Princeton University, AT&T Bell Laboratories and Bellcore.

DIMACS is an NSF Science and Technology Center, funded under contract STC–91–19999; and also receives support from the New Jersey Commission on Science and Technology.

Reconstructing phylogenetic trees when sites are dependent

Simon Tavaré and Yinsuo Feng
Department of Mathematics, University of Southern California, Los Angeles, CA 90089-1113

Introduction

Classical methods for reconstructing phylogenetic trees from DNA sequence data assume that the sites have evolved independently of one another; cf. [5]. This assumption is clearly at odds with the data in many cases [1, 13]. The natural question is: Does the dependence matter? Several authors [9, 10, 11, 12] have modeled the evolution of sequences subject to constraints induced by secondary structure. Our work goes in a different direction: We have developed a class of models for sequence evolution for which the stationary distribution at a typical node in the tree can be essentially *any* probability distribution over sequence space. The familiar models with independent sites are special cases.

The new process is based on a putative DNA repair mechanism [6] that corrects potential mutations in such a way as to produce the required stationary distribution. This mechanism is related to the Markov chain Monte Carlo method [7]. For some special cases of the model, we have assessed the effect of the dependence in reconstructing trees of four species when using methods designed for independent sites. For these cases, the existing methods perform well except in the notoriously difficult case of widely different rates in different branches of the tree. Felsenstein's 'long branches attract' phenomenon [2] is further accentuated in this case.

The model

We describe a simple version of a reversible Markov model for sequence evolution that has an arbitrary stationary distribution. We assume the (aligned) sequences are s nucleotides long, and denote a typical sequence by $\boldsymbol{i} = (i_1, \ldots, i_s)$. Let $\pi(\boldsymbol{i}) > 0$ be the stationary probability of sequence $\boldsymbol{i}$. To model the evolution of the sequence along a branch of the tree, suppose that potential mutations occur at rate λ, a site being chosen uniformly to change according to a transition matrix $P = (p(i,j))$. This produces a new sequence $\boldsymbol{j}$ that differs from $\boldsymbol{i}$ in at most a single coordinate. The

probability that $\boldsymbol{i}$ changes to $\boldsymbol{j}$ we denote by $m(\boldsymbol{i},\boldsymbol{j})$. Define the Hastings ratio by $h(\boldsymbol{i},\boldsymbol{j}) = \min(1, \pi(\boldsymbol{j})m(\boldsymbol{j},\boldsymbol{i})/\pi(\boldsymbol{i})m(\boldsymbol{i},\boldsymbol{j}))$ if $\pi(\boldsymbol{i})m(\boldsymbol{i},\boldsymbol{j}) > 0$, and $= 1$ otherwise. The error correction mechanism then corrects $\boldsymbol{j}$ back to $\boldsymbol{i}$ with probability $1 - h(\boldsymbol{i},\boldsymbol{j})$, and otherwise accepts the candidate mutation $\boldsymbol{j}$. The distribution $\pi(\cdot)$ is taken to reflect the structure of the sequences in the present day sample. For example, such sequences often appear to behave like low order homogeneous Markov chains [1, 13], for which

$$\pi(\boldsymbol{i}) = \pi_0(i_1)r(i_1,i_2)\cdots r(i_{s-1},i_s),$$

for some (strictly positive) transition matrix $R = (r(i,j))$.

An example

We are particularly interested in the behavior of maximum likelihood methods [3] for reconstructing phylogenetic trees. We consider the model where P has identical rows with elements $\pi_0(\cdot)$, and $R = \alpha I + (1-\alpha)P$, for some $0 \leq \alpha \leq 1$. The marginal distribution of any particular site in a given sequence is $\pi_0(\cdot)$. The strength of the dependence along a given sequence is measured by α: when $\alpha = 0$ the sequences behave as though the sites evolve independently (and the model then reduces to the one developed in [3]), and as $\alpha \to 1$, the sequences become more and more dependent.

We simulated 500 sets of sequence data evolving along a four-taxon tree according to this model using a variety of different branch lengths (cf. [8]). For each run, we used the maximum likelihood method of [3] to estimate the underlying tree topology, and recorded how many times the correct tree was chosen. Note that the reconstruction method assumes independent sites, and its assumptions are precisely those of the model with $\alpha = 0$. This value serves to calibrate the behavior of the method for other values of the dependence parameter α. In [4] the simulation results are described in detail. Suffice it to say here that when the central branch and one branch on either side of it are short and the other two branches being relatively very long (cf. [2]), then the dependence makes the error rate even larger than when $\alpha = 0$. On the other hand, for other tree shapes the effect of dependence is to *improve* the reliability of the independent sites method for a range of values of α. Much of this phenomenon can be attributed to the way in which the error correction mechanism reduces the real rate of substitutions in the tree.

Discussion

It should be clear that the model outlined above can be generalized in many ways by altering the candidate mutation mechanism and the stationary distribution as required. Further details are given in [4], together with a variety of simulation results. It is not yet clear what general effects such complex models have on phylogeny reconstruction methods that assume independent sites. A challenging area for research is the development of computationally feasible maximum likelihood methods that take account of the dependence.

References

1. Borodovsky, M.Y., Sprizhitsky, Y., Golovanov, E., and Alexandrov, A. (1986) Statistical patterns in the primary structures of functional regions in the genome of E. coli: II Nonuniform Markov models. *Mol. Biol.*, **20**, 1024-1033.

2. Felsenstein, J. (1978) Cases in which parsimony and compatibility methods will be positively misleading. *Syst. Zool.*, **27**, 401-410.

3. Felsenstein, J. (1981) Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.*, **17**, 368-376.

4. Feng, Y. (1995) The effects of dependence among sites in phylogeny reconstruction. Unpublished Master of Science Thesis, Mathematics Department, University of Southern California.

5. Goldman, N. (1993) Statistical tests of models of DNA substitution. *J. Mol. Evol.*, **36**, 182-198.

6. Goodman, M.F., Creighton, S., Bloom, L.B., and Petruska, J. (1993) Biochemical basis of DNA replication fidelity. *Critical Reviews in Biochemistry and Molecular Biology*, **28**, 83-126.

7. Hastings W.K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97-109.

8. Huelsenbeck, J.P. and Hillis, D.M. (1993) Success of phylogenetic methods in the four-taxon case. *Syst. Biol.*, **42**, 247-264.

9. Muse, S.V. (1994) Evolutionary analysis of DNA sequences subject to constraints on secondary structure. *Genetics*, **139**, 1429-1439.

10. Rzhetsky, A. (1995) Estimating substitution rates in ribosomal RNA genes. *Genetics*, in press.

11. Schöniger, M., and von Häseler, A. (1994) A stochastic model for the evolution of autocorrelated DNA sequences. *Mol. Phylogenet. Evol.*, **3**, 240-247.

12. Tillier, E.R.M. and Collins, R.A. (1995) Neighbor joining and maximum likelihood with RNA sequences: addressing the interdependence of sites. *Mol. Biol. Evol.*, **12**, 7-15.

13. Watterson G.A. (1992) A stochastic analysis of three viral sequences. *Mol. Biol. Evol.*, **9**, 666-677.