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.

Contents

Introduction

Simon Tavaré
Department of Mathematics
University of Southern California
Los Angeles, CA 90089-1113
stavare@gnome.usc.edu

Phylogeny reconstruction remains a dynamic field with many challenging problems, both theoretical and applied. This report attempts to capture some of the recent developments, as represented by the speakers at the DIMACS workshop.

The organizing committee included Michael Bulmer, Martin Farach, Peter Smouse and Bob Vrijenhoek from Rutgers, Simon Tavaré from USC, and Tandy Warnow from the University of Pennsylvania. We are indebted to DIMACS for their help in organizing the workshop and for arranging for publication of the proceedings. We would like to thank the participants, approximately 100 phylogeny enthusiasts, for making the workshop such a success.

The work of the organizing committee, and the success of the workshop, was greatly enhanced by the support of Pat Toci and Sandy Barbu who managed all the local arrangements. Sandy Barbu converted a wide variety of write-ups into the coherent format of the final Technical Report.

Finally, we thank Walter Fitch (UC Irvine), a pioneer in the field of phylogenetics, for getting the workshop off to a great start. We gave him the difficult task of providing 'An introduction to molecular evolution for mathematicians, statisticians and computer scientists,' a challenge he met with great skill and energy!

This report can be obtained electronically at:

http://dimacs.rutgrs.edu.TechnicalReports/1995.html

as:

95-48 (abstract) Proceedings of Phylogeny Workshop

Simon Tavare, Host

Instructions on how to get the report are at:

http://dimacs.rutgers.edu/TechnicalReports/obtain.html

Sorting by Reversals: Genome Rearrangements in Plant Organelles[1]

Vineet Bafna
DIMACS Center
P. O. Box 1179
Piscataway, NJ 08855
`bafna@dimacs.rutgers.edu`

Abstract Sequence comparison in computational molecular biology is a powerful tool for deriving evolutionary and functional relationships between genes. However, classical alignment algorithms handle only local mutations (i.e. insertions, deletions and substitutions of nucleotides) and ignore global rearrangements (i.e. inversions and transpositions of long fragments). As a result, applications of sequence alignment in analyzing highly rearranged genomes (i.e. herpes viruses or plant mitochondrial DNA) are very limited and may lead to contradictions in molecular phylogeny studies since different genes give rise to different evolutionary trees. The paper describes the problem of *genome* comparison versus classical *gene* comparison and presents algorithms to analyze rearrangements in genomes evolving by inversions. In the simplest form the problem corresponds to *sorting by reversals*, i.e. sorting of an array using reversals of arbitrary fragments. We describe algorithms to analyze genomes evolving by *inversions* and discuss applications of these algorithms in molecular evolution.

Introduction

Sequence alignment is often the first step in molecular evolution studies. However, in many cases sequence alignment is very unreliable thus making further evolutionary tree reconstruction almost impossible. For example, the similarity between many genes in herpes viruses is so low that it is frequently indistinguishable from the background noise [1, 8]. As a result, the classical methods of *sequence comparison* frequently lead to ambiguous conclusions for such highly diverged genomes and alternative methods are sought. Since it is often found that the order of genes is much more conserved than the DNA sequence [7] an approach based on comparison of *gene orders* [26, 27] versus traditional comparison of *DNA sequences* seems to be a method of choice for many "hard-to-analyze" genomes.

Surprisingly enough, similar problems exist for such conserved genomes as plant mitochondrial DNA(mtDNA). The point mutation rate in plant mtDNA is estimated to be 100 times slower than in animal mtDNA and many genes are $99\% - 99.9\%$ identical in related species [23]. However, although there is little change in the DNA sequence of the plant mitochondrial genes from one species to another, there is rapid and extensive change in gene arrangements.

Studies of gene arrangements in plant organelles were pioneered by Palmer and co-authors who compared gene orders to analyze plant molecular evolution. Every such study involves solving a

[1]This research was carried out while the author was a student at Penn State. Partially supported by NSF grant CCR-9308567 and NIH grant 1R01 HG00987 to Pavel A. Pevzner, and by DIMACS (Center for Discrete Mathematics and Theoretical Computer Science), a National Science Foundation Science and Technology Center under NSF contract STC-8809648.

combinatorial "puzzle" to find a shortest series of genome rearrangement to transform one genome into another. Fig. 0.1 presents the order of conserved blocks in *Tobacco* and *Lobelia fervens* chloroplast DNA(cpDNA) with a hypothetical sequence of rearrangement events [17] during evolution of *Lobelia fervens* from a tobacco-like ancestral genome.

In the cases of genomes consisting of small number of "conserved blocks" Palmer and co-authors were able to find the most parsimonous scenarios for rearrangements. However, even for genomes consisting of more than 10 blocks, exhaustive search over all potential solutions is intractable. On the other hand, similar problems are rapidly arising in many different areas, including evolution of plant cpDNA [13, 25] plant mtDNA [6, 23], animal mtDNA [12, 27], virology [10, 18], *Drosophila* genetics [5, 31] and comparative physical mapping [19]. The goal of this paper is to describe an approach to automating such analysis by estimating the *inversion distance* between two genomes.

Genomes evolve by inversions and transpositions as well as more simple operations of deletion, insertion and duplication of fragments. However, inversions seem to be the dominant form of rearrangement, and in fact some genomes (for example, many plant mitochondrial DNA) are believed to rearrange almost solely by inversions [23] and less frequently by duplications. A combinatorial problem of *sorting by reversals* corresponding to genome rearrangements by inversions was introduced by Kececiouglu and Sankoff, 1993 [14] who suggested the first "performance guarantee" algorithm to analyze genome rearrangements. The problem was further studied by Bafna and Pevzner, 1993 [2] who improved the Kececiouglu and Sankoff algorithm and by Kececiouglu and Sankoff, 1994 [15] who analyzed the performance of sorting by reversals on simulated data. More recently, Hannenhalli and Pevzner, 1995 [11] devised a polynomial–time algorithm for sorting *signed* fragments by reversals.

In the very first computational studies of genome rearrangements Watterson et al., 1982 [30] introduced the notion of *breakpoint* (disruption of gene order) and noticed some correlations between the reversal distance and the number of breakpoints (in fact, Sturtevant and Dobzhansky, 1936 [29] implicitly discussed these correlations 60 years ago!). However, the estimate of reversal distance in terms of breakpoints is very rough and does not provide an accurate bound. The important result of Bafna and Pevzner, 1993 [2] is that there is another "hidden" parameter which estimates reversal distance with much greater accuracy. In the following discussion, we define the model formally and present the results of computational experiments in which we analyze previously postulated scenarios of rearrangements in plant organelles.

Sorting By Reversals

In the problem we consider, the order of genes in two organisms is represented by permutations $\pi = (\pi_1\pi_2\dots\pi_n)$ and $\sigma = (\sigma_1\sigma_2\dots\sigma_n)$. A *reversal* $\rho(i,j)$ of an interval $[i,j]$ is the transformation

$$\rho: \quad (\pi_1\dots\pi_{i-1}\pi_{\mathbf{i}}\pi_{\mathbf{i+1}}\cdots\pi_{\mathbf{j-1}}\pi_{\mathbf{j}}\pi_{j+1}\dots\pi_n)$$

$$\rightarrow (\pi_1\dots\pi_{i-1}\pi_{\mathbf{j}}\pi_{\mathbf{j-1}}\cdots\pi_{\mathbf{i+1}}\pi_{\mathbf{i}}\pi_{j+1}\dots\pi_n)$$

which has the effect of reversing genes $\pi_{\mathbf{i}}, \pi_{\mathbf{i+1}}, \dots \pi_{\mathbf{j-1}}, \pi_{\mathbf{j}}$. Given permutations π and σ, the *reversal distance problem* is to find a shortest series of reversals transforming genome π into genome σ, i.e. to find a series of reversals $\rho_1, \rho_2, \dots, \rho_t$ such that $\pi \cdot \rho_1 \cdot \rho_2 \cdots \rho_t = \sigma$ and t is minimum. We call t the *reversal distance* between π and σ. Without loss of generality we can assume that σ is the *identity* permutation $\imath = (1, 2, \ldots, n)$ and in the following we consider a problem of transforming π into $\imath$. *Sorting π by reversals* is the problem of finding reversal distance $d(\pi)$, between π and $\imath$. As an example, Fig. 0.1, if considered from the bottom to the top presents sorting of *Lobelia fervens* permutation $\pi = 71245368$ by 5 reversals $\rho(6,7), \rho(4,7), \rho(5,7), \rho(2,7), \rho(1,7)$.

Let $i \sim j$ if $i = j + 1$ or $i = j - 1$. Extend a permutation $\pi = \pi_1 \pi_2 \ldots \pi_n$ by adding $\pi_0 = 0$ and $\pi_{n+1} = n + 1$. We call a pair of consecutive elements π_i and π_{i+1}, $0 \le i \le n$, of π an *adjacency* if $\pi_i \sim \pi_{i+1}$, and a *breakpoint* if $\pi_i \not\sim \pi_{i+1}$. Denote the number of adjacencies and breakpoints in π as $a(\pi)$ and $b(\pi)$ correspondingly. Clearly $a(\pi) + b(\pi) = n + 1$. For example, *Lobelia fervens* cpDNA corresponds to the permutation $\pi = 71245368$. We let $\pi_0 = 0$ and $\pi_9 = 9$. Then $(1,2), (4,5)$ and $(8,9)$ are the three adjacencies, while the remaining pairs $(0,7), (7,1), (2,4), (5,3), (3,6)$ and $(6,8)$ are breakpoints. For the *Lobelia fervens* permutation, $a(\pi) + b(\pi) = 3 + 6 = 9$.

As the identity permutation $\imath$ has no breakpoints, sorting by reversal corresponds to eliminating breakpoints. An observation that every reversal can eliminate *at most* 2 breakpoints immediately implies that

$$d(\pi) \ge \frac{b(\pi)}{2} = \frac{n + 1 - a(\pi)}{2} \tag{0.1}$$

The above estimate is not accurate. In order to improve the bound on $d(\pi)$, define an edge-coloured *graph* $G(\pi)$ with $n+2$ vertices $0, 1, \ldots, n, n+1$. Join vertices i and j by a *black* edge if i and j occupy neighboring positions in π. We join vertices i and j by a *gray* edge if $i \sim j$. It can be shown that the edges of $G(\pi)$ can be decomposed into a collection of edge-disjoint *alternating cycles* (i.e. cycles whose edges change colors in every vertex). Let $c(\pi)$ be the number of cycles in a maximal cycle decomposition. Bafna and Pevzner, 1993 [2] show that

$$d(\pi) \ge n + 1 - c(\pi) \tag{0.2}$$

The bound (0.2) in terms of maximal cycle decomposition is much tighter than the bound (0.1) in terms of the number of breakpoints/adjacencies. In fact, for all biological examples we studied $d(\pi) = n + 1 - c(\pi)$ thus reducing reversal distance problem to maximal cycle decomposition problem. Further, in those examples, $c(\pi)$ is very small implying that bound (0.1) is considerably weaker than bound (0.2). Based on the notion of "maximal cycle decomposition", Bafna and Pevzner, 1993 [2] developed a "performance guarantee" algorithm for sorting by reversals which constructed (provably) shortest evolutionary scenarios for all biological examples we studied.

Table 0.1 presents the results of our program on a series of experimental data by Palmer et al. [13, 17, 20, 21, 22, 23, 24, 28]. For example, Palmer and Herbon, 1988 [23] postulated 11 rearrangements for the comparison B. nigra vs. B. campestris while our program produces a sequence of nine inversions which transform one sequence into the other. At the same time it also computes a lower bound of 9, thereby *proving* that our program found the most parsimonious sequence of reversals. More recently, Hannenhalli and Pevzner, 1995 [11] investigated the structure of cycle graphs intensively, and discovered parameters that give an exact formula for the reversal distance. This naturally leads to a polynomial–time algorithm for computing the reversal distance for signed permutations.

Conclusion

This paper is a condensed version of [3], which also contains some other examples where these concepts have been applied, such as comparative human-mouse mapping. Technical details, including the algorithm are presented in [2]. The concept of cycle graphs of permutation turns out to be useful in estimating distance for other form of rearrangements as well, including transpositions [4] and translocations [9, 16]. However, it is not clear if similar techniques can be used in the case that a large number of duplications have occured in the genomes.

While many algorithmic problems remain, it is our hope that this research will provide biologists with a useful tool for making evolutionary hypotheses based on rearrangements, and will stimulate further research in the area.

Bibliography

[1] J. Albrecht, J. Nicholas, D. Biller, K. R. Cameron, B. Biesinger, C. Newman, S. Wittmann, M. A. Craxton, H. Coleman, B. Fleckenstein, and R. W. Honess. Primary structure of the herpesvirus saimiri genome. *Journal of Virology*, 66:5047–5058, 1992.

[2] V. Bafna and P. Pevzner. Genome rearrangements and sorting by reversals. In *34th IEEE Symp. on Foundations of Computer Science*, pages 148–157, 1993.

[3] V. Bafna and P. Pevzner. Sorting by reversals: Genome rearrangements in plant organelles and evolutionary history of X chromosome. *Molecular Biology and Evolution*, 12:239–246, 1995.

[4] V. Bafna and P. Pevzner. Sorting permutations by transpositions. In *The sixth annual ACM-SIAM symposium on discrete algorithms*, pages 614–623, 1995.

[5] H. L. Carson. Chromosomal sequences and interisland colonization in Hawaiian Drosphila. *Genetics*, 103:465–482, 1983.

[6] C. Fauron and M. Havlik. The maize mitochondrial genome of the normal type and the cytoplasmic male sterile type T have very different organization. *Current Genetics*, 15:149–154, 1989.

[7] N. Franklin. Conservation of genome form but not sequence in the transcription antitermination determinants of bacteriophages $\lambda, \phi 21$ and $P22$. *Journal of Molecular Evolution*, 181:75–84, 1985.

[8] A. M. Griffin and M. E. G. Boursnell. Analysis of the nucleotide sequence of DNA from the region of the thymidine kinase gene of infectious laryngotracheitis virus; potential evolutionary relationships between the herpesvirus subfamilies. *Journal of General Virology*, 71:841–850, 1990.

[9] S. Hannenhalli. A polynomial time Algorithm for computing translocation distance between genomes. In *Combinatorial Pattern Matching(to appear)*, 1995.

[10] S. Hannenhalli, C. Chappey, E. Koonin, and P. Pevzner. Scenarios for genome rearrangements: Herpesvirus evolution as a test case. In *Proc. of 3rd Intl. Conference on Bioinformatics and Complex Genome Analysis*, 1994 (to appear).

[11] S. Hannenhalli and P. Pevzner, Transforming Cabbage into Turnip, In *27th Annual ACM symposium on Theory of Computing (to appear)*, 1995.

[12] R. J. Hoffmann, J. L. Boore, and W. M. Brown. A novel mitochondrial genome organization for the blue mussel, *Mytilus edulis*. *Genetics*, 131:397–412, 1992.

[13] S. B. Hoot and J. D. Palmer. Structural rearrangements including parallel inversions within the chloroplast genome of anemone and related genera. *Journal of Molecular Evolution*, 38:274–281, 1994.

[14] J. Kececioglu and D. Sankoff. Exact and approximation algorithms for the inversion distance between two permutations. In *Proc. of 4th Ann. Symp. on Combinatorial Pattern Matching*, Lecture Notes in Computer Science 684, pages 87–105. Springer Verlag, 1993. (Extended version appeared in Algorithmica 13: 180-210, 1995).

[15] J. Kececioglu and D. Sankoff. Efficient bounds for oriented chromosome inversion distance. In *Lecture notes in computer science*, volume 807, pages 307–325, 1994.

[16] J. D. Kececioglu and R. Ravi. Of mice and men: Evolutionary distances between genomes under translocations. *The sixth annual ACM-SIAM symposium on discrete algorithms*, 1995.

[17] E. B. Knox, S. R. Downie, and J. D. Palmer. Chloroplast genome rearrangements and evolution of giant lobelias from herbaceous ancestors. *Mol. Biol. Evol.*, 10:414–430, 1993.

[18] E. V. Koonin and V. V. Dolja. Evolution and taxonomy of positive-strand RNA viruses: implications of comparative analysis of amino acid sequences. *Crit. Rev. Biochem. Molec. Biol.*, 28:375–430, 1993.

[19] M. F. Lyon. X-Chromosome Inactivation and the Location and Expression of X-linked Genes. *Am. J. Hum. Genet.*, 42:008–016, 1988.

[20] C. A. Makaroff and J. D. Palmer. Mitochondrial DNA rearrangements and transcriptional alterations in the male sterile cytoplasm of Ogura radish. *Molecular Cellular Biology*, 8:1474–1480, 1988.

[21] B. G. Milligan, J. N. Hampton, and J. D. Palmer. Dispersed repeats and structural reorganization in subclover chloroplast DNA. *Molecular Biology Evolution*, 6:355–368, 1989.

[22] J. Palmer and L. Herbon. Unicircular structure of the Brassica hirta mitochondrial genome. *Current Genetics*, 11:565–570, 1987.

[23] J. D. Palmer and L. A. Herbon. Plant mitochondrial DNA evolves rapidly in structure, but slowly in sequence. *Journal of Molecular Evolution*, 27:87–97, 1988.

[24] J. D. Palmer, B. Osorio, and W. F. Thompson. Evolutionary significance of inversions in legume chloroplast DNAs. *Current Genetics*, 14:65–74, 1988.

[25] L. A. Raubeson and R. K. Jansen. Chloroplast DNA evidence on the ancient evolutionary split in vascular land plants. *Science*, 255:1697–1699, 1992.

[26] D. Sankoff, R. Cedergren, and Y. Abel. Genomic divergence through gene rearrangement. In *Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences*, chapter 26, pages 428–438. Academic Press, 1990.

[27] D. Sankoff, G. Leduc, N. Antoine, B. Paquin, B. F. Lang, and R. Cedergren. Gene order comparisons for phylogenetic inference: Evolution of the mitochondrial genome. *Proc. Natl. Acad. Sci. USA*, 89:6575–6579, 1992.

[28] S. H. Strauss, J. D. Palmer, G. T. Howe, and A. H. Doerksen. Chloroplast genomes of two conifers lack a large inverted repeat and are extensively rearranged. *Proceedings of National Academy of Sciences, U.S.A*, 85:3598–3902, 1988.

[29] A. H. Sturtevant and T.Dobzhansky. Inversions in the third chromosome of wild races of *drosophila pseudoobscura*, and their use in the study of the history of the species. *Proc. Nat. Acad. Sci.*, 22:448–450, 1936.

[30] G. A. Watterson, W. J. Ewens, T. E. Hall, and A. Morgan. The chromosome inversion problem. *Journal of Theoretical Biology*, 99:1–7, 1982.

[31] J. Whiting, M. Pliley, J. Farmer, and D. Jeffery. In situ hybridization analysis of chromosomal homologies in *Drosophila melanogaster* and *Drosophila virilis*. *Genetics*, 122:99–109, 1989.

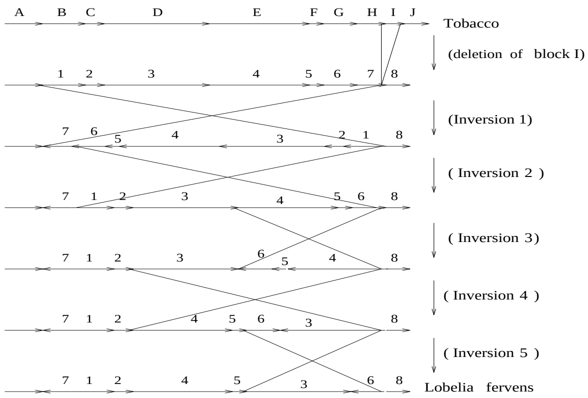

Figure 0.1:

Source	DNA	n	Postulated Distance	Lower Bound	Upper Bound	Reference
B. nigra vs. B. campestris	mt	12	11	9	9	Palmer and Herbon,1988
B.hirta vs. B.campestris	mt	12	10	10	10	Palmer and Herbon,1987
Ogura vs. normal radish	mt	12	10	8	8	Makaroff and Palmer, 1988
R. sativa vs. B. campestris	mt	15	14	13	13	Palmer and Herbon,1988
radiata pine vs. petunia-mung	cp	5	4	4	4	Strauss et al., 1988
Tobacco vs. Lobelia fervens	cp	7	5	5	5	Knox et al, 1993
Anemone vs. Clematis	cp	8	6	6	6	Hoot and Palmer, 1994
subclover vs. alfafa	cp	10	8	8	8	Milligan et al.,1989
pea vs. soybean	cp	13	8	8	8	Palmer et al.,1988

Table 0.1: Reversal distances for various cpDNA and mtDNA. Each row compares a pair of plant organells DNA consisting of n ($5 \leq n \leq 15$) blocks, as described in the third column. The fourth column describes the distance postulated in previous studies, while the fifth and sixth columns describe the lower and upper bound produced by the algorithm.

The Perils of Ignoring Rate Heterogeneity: Inconsistency of Topology Estimation Methods When Substitution Rates Vary Across Characters

Joseph T. Chang
Yale University Statistics Department
Box 208290 Yale Station
New Haven, CT 06520-8290
chang@brandy.stat.yale.edu

The following are excerpts from the introduction and references of a full paper that is available from the author.

1 Introduction

Markov models of character changes have been well studied and extensively used in analyzing character data such as DNA sequences. A simplifying assumption that has typically been made is that each character evolves according to precisely the same Markov model. In particular, this "homogeneity assumption" incorporates the assertion that, at any given point in the evolutionary tree, all of the different characters evolve at the same rate. Unfortunately, this assumption is known to be very unrealistic. At one extreme are "invariable" characters that are so essential to the functioning of the organism that no change in the character could survive. The rate of evolutionary change of such characters could be said to be zero. On the other hand, there are characters that can experience certain changes without modifying any amino acid sequence produced. Such characters, which may include characters in noncoding regions as well as the third codon positions in coding regions, are free to change at a relatively rapid rate.

A fundamental statistical criterion in judging the adequacy of an inference method is that of consistency. A tree topology reconstruction method is consistent under a certain model of evolution if, when given more and more characters generated according to the assumed model, the reconstructed topology is certain to converge to the correct topology. Some distance methods and maximum likelihood methods for topology reconstruction have been shown to be consistent under Markov models having homogeneous evolution across characters. The purpose of this paper is to present examples that show that these methods may fail to be consistent even under very simple models that allow for rate variation across characters.

Markov models have a significant history in phylogeny reconstruction, and they continue to play an important role. The articles [7], [8], and [2] are recommended for an overview of Markov models and other methods in phylogeny reconstruction. The Markov assumption, roughly speaking, asserts that character changes are independent in different branches of the tree. This has been viewed as an appealing compromise, giving theoretical and computational tractability while remaining relatively palatable scientifically.

However, the homogeneity assumption that each character evolves according to the same Markov model is worrisome. Beginning with the work of Fitch and Margoliash [9], a substantial literature has developed evidence for rate variation across characters as well as statistical approaches for taking such rate variation into account. A recent and convincing statement of evidence is provided by Shoemaker and Fitch [14], whose study may be consulted for further references. A variety of statistical models and methods for incorporating rate heterogeneity have been proposed recently; these include the work of Churchill, Navidi, and von Haeseler [5]; Jin and Nei [10]; Kelly and Rice [11]; Lundstrom, Tavaré, and Ward [13]; Sidow, Nguyen, and Speed [15]; and Yang [18].

It is natural to ask how necessary it is to go to the effort of developing and using such techniques. If we use standard methods developed for models that assume homogeneous evolution across characters, how can we be led astray when that assumption is false? One form of deception that has been established is a systematic bias in the estimation of branch lengths; see Fitch [9] and Kelly and Rice [11]. The recent studies of Tateno, Takezaki, and Nei [16] and Kuhner and Felsenstein [12]. have used simulation to investigate the performance of phylogenetic methods when substitution rates vary across sites. These are extensive studies that make progress toward quantifying the extent to which rate heterogeneity degrades the accuracy of several methods.

In this paper we focus on the question: Can we consistently recover the tree topology? We present examples that show that distance and maximum likelihood methods that are consistent for the homogeneous Markov model can fail to be consistent when rates of evolution vary across characters. Our aim is to present simple examples that establish the existence of the phenomenon and give some understanding of it without extensive analytic calculations or computer simulations. A more extensive analysis of issues such as conditions under which the phenomenon occurs and how common it might be will be left for later work.

The distance method of Barry and Hartigan [1] is consistent for the homogeneous Markov model; see Chang and Hartigan [4]. Cavender and Felsenstein [3] gave a numerical example that showed that their closely related method may be inconsistent when different characters evolve at different rates. This paper shows that in fact the presence of invariable characters is enough to admit examples for which the methods of Barry and Hartigan [1] and Cavender and Felsenstein [3] become inconsistent.

Not surprisingly, the maximum likelihood method of Felsenstein [6] is consistent for the homogeneous Markov model. One would anticipate this since, as shown by Wald [17], a maximum likelihood method developed for a given family of distributions is, subject to mild regularity conditions, consistent when used for that family. Here we show by example that the maximum likelihood method of [6] may be inconsistent when invariable characters or more general forms of rate heterogeneity are present.

In a sense, we are testing these methods "unfairly" by applying them to classes of models that are more general than those for which the methods were designed. Thus, it is not altogether surprising that the methods may become inconsistent under these circumstances. On the other hand, it does not seem obvious that they should do so either. In particular, all we ask of the methods is that they identify the tree topology, which is a rather minimal request in terms of detailed information. One might reasonably harbor some hope for the methods to be able to do at least this much. After all, even though we allow rates and branch lengths to vary across characters, we assume that all characters evolve according to one common tree topology. If each character contributes some information about the tree topology that they all share, it seems reasonable to hope that as more and more characters accumulate, the data could reveal to us the common topology with more and more certainty. The examples presented in this paper show that the distance and maximum likelihood methods under consideration cannot reliably extract that information from the data.

Bibliography

[1] D. Barry and J. A. Hartigan. Asynchronous distance between homologous DNA sequences. *Biometrics*, 43:261–276, 1987.

[2] D. Barry and J. A. Hartigan. Statistical analysis of hominoid molecular evolution. *Statistical Science*, 2:191–210, 1987.

[3] J. A. Cavender and J. Felsenstein. Invariants of phylogenies in a simple case with discrete states. *Journal of Classification*, 4:57–71, 1987.

[4] J. T. Chang and J. A. Hartigan. Reconstruction of evolutionary trees from pairwise distributions on current species. In *Computer Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, pages 254–257, 1991.

[5] G. A. Churchill, A. von Haeseler, and W. C. Navidi. Sample size for a phylogenetic inference. *Mol. Biol. Evol.*, 9:753–769, 1992.

[6] J. Felsenstein. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Syst. Zool.*, 22:240–249, 1973.

[7] J. Felsenstein. Statistical inference of phylogenies. *J. R. Statist. Soc. A*, 146:246–272, 1983.

[8] J. Felsenstein. Phylogenies from molecular sequences: Inference and reliability. *Annu. Rev. Genet.*, 22:521–565, 1988.

[9] W. M. Fitch and E. Margoliash. A method for estimating the number of invariant amino acid coding positions in a gene using cytochrome c as a model case. *Biochem. Genet.*, 1:65–71, 1967.

[10] L. Jin and M. Nei. Limitations of the evolutionary parsimony method of phylogenetic analysis. *Mol. Biol. Evol.*, 7:82–102, 1990.

[11] C. Kelly and J. Rice. Modeling molecular evolution: A heterogeneous rate analysis, 1994. To appear in *Math. Biosci.*

[12] M. K. Kuhner and J. Felsenstein. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol. Biol. Evol.*, 11:459–468, 1994.

[13] R. Lundstrom, S. Tavaré, and R. H. Ward. Modeling the evolution of the human mitochondrial genome. *Math. Biosci.*, 112:319–335, 1992.

[14] J. S. Shoemaker and W. M. Fitch. Evidence from nuclear sequences that invariable sites should be considered when sequence divergence is calculated. *Mol. Biol. Evol.*, 6:270–289, 1989.

[15] A. Sidow, T. Nguyen, and T. P. Speed. Estimating the fraction of invariable codons with a capture-recapture method. *J. Mol. Evol.*, 35:253–260, 1992.

[16] Y. Tateno, N. Takezaki, and M. Nei. Relative efficiencies of the maximum-likelihood, neighbor-joining, and maximum-parsimony methods when substitution rate varies with site. *Mol. Biol. Evol.*, 11:261–277, 1994.

[17] A. Wald. Note on the consistency of the maximum likelihood estimate. *Ann. Math. Statist.*, 20:595–601, 1949.

[18] Z. Yang. Maximum likelihood estimation of phylogeny from dna sequences when substitution rates vary over sites. *Mol. Biol. Evol.*, 10:1396–1401, 1993.

Phylogenetic Networks

Gary A. Churchill
Biometrics Unit, 337 Warren Hall
Cornell University
Ithaca, NY 14853-7801
gary@amanita.biom.cornell.edu

1 Introduction

It is a generally accepted principle in molecular evolution that relationships among gene sequences sampled from a group of organisms can be represented as a bifurcating tree. Indeed phylogenetic trees almost always provide an adequate representation for the history of sequences sampled from moderately divergent species. However there are some situations for which a more general representation is required. In this note, network phylogenies are proposed as a graphical summary of relationships among a set of sequences.

Inference procedures for network phylogenies have been described by von Haeseler and Churchill (1993) for the case of binary character states and a symmetric model of character change. Their model, summarized below, is based on a multinomial likelihood for the observed character states configurations. A distance based method for inferring networks has been described by Bandelt and Dress (1990) and applied to viral sequences by Dopazo et al. (1990).

2 The Network Model

Suppose we have a set of K binary sequences each of length N with homologous sites aligned in columns. Thus the data are a matrix with elements $s_{ij} \in \{0,1\}$ where i is the species index and j is the site index. Each site j forms a K–dimensional vector of zeroes and ones, a binary character–state configuration. We assume that the evolutionary process acts independently and identically at each site and that changes between the states zero and one are symmetric. Thus two configuration are equivalent if one can be obtained from the other by reversing the labeling of the zero and one states.

A split is a bipartition of the K species into two disjoint sets and can be represented by the set S, such that $1 \in S$, and the set $S^c = \{1,\ldots,K\} \setminus S$. Each of the 2^{K-1} non–equivalent configurations corresponds to a split on the species. A network model $M = (\mathcal{S}, \theta)$ is defined by a set of splits and a parameter vector $\theta = \{\theta_S | S \in \mathcal{S}\}$, where θ_S is the probability that, at a given site, an odd number of substitutions separate the sequences in S from those in S^c. The parameter vector θ is analogous to Hendy and Penny's (1989) $\mathbf{p}$ vector, the probability of character changes along an edge of a tree.

A network model is a tree if and only if all splits in $\mathcal{S}$ are pairwise compatible. Splits S_i and S_j are pairwise compatible if there exists a $B_i \in \{S_i, S_i^c\}$ and a $B_j \in \{S_j, S_j^c\}$ with $B_i \cap B_j = \emptyset$ (Buneman (1971)). For example, consider the model M_1, with $\mathcal{S} = \{S_1, S_2, S_3, S_4\}$, where $S_1 = \{1\}, S_2 = \{1,3,4\}, S_3 = \{1,2,4\}, S_4 = \{1,2,3\}$,. This is the familiar four species star phylogeny. A new model M_2 is created by adding the split $S_5 = \{1,2\}$. This is the four species tree that pairs species 1 and 2 versus species 3 and 4. If we now obtain a third model M_3 by adding the split $S_6 = \{1,3\}$ the result is a network. The splits S_5 and S_6 are not compatible.

3 Applications

It is important to make a distinction between networks as a represenation of statistical uncertainty in the inference of a true tree-like phylogeny and networks as representation of an underlying biological phenomenon.

Statistical Networks. It is often the case that there is insufficient information in a sequence set to determine a unique and fully bifurcating tree with certainty. The collection of trees consistent with a dataset, in the sense that each cannot be rejected by a hypothesis test, consititutes a confidence interval (Cavender 1978, Felsenstein 1988). Unfortunately such collections may be large and hard to visualize. A set of trees can be combined by taking the union of all splits in the trees and forming the corresponding network. Conversely, a network can be decomposed into a set of trees each composed of a maximal set of compatable splits. A minimal network that contains the set of trees compatable with the data is a confidence set with a compact representation as a set of splits. The network confidence set is more precise than than a multifurcating tree.

Biological Networks. Not all relationships among sets of sequences are treelike. First, there are increasing numbers of examples of horizontal transfers of genetic information across species boundaries (reviewed by Syvanen, 1994) including evidence for ancient fusions of whole genomes (Golding and Gupta, 1995). Second, the events that take place at the bifurcation in a phylogenetic tree can be very complex. Population subdivision with some gene flow and/or hybridization can result in gene trees that do not agree with species trees. Third, when we are looking at sequence derived from within a species, different segments of a sequence are likely to have unique histories due to recombination. Ideally we would like to identify the recombination breakpoints and assign a tree–like history to each segment. However, the amount of sequence variation may limit our ability to make such inferences. This list of situations that would give rise to network relationships is not exhaustive.

4 Alternative Models

The network model allows multiple events on non–compatable splits to act on the same site in a sequence. This approach contradicts the notion that each site should have a tree–like history. Alternative models could be developed if we insist on tree–like evolution at each site but wish to allow the tree to vary from site to site.

Mixture Model. In a mixture model, each site in the aligned sequence set is assumed to have an unknown phylogenetic tree that is drawn from a finite set $\mathcal{T} = \{\tau_1, \ldots, \tau_k\}$ of possible trees. The tree $\tau_{k(i)}$ corresponding to site i is drawn independently with probability $\alpha_{k(i)}$ from the set $\mathcal{T}$. If the set of possible trees is known, the resulting likelihood is a multinomial mixture and can be maximized using standard methods for mixture estimation. Estimation of the set $\mathcal{T}$ represents a much more challenging problem. A reasonable starting point might be obtained by fitting a network model and taking $\mathcal{T}$ to be the set of subtrees of the network.

Block Models. When a small number of recombination or gene conversion events have occurred in the history of the sequence set, blocks of adjacent sites will share identical tree–like histories. Sawyer's

(1989) test for gene conversion is based on detection of runs of compatable configurations. The problem of reconstructing the trees and the recombination break points has been studied by Hein (1990) using parsimony methods. The model described by Hein is implicitly a hidden Markov model. Related applications of hidden Markov models to sequence data have been described by Churchill (1989) and by Felsenstein and Churchill (1995). The hidden Markov model assumes that topology changes along the aligned sequence set form a Markov chain. Unfortunately, segments generated by standard models of recombination are not Markovian. Robust properties of the hidden Markov model suggest that it may provide a tractable approximation to more realistic block models.

Rate Heterogeneity. Variation in substitution rates between sites can induce an apparent network structure in sequence data. When rate heterogeneity is present, estimated rates assuming a homogeneous model will be an average of the actual rates. Sites with a high rate of change are likely to show an apparent excess of parallel changes in independent lineages. The introduction of invariable sites (Shoemaker and Fitch, 1989) into the network model will reduce the number of statistically significant non–compatable splits. More general models of rate heterogeneity should be considered in conjuction with a network phylogeny. In principle, it should be possible to distinguish rate heterogenity from network structure as heterogeneity should inflate the significance of all splits equally.

5 Summary

Network models extend the classical model of phylogeny (e.g. Felsenstein 1981) by allowing for a more general dependence structure among gene sequences. Networks can be interpreted statistically as mixtures of tree–like models and can be useful in modeling a variety of evolutionary scenarios. Further work is needed to develop mixture models, block models, models of rate heterogeneity and associated methods for statistical analysis of network relationships among sequences.

Bibliography

[1] Bandelt HJ, Dress AWM (1990) A canonical decomposition theory for metrics on a finite set. Preprint 90–032, SFB 343, Universität Bielefeld.

[2] Buneman P (1971) The recovery of trees from measures of dissimilarity. In: Hodson FR, Kendall DG, Tantu P (eds) Mathematics in the archaeological and historical science. Proc. of the Anglo–Romanian–Conference 1970. Univ. Press Edinburgh. pp 387–395.

[3] Cavender J (1978) Taxonomy with confidence. Math. Biosc. 40:271–280.

[4] Churchill GA (1989) Stochastics Models for heterogeneous DNA sequences Bull. Math. Biol.

[5] Dopazo J, Dress A, von Haeseler A (1990) Split decomposition: A new technique to analyze viral evolution. Preprint 90–037, Sonderforschungsbereich 343 Diskrete Strukturen in der Mathematik. Universität Bielefeld.

[6] Felsenstein J (1981) Evolutionary trees from DNA sequences: A maximum likelihood approach. J. Mol. Evol. 17:368–376.

[7] Felsenstein J (1988) Phylogenies from molecular sequences: Inference and reliability. Annu. Rev. Genet. 22:521–565.

[8] Felsenstein J, Churchill GA (1995) Mol. Biol. Evol. (submitted)

[9] Golding B, Gupta (1995) Mol. Biol. Evol.

[10] von Haeseler A, Churchill GA (1993) Network models for sequence evolution. J. Mol. Evol. 37:77–85.

[11] Hein J (1990) reconstructing evolution of sequences subject to recombination using parsimony. Math. Biosci. 98:185–200.

[12] Hendy MD, Penny D (1989) A framework for the quantitative study of evolutionary trees. Syst. Zool. 38:297–309.

[13] Sawyer S (1989) Statistical test for detecting gene conversion. Mol. Biol. Evol. 6:526–538.

[14] Shoemaker JS, Fitch WM (1989) Evidence from nuclear sequences that invariable sites should be considered when sequence divergence is calculated. Mol. Biol. Evol. 6:270–289.

[15] Syvanen M (1994) Horizontal gene transfer: Evidence and possible consequences. Ann. Rev. Genet. 28:237–261.

Intraspecific Cladogram Estimation

Keith A. Crandall
Department of Zoology
University of Texas
Austin, TX 78712-1064
crandall@phylo.zo.utexas.edu

Recent advances in population genetic theory, especially coalescent theory, coupled with an explosion of molecular techniques have allowed detailed phylogenetic information at the population level. Such genealogical relationships are termed gene trees, allele trees, or haplotype trees (I use haplotype throughout this paper) in which different haplotypes are the operational taxonomic units. These haplotype trees can be used to test numerous hypotheses concerning a variety of biological phenomena, from detecting natural selection to measuring migration rates [4]. Yet the use of haplotype trees in testing various hypotheses depends upon the accuracy in reconstructing the true history. Traditional methods, such as maximum parsimony, were developed to estimate relationships of higher order taxa (e.g. species, genera, etc.). These methods make certain assumptions that are generally valid in the higher order case, but become invalid, or at least suspect, at the population genetic level. These assumptions include the lack of ancestral haplotypes in the population, no recombination within the gene region being examined, and a bifurcating topology of relationships. Furthermore, these data typically have lower levels of variation, thus providing fewer characters for phylogenetic analysis. The method presented in this paper does not make these assumptions and has its statistical power when haplotypes share many sites and differ by few. Additionally, it provides a robust hypothesis testing framework within which one can test a variety of population genetic hypotheses.

Cladogram Estimation.

The procedure of Templeton et al. [15] estimates the set of cladograms that portray connections among haplotypes where these connections have a high cumulative probability (≥ 0.95) of being true; thereby constructing a probable set of alternative cladograms. This estimation algorithm is based on a parsimony criterion with a statistical procedure to evaluate the limits of the parsimony assumption. Under consideration by the method is the number of mutational events causing a single nucleotide difference between two haplotypes with a difference caused by a single mutational event being the parsimonious state and a difference cause by multiple mutational events being the nonparsimonious state. Thus the term parsimony refers to the minimum number of mutations separating two individual sequences rather than a global minimum tree length based on shared derived characters.

Cladogram Probabilities.

Crandall and Templeton [3] have shown that results from coalescent theory are valuable in refining probabilities for alternative cladograms estimated by the method above. Coalescent theory predicts that alleles of recent origin should be low in frequency and localized on the tips of cladograms (Donnelly

& Tavaré [6]). Crandall and Templeton [3] testing this predication using a collection of restriction site data sets from the literature. This hypothesis was supported, i.e. haplotypes of low frequency preferentially occurred on the tips of cladograms while those of high frequency occurred in the interior. Furthermore, their empirical based relationship of haplotype type frequency to tip probability can be used to refine probabilities of alternative cladograms [3]; Crandall et al. [5]). A second relationship suggested by the theoretical work of Donnelly and Tavaré [6] and supported by the empirically based tests of Crandall and Templeton [3] is the expectation that rare haplotypes should be preferentially connected to common haplotypes rather than other rare haplotypes. This relationship can also be used to refine alternative cladogram probabilities by weighting each connection by the frequency of the alternative haplotypes in the population over the sum of these frequencies (Crandall & Templeton [3]; Crandall et al. [5]).

Accuracy of the Cladogram Estimation Procedure.

Crandall [1] utilized a published data set of restriction sites from the bacteriophage T7 generated to produce a known phylogeny (Hillis et al. [10]). These restriction site data were subsampled with replacement for sixteen sites for each of the sixteen taxa in the phylogeny (including ancestral node taxa). Fifty data sets were generated by this subsampling and cladograms were estimated using the procedure of Templeton et al. [15], as well as bootstrapping with maximum parsimony. The results showed that the Templeton et al. [15] method perform near the stated confidence limits and that this method outperform bootstrapping with maximum parsimony (Crandall [1]). Thus, this method accurately reconstructs relationships when low levels of diversity exist within a data set and outperforms at least one traditional technique under these conditions. Results from simulation studies suggest that other traditional phylogeny reconstruction techniques also perform poorly when few characters are available (Huelsenbeck & Hillis [11]).

Nested Analyses.

Templeton et al. [14] and Templeton and Sing [17] have developed statistical procedures for detecting significant associations between phenotype and genotype within a cladogram framework. Their procedures utilize the cladogram structure from the above estimation procedure to define a nested statistical design; thereby allowing the clustering of individuals based on genotype rather that phenotype. The statistical analysis allows ambiguity in the cladogram estimation and is compatible with either quantitative or categorical phenotypes.

The nesting procedure consists of nesting $n-$step clades within $(n+1)-$step clades, where n refers to the number of transitional steps used to define the clade. By definition, each haplotype is a $0-$step clade. The $(n+1)-$step clades are formed by the union of all $n-$step clades that can be joined together by $(n+1)$ mutational steps. The nesting procedure begins with tip clades, i.e. those clades with a single mutational connection and proceeds to interior clades. The nesting procedure results in hierarchical nests with nesting level directly correlated to evolutionary time, i.e. the lower the nesting level the more recent the evolutionary events relative to higher nesting levels. The nesting design can then be used to test for significant associations of phenotype and genotype by either a nested analysis of variance

(NANOVA) for continuous data (Templeton et al. [14]) or a permutation chi-squared contingency test for categorical data (Roff & Bentzen [12];Templeton & Sing [17]).

Applications.

The cladogram estimation procedure as well as the nested analyses have been utilized in many contexts, especially in human population genetics. This is not surprising since human data sets typically represent individuals with very low levels of variation between sequences and are subject to recombination. The method has been used to explore viral transmission (Crandall [2]),to examine associations of risk factors in coronary artery disease to underlie candidate genes (Sing et al. [13]; Hallman et al. [7], Haviland et al. [8], Haviland et al. [8]), and to partition population substructure from population history (Templeton et al. [16]). Clearly, this broad range of applications signifies the utility of the approach as well as its appeal to researchers working within species.

Bibliography

[1] Crandall, K. A., Intraspecific cladogram estimation: Accuracy at higher levels of divergence. *Systematic Biology*, 43: 222–235, 1994.

[2] Crandall, K. A., Intraspecific phylogenetics: Support for dental transmission of human immunodeficiency virus. *Journal of Virology*, 69:2351–2356, 1995.

[3] Crandall, K. A. and Templeton, A. R., Empirical tests of some predictions from coalescent theory with applications to intraspecific phylogeny reconstruction. *Genetics*, 134:959–969, 1993.

[4] Crandall, K. A. and Templeton, A. R., Applications of intraspecific phylogenetics. In: New Uses for New Phylogenies (Harvey, P. H., Leigh Brown, A. J. & Maynard Smith, A. J., eds). pp. in press. Oxford University Press, Oxford, England, 1995.

[5] Crandall, K. A., Templeton, A. R. and Sing, C. F., Intraspecific phylogenetics: problems and solutions. In: Models in Phylogeny Reconstruction (Scotland, R. W., Siebert, D. J. & Williams, D. M., eds). pp. 273-297. Clarendon Press, Oxford, England, 1994.

[6] Donnelly, P. and Tavaré, S., The ages of alleles and a coalescent. Advances in Applied Probability 18:1-19, 1986.

[7] Hallman, D. M., Visvikis, S., Steinmetz, J. and Boerwinkle, E., The effect of variation in the apolipoprotein B gene on plasma lipid and apolipoprotein B levels. I. A likelihood-based approach to cladistic analysis. Annals of Human Genetics 58:35-64, 1994.

[8] Haviland, M. B., Ferrell, R. E. and Sing, C. F., A cladistic analysis of the relationship between variation in the low density lipoprotein receptor gene region and interindividual variation in plasma lipid, lipoprotein and apolipoprotein levels. American Journal of Human Genetics in press, 1995.

[9] Haviland, M. B., Kessling, A. M., Davignon, J. and Sing, C. F., Cladistic analysis of the apolipoprotein AI-CIII-AIV gene cluster using a healthy French Canadian sample. I. Haploid analysis. Annals of Human Genetics in press, 1995.

[10] Hillis, D. M., Bull, J. J., White, M. E., Badgett, M. R. and Molineux, I. J., Experimental phylogenetics: Generation of a known phylogeny. Science 255: 589-591, 1992.

[11] Huelsenbeck, J. P. and Hillis, D. M., Success of phylogenetic methods in the four-taxon case. Systematic Biology 42: 247-264, 1993.

[12] Roff, D. A. and Bentzen, P., The statistical analysis of mitochondrial DNA polymorphisms: Chi-square and the problem of small samples. Molecular Biology and Evolution 6: 539-545, 1989.

[13] Sing, C. F., Haviland, M. B., Zerba, K. E. and Templeton, A. R., Application of cladistics to the analysis of genotype-phenotype relationships. European Journal of Epidemiology 8: 3-9, 1992.

[14] Templeton, A. R., Boerwinkle, E. and Sing, C. F., A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. I. Basic theory and an analysis of alcohol dehydrogenase activity in Drosophila. Genetics 117: 343-351, 1987.

[15] Templeton, A. R., Crandall, K. A. and Sing, C. F., A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. Genetics 132: 619-633, 1992.

[16] Templeton, A. R., Routman, E. and Phillips, C. A., Separating population structure from population history: A cladistic analysis of geographical distribution of mitochondrial DNA haplotypes in the Tiger Salamander, Ambystoma tigrinum. Genetics: submitted 1995.

[17] Templeton, A. R. and Sing, C. F., A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. IV. Nested analyses with cladogram uncertainty and recombination. Genetics 134: 659-669, 1993.

A nonlinear invariant for a model without the homogeneity of sites assumption

Vincent Ferretti[1]
David Sankoff

Centre de recherches mathematiques
Université de Montréal
CP 6128 Succursale Centre-Ville
Montreal, Quebec
Canada H3C 3J7
ferretti@ere.umontreal.ca
sankoff@ere.umontreal.ca

We wish to infer the branching stucture of the evolutionary tree $\boldsymbol{T}$ of a group of N observed species. All we know about $\boldsymbol{T}$ is that it contains N terminal vertices, each representing one observed species.

As data, we have N aligned nucleic sequences of length n on the state space $\boldsymbol{\sigma} = \{1, \ldots, k\}$, one from each species. For each position i of this alignment, there is an (unknown) set $\boldsymbol{M}_i = \{\boldsymbol{m}_{XY} : XY \text{ is an edge of } \boldsymbol{T}\}$, where $\boldsymbol{m}_{XY}$ is the $k \times k$ substitution matrix of a continuous time Markov process. We assume that the set $\boldsymbol{M}_i$, the set of generalized rates of evolution at the position i, is included in a semi-group $\boldsymbol{S}$ of positive determinant matrices, and $\det \boldsymbol{m}_{XY}$ may be considered as a generalized length for the edge XY of the tree. The parametric form of $\boldsymbol{S}$ represents a model for the evolutionary process.

Given a set of substitution matrices $\boldsymbol{M}_i$ for all n values of i, containing one matrix for each edge of $\boldsymbol{T}$, the vector $\boldsymbol{f}$, containing the expected frequencies, across all positions of the alignment, of each of the k^N possible N-tuples (or configurations) of states, is called the expected spectrum.

A phylogenetic invariant for a tree $\boldsymbol{T}$ is a polynomial function Q of the expected spectrum $\boldsymbol{f}$ that is identically zero for any choice of sets $\boldsymbol{M}_i$, $1 \leq i \leq n$, and variable (and nonzero) for all other trees with N terminal vertices. If the value of Q based on the *observed* frequencies of the configurations along the alignment as estimators of the expected spectrum $\boldsymbol{f}$ is clearly closer to zero than the invariants corresponding to other trees, then $\boldsymbol{T}$ is probably the correct tree, in the sense that the data are generated on $\boldsymbol{T}$ according to the model of evolution adopted.

Two distinct traditions have emerged in this field. The class of linear invariants has been characterized exhaustively for several models in a series of papers by Lake, Cavender, and Fu and colleagues (see references in [3]). The results on other polynomial invariants, because of their nonlinearity, are less systematic, but many problems have been studied. Phylogenetic comparison has been extended for large N and k, and a wide variety of models have been investigated, for many of which there are no linear invariants. This includes models that are asymmetric, others where evolution in adjacent positions is not independent, and others that can be described as random walks on Abelien groups (see references in [1, 2]).

[1]Research Supported by grants from the Natural Sciences and Engineering Research Council of Canada and the Canadian Genome Analysis and Technology program.

One of the advantages often cited for linear invariants in practical applications is that they are not sensitive to inhomogeneities in rates of evolution at different sequence positions while polynomial invariants are only valid for sequences where homogeneity is strictly observed. In this paper, however, we set up a model for evolution where the positions fall into two distinct classes (e.g. RNA secondary structure stems versus single-stranded regions, first two positions of a codon versus the third, mRNA versus non-coding RNA) and find a cubic invariant that is valid, no matter what the proportion of positions in each class.

The two different "rates" are actually two different sets $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ of matrices, and evolution at all positions proceeds either according to one or the other. For example, all the matrices in $\boldsymbol{M}_1$ may have a large determinant, i.e. represent slowly evolving positions, while $\boldsymbol{M}_2$ may represent the faster rates at other positions, though the "fast-slow" distinction between the two sets of position need not to be the same for all branches of $\boldsymbol{T}$. For each θ, $0 \leq \theta \leq 1$, we look for a nonlinear invariant for a spectrum $\boldsymbol{f}$ of form

$$\theta \boldsymbol{f}_{\boldsymbol{M}_1} + (1-\theta)\boldsymbol{f}_{\boldsymbol{M}_2}.$$

We focus on the tree $\boldsymbol{T}$ where species A and B are grouped together and species C and D are grouped together. We use the Jukes & Cantor model on the state space $\boldsymbol{\sigma} = \{1, \ldots, k\}$ defined by the semigroup of transition matrices of the form $\boldsymbol{m}_{XY} = a\boldsymbol{J} + (1-ka)\boldsymbol{I}$, where $a < 1-(k-1)a$. Denote:

$$\begin{array}{lll}
f_0 = \sum f(i,i,i,i) & f_1 = \sum f(i,i,i,j) & f_2 = \sum f(i,i,j,i) \\
f_3 = \sum f(i,i,j,j) & f_4 = \sum f(i,j,i,i) & f_5 = \sum f(i,j,i,j) \\
f_6 = \sum f(i,j,j,i) & f_7 = \sum f(i,i,j,j) & f_8 = \sum f(i,i,j,l) \\
f_9 = \sum f(i,j,i,l) & f_{10} = \sum f(i,j,l,i) & f_{11} = \sum f(i,j,l,l) \\
f_{12} = \sum f(i,j,l,j) & f_{13} = \sum f(i,j,j,l) & f_{14} = \sum f(i,j,l,m)
\end{array}$$

where $\boldsymbol{f}$ contains the expected frequencies of the configurations at species A,B,C, and D, respectively, across all positions of the alignment. The summations are taken over all combinations of terms in σ.

We found that all quadratic invariants for $\boldsymbol{T}$ are only trivially invariant since they can be factored as $f_i \times L$, where L is a linear invariant for $\boldsymbol{T}$. We then carried our search for invariants to the next highest degree of polynomial, degree 3, or cubic. We found one that could not be written as a function of linear invariants:

$$\begin{array}{l}
C = \\
(f_2 - f_1)\{(f_{11} - [k-2]f_4)(f_9 - f_{13}) + [k-2]([k-2]f_5 - f_{12})(f_4 - f_7)\} \\
+(f_8 - [k-2]f_1)\{(f_{11} - [k-2]f_4)(f_6 - f_5) + (f_4 - f_7)(f_{10} - f_9)\} + \\
(f_3 - [k-1]f_0)\{[k-2](f_6 - f_5)([k-2]f_5 - f_{12}) + (f_{10} - f_9)(f_{13} - f_9)\}
\end{array}$$

The importance of the invariant C lies not in its potential use for phylogenetic inference, since we already have linear invariants for this model. Its existence, however, refutes the common assumption that nonlinear invariants are restricted by the unrealistic assumption of rate homogeneity.

The fact that C does not depend on θ is remarkable. Not only do we not need to know which positions have the same rate, but we do not even have to know how many there are positions of each type.

Our method could be used to search for invariants for a model in which the distribution of rates can be modeled by a one- or two-parameter distribution. The next step would be to find a heterogeneous rate context in which there are no linear invariants, but one or more nonlinear ones.

Bibliography

[1] Ferretti, V. and Sankoff, D., "Phylogenetic invariants for more general evolutionary models," to appear in *Journal of Theoretical Biology*, 1995.

[2] Ferretti, V. and Sankoff, D., "A remarkable nonlinear invariant for evolution with heterogeneous rates," *Centre de recherches mathématiques, Technical Report*, 1995

[3] Steel, M. A. and Fu, Y. X., "Classifying and counting linear phylogenetic invariants for the Jukes & Cantor model," to appear in *Journal of Computational Biology*, 1995.

Estimation of Parameters of Evolutionary Models, with Applications in the Analysis of Large Data Sets

Nick Goldman

Department of Genetics
University of Cambridge
Downing Street
Cambridge CB2 3EH, England
N.Goldman@gen.cam.ac.uk

Various stochastic models are currently used to describe the evolution of DNA sequences, in particular in maximum likelihood and distance matrix methods in phylogenetics. These are generally Markov process models permitting substitution of a nucleotide (A, G, C or T) at a site in a sequence by any other base. Models of gradually increasing complexity have been introduced and are currently in use, from the simplest model in which all replacements occur at the same rate [1][1]), via models which allow for inequality of rates of transitions—A $\leftrightarrow$ G, C $\leftrightarrow$ T—and transversions—A, G $\leftrightarrow$ C, T—(e.g. [2]) or inequality of equilibrium frequencies of nucleotides (e.g. [1]) to models which incorporate all these features (e.g. [4]) and even include variation across nucleotide sites of the rate of evolution, modelled by a Gamma distribution of rates with controllable shape parameter[5]. The introduction of the more complex models, with their biologically-meaningful parameters, is intended to produce more realistic models that better fit real data.

Recently it has become possible to test these models, typically using computer-intensive parametric bootstrap techniques due to the failure of traditional asymptotic (χ^2) statistics[6, 7]. A typical finding is that current models are not good descriptions of the processes of molecular evolution for a given dataset, but that the most-complex models are significantly better than simpler ones[6, 8]. Further study has shown that the consequences of the failings of these models can be significant in biological applications. Used in phylogenetic inference, they can lead to incorrect evolutionary trees being inferred and to branch lengths of trees being systematically underestimated[8, 9].

Nevertheless, the ability to test and compare different models can lead to better understanding of *processes* of molecular evolution, as those biological features which are modelled by particular parameters that are shown (by statistical testing) to be important may be taken to be of biological significance. It is now clear that base frequencies, transition/transversion rate ratios and rate variation across sites are all important, and can be reliably estimated for datasets of interest[6, 8, 9]. It is relatively straightforward to perform likelihood ratio tests or even simpler z-tests to assess the necessity of these parameters[6, 9], and it is also readily possible to get an idea of the applicability of the 'molecular clock' (i.e. is there evidence that rates of evolution are constant throughout the whole tree?)[9].

It should be noted that the above contains slight simplifications, as the dependence of parameter estimates on one-another is not straightforward[9]. However, the increasing confidence with which we can make inferences about patterns of molecular evolution is beginning to suggest strategies for analyzing datasets that are too large to permit simultaneous choice of accurate model, estimation of the parameters of that model, and inference of phylogeny (as is theoretically possible using maximum likelihood methods). Since numerous studies have shown that accuracy of estimated trees is more affected by the accuracy of the model of base substitution used, with appropriate parameter values, than by the method of estimation employed (e.g. maximum likelihood, or the Fitch-Margoliash or neighbor-joining distance matrix methods), we might perform initial analyses to choose the best model and suitable parameter values—generally not too time-consuming—and then 'fix' this model and use it in the estimation method of choice.

Bibliography

[1] Jukes, T.H. & Cantor, C.R. Pp. 21–132 *in*: Mammalian Protein Metabolism (ed. Munro, H.N.). Academic Press, New York (1969).

[2] Kimura, M. *J. Mol. Evol.* **16**, 111–120 (1980).

[3] Felsenstein, J. *J. Mol. Evol.* **17**, 368–376 (1981).

[4] Hasegawa, M., Kishino, H. & Yano, T. *J. Mol. Evol.* **22**, 160–174 (1985).

[5] Yang, Z. *Mol. Biol. Evol.* **10**, 1396–1401 (1993).

[6] Goldman, N. *J. Mol. Evol.* **36**, 182–198 (1993).

[7] Goldman, N. *J. Mol. Evol.* **37**, 650–661 (1993).

[8] Yang, Z., Goldman, N. & Friday, A. *Mol. Biol. Evol.* **11**, 316–324 (1994).

[9] Yang, Z., Goldman, N. & Friday, A. *Syst. Biol.* , in press.

A Survey of Tree Consensus Criteria and Algorithms

Sampath K. Kannan
Department of Computer and Information Science
University of Pennsylvania
Philadelpia, PA 19104
kannan@central.cis.upenn.edu

In broad terms, the algorithmic problem of reconstructing evolutionary trees is to go from observed data on a set of species, to a tree (describing the evolution of the given set of species from a common ancestor) that "best fits" the given data.

This specification of the problem is far from unambiguous. In practice many different data sets are used for the same set of species. Several different optimization criteria are used, some appropriate for character-based data and others for distance-based data. Even if we fix a particular data set and a particular optimization criterion, since optimality is merely an attempt to capture reality, attention should be payed to trees that come close to the optimum and at the very least to all trees that achieve the optimum.

For all of the above reasons we typically have a number of trees purportedly describing the evolution of a set of species. The goal of tree consensus algorithms is to extract common information and produce one tree from this set of trees. In order to design a good consensus function mapping a set of trees to a tree, we first need to identify desirable criteria that such a function should meet.

Arrow's Theorem

The study of such consensus methods was perhaps first undertaken in mathematical economics where Arrow considered the problem of constructing a function that mapped a set of individual preferences (given by total orders among a set of options) into a consensus social choice (given by a total order on the same options). Arrow[1] showed that there are no consensus functions satisfying a very reasonable set of criteria.

Arrow's theorem motivated several pieces of work on tree consensus that tried to axiomatize the requirements on consensus functions and construct functions satisfying these requirements.

Following Arrow, researchers also attempted to express the requirements of a tree consensus function (and in general, consensus functions on other objects) in terms of primitive substructures of trees.

Primitive substructures of trees

Given a rooted tree T with leaf set S, associated with each node v is a subset $R(v) \subset S$ representing the set of leaves under v. These sets are called *clusters*. The cluster encoding $C(T)$ of T is the set of all clusters of T. Clusters are the most popular primitive substructure on which tree consensus methods have been based. Other primitive substructures include n_k*-taxon statements*, *nestings*, and *triples* the last of which is relevant to our work on *local consensus*.

Given a rooted tree T, if $x, y, z \in S$ such that the least common anncestor of x and y lies below the least common ancestor of x and z (and y and z) we denote this information by $(x, y), z$. If all pairwise least common ancestors of x, y and z are at node v, we denote this by (x, y, z). It is easy to see that given the structure of all the triples, the tree is uniquely determined just as it is from the cluster encoding.

Tree consensus criteria

Neutrality: For consensus function f if $f(T_1, \ldots, T_k) = T$ and $f(T'_1, \ldots, T'_k) = T'$ and x and y are primitive substructures such that $\{i | x \in T_i\} = \{j | y \in T'_j\}$, then $x \in T \Leftrightarrow y \in T'$.

Monotone Neutrality: This is a stronger condition than neutrality stating that if the set of indices of trees in which y is present is a superset of the set of indices of trees in which x is present, then if x is in the consensus tree, y also must be in the consensus tree.

Pareto Optimality: This condition requires that any primitive substructure that is present in all of the input trees must be present in the consensus tree.

Faithfulness: This criterion only applies to the situation when the primitive substructures are sets. (Thus it applies to clusters but not to triples.)

If $f(T_1, \ldots, T_k) = T$ and S_i is a primitive substructure in T_i for $i = 1, 2, \ldots, k$, then there exists a substructure S in T such that $\cup_{i=1}^{k} S_i \subseteq S \subseteq \cap_{i=1}^{k} S_i$.

Faithfulness may be seen to be a consequence of two natural criteria — independence of irrelevant alternatives and Pareto optimality.

Unfortunately, it has been shown by Neumann[5] that there are no tree consensus rules that are faithful and neutral.

Tree Consensus Criteria: The *strict consensus* criterion requires that the output tree contain exactly those clusters that are contained in all of the input trees. Day[3] gave an elegant $O(kn)$ time algorithm for constructing the strict consensus of k trees on n species.

Variants of strict consensus include the *majority consensus* (which requires that the output tree contain exactly those clusters that are present in more than half of the input trees) and the *median procedure* [2] (which looks for tree(s) that minimize a certain metric).

All of these consensus functions are neutral but not faithful. Intuitively, faithful criteria are criteria that preserve information that is present in the input trees.

Most of the known faithful criteria are expressed in terms of intersection rules[5]. One such rule is the Durchschnitt rule which is defined as follows. Define the height of a node in a tree (and of the cluster defined by the node) in the usual way — leaves have height 0 and internal nodes have height one greater than the maximum height of their children. Then the requirement is that any cluster that can be obtained by intersecting clusters of equal height, one from each input tree, should be present in the output tree. This rule can be generalized to other height functions and all of these rules can be shown to be faithful[5]. A popular rule called the Adam's consensus which is based on looking at nestings is very similar to some of these intersection rules.

Local Consensus Rules:[4] We are faced with the choice between neutral rules and faithful rules if we want a consensus rule to be defined on all inputs. Local consensus rules are a family of rules that contain both faithful and neutral rules. All of the local consensus rules can be efficiently computed.

A particular local consensus rule is defined by prescribing the topology of a triple in the output tree as a function of the topologies of that triple in each of the input trees. We will use the notation $((a, b), c)$ to indicate that a and b share a more recent least common ancestor in a tree in comparison with c. We will use (a, b, c) to indicate that each pair among a, b and c has the same least common ancestor as the other pairs. (In this case the least common ancestor v must have at least 3 children.)

An example of a local consensus rule for consensus of k trees is $f(((a, b), c), ((a, b), c), \ldots, ((a, b), c)) = ((a, b), c)$. This rule says that if all k input trees agree on the topology of a triple, then the output should have the same topology for the triple. This gives the biologist the ability to plug in a particular local consensus rule by specifying the required behaviour on each k-tuple of triple topologies.

We define some specific local consensus rules that seem natural. The *binary local consensus* rule applies to the case where all the input trees are binary. It essentially says that when all input trees agree on a triple, then the output tree should also have the same topology for that triple. On the other hand, when the input trees disagree then the output tree should leave the triple unresolved. Two generalizations of this rule are possible for non-binary input trees and these are called the *optimistic local consensus* and the *pessimistic local consensus* respectively. These trees may not always exist, but when they do they are neutral.

To define local consensus rules where a consensus tree always exists, we bring in a relaxed version of a local consensus rule. A relaxed rule simply does not specify the desired topology for some tuples of input topologies and called a *partial local consensus rule.* We could also have optimization versions of partial local consensus rules where for some tuples of input topologies we are given exact specifications on the desired output topology. On other tuples of input topologies we are given a *desired* output topology and our goal is to maximize the number of triples whose output topology conforms to the desired one while making sure that triples with prescribed topologies do have that topology in the output.

Two such natural optimizing rules are the *relaxed discord local consensus* and the *relaxed accord local consensus.* In the former, we require that for any triple on which all input trees agree, the output tree also has the same topology whereas for any triples on which the input trees disagree, we desire that the output tree has the unresolved topology. The latter rule reverses these conditions.

The relaxed discord and relaxed accord trees always exist. While the first rule is neutral, the second is faithful.

For all of the rules described above we have exact characterizations of what the output tree should look like if it exists. Using these, we compute the unique tree that would satisfy the rule if it exists and then in a second phase verify that the tree we have constructed does indeed satisfy the rule.

For all the rules described above we have linear time algorithms to compute and verify the desired output tree. The one exception is the relaxed accord local consensus which takes $O(n^2)$ time to compute.

Choices of local consensus rules may be made by the biologist based on detailed knowledge of the quality of the data. Also choosing a rule is more immediate and less abstract than choosing the criterion that one wants a consensus function to satisfy.

Many variants to the consensus approach such as the maximum agreement subtree (MAST), median procedure, tree compatibility, etc., can be seen to be connected to particular local consensus rules.

Bibliography

[1] K. Arrow, Social Choice and Individual Values, John Wiley and Sons, New York, 1951.

[2] Barthélemy, J. and McMorris, F. *The median procedure for n-Trees*, Journal of Classification, 3:329-334 (1986).

[3] W.H.E. Day, Optimal algorithms for comparing trees with labeled leaves. *Journal of Classification*, 2;7-28 (1985).

[4] S. Kannan, T. Warnow, and S. Yooseph, Computing the Local Consensus of Trees, Proc. Symposium on Discrete Algorithms, San Francisco, Jan 1995.

[5] D.A. Neumann, Faithful Consensus Methods for n-Trees, *Mathematical Biosciences*, 63:271 – 287 (1983).

A Biologically Meaningful Model for Comparing Molecular Phylogenies

Boris Mirkin
DIMACS
Rutgers University
P.O.Box 1179
Piscataway, NJ 08855-1179, USA.
mirkin@dimacs.rutgers.edu

Ilya Muchnik
RUTCOR
Rutgers University
Bush Campus
New Brunswick, NJ 08903, USA.
muchnik@rutcor.rutgers.edu

Temple F. Smith
Biomolecular Engineering Research Center
College of Engineering
Boston University, Boston MA 0221
tsmith@bu.edu

Key words: phylogeny tree, gene evolutionary tree, inconsistency, dissimilarity of the trees, duplication/speciation/loss modeling.

The current strategy for the phylogeny reconstruction is based on the separate consideration of different gene families which are represented by so called homologous sequences (Li and Graur [4], Baum [3]). The homologous sequences are supposed to be descendants of a common ancestor in such a way that the more similar sequences correspond to the more recent common predecessors. An evolutionary tree constructed for a particular family or super-family serves as a structural support for analysis of the history of corresponding functions. This can be considered as the first step in the following two-step strategy for reconstructing evolutionary tree for species.

1. Constructing a phylogenetic tree for each of the gene families independently.

2. Combining particular gene trees into a unique tree representation.

This talk concerns Step 2: combining different, often contradictory, gene trees into a unique evolutionary tree. The problem is considered in the literature as the problem of reconciling different phylogenetic trees. Existing mathematical methods consider phylogenetic trees as formal mathematical objects and try to find as much alike in the structures as possible (Adams [1], Barthélemy, Leclerc and Monjardet [2]). The shortcoming of this approach is that the concepts of “similar structures” involved are purely formal, having no independent biological meaning. For instance, there is no formal difference between inconsistencies that arise from different techniques applied to the same data or the difference in the data themselves.

We introduce a biologically meaningful model for embedding a particular gene tree into the general evolutionary tree, via duplication/speciation/loss events (as considered in Ohno [5], Li and Graur [4]). Based on the model, an asymmetric dissimilarity measure for the trees is defined. The measure is compared with another one defined in terms of a mapping procedure for the trees. A graph-theoretic reformulation of the measure is derived.

Rooted binary tree is considered as a set of nested subsets of its leaf set. The major procedure of embedding a "gene" G phylogeny into a "species" phylogeny T (where the leaf set J of G is assumed included into the leaf set I of T) is based on the following.

The trees G with the root g and T with the root t are referred to as *root-consistent* if each of the sets of the root children $a(g), b(g)$ in G is contained in a set $a(t)$ or $b(t)$ of the root children in T: the root divergence is the same in both of the trees, in this case. If the trees are not root-consistent, a duplication of G is postulated in the root t to explain the inconsistence based on the following principle.

Duplication/Speciation Principle. The duplication event in the gene corresponding to tree G has occured at the root of evolutionary tree T and has evolved in T in such a way that contemporary organisms corresponding to the leaves in $a(g)$ have one of the divergent alleles while the leaves in $b(g)$ the other.

Based on this postulate, we can define all the events related to the duplication in the evolutionary tree T in a unique way with a special "sign labeling" procedure leading to a complete description of the history of the duplication event across T (along with indication of corresponding speciation/loss events).

Embedding tree G into T consists in sequential embedding all its subtrees into corresponding contractions of T via the duplication model above. The total number of duplications/losses/information gaps after embedding done is considered an inconsistency measure C(T/G).

Two ways for reformulation the duplication events defined are considered as based on: (1) mapping of tree G into T, and (2) incompatible pairs of nodes (g,t), $g \in G$, $t \in T$.

The mapping $M : G \to T$ is defined as follows: For any interior node $g \in G$, its image $M(g)$ is the node of T having the smallest cluster satisfying the following inclusion: $g \subseteq M(g)$.

It is proved that the duplications correspond to such interior nodes $g \in G$ that $M(g) = M(a(g))$ or/and $M(g) = M(b(g))$. The inconsistency function also could be reformulated in terms of the mapping, although the speciation/loss events cannot be assigned to the corresponding duplications properly when only the mapping is utilized.

Another concept represents the duplication events more properly.

Let us refer to $g \in G$ and $t \in T$ as *incompatible* nodes, if the following holds:

$$g \not\subseteq t \text{ and } a(g) \cap t \neq \emptyset \text{ and } b(g) \cap t \neq \emptyset.$$

The maximal incompatible pairs (g,t) are proved to correspond to the duplications related to subtrees with the roots g and $p(t)$ where $p(t)$ is the parent of t in T. Moreover, the following equality is proved:

$$C(T/G) = I(G,T) + 3M(G,T)$$

where $I(G,T)$ ($M(G,T)$) is the cardinality of the set of the (maximal) incompatible pairs. The equality shows that the inconsistency measure counts any maximal incompatible pair four times (once in $I(G,T)$ and thrice in $3M(G,T)$) while ordinary incompatible pairs (proved to correspond unspeciated ancestors) are counted only once. In particular, when any incompatible pair is maximal, $C(T/G) = 4I(G,T)$.

The properties of the measure proved can be employed in the problem of the species tree reconstruction from a set of different gene trees, as based on minimization of the overall inconsistency measure.

Bibliography

[1] Adams, E.N. III., N-Trees as Nestings: Complexity, Similarity and Consensus, *Journal of Classification*, **3** 1986, 299–317.

[2] Barthélemy, J.P., Leclerc, B., and Monjardet, B., On the Use of Ordered Sets in Problems of Comparison and Consensus of Classifications, *Journal of Classification*, **3** 1986, 187-224.

[3] Baum, B.R., Combining Trees as a Way of Combining Data Sets for Phylogenetic Inference, and the Desirability of Combining Gene Trees, *Taxon*, **41** 1992, 3-10.

[4] Li, W.-H., and Graur, D., *Fundamentals of Molecular Evolution*, Sunderland: Sinauer Associates, 1991.

[5] Ohno, S.,*Evolution by Gene Duplication*. Springer-Verlag, Berlin, 1970.

Maximum likelihood approaches to sequence-based phylogenetic analysis: Some histories would make the present much less improbable than others

Gary J. Olsen
Department of Microbiology
University of Illinois
Urbana, IL 61801
gary@phylo.life.uiuc.edu

Molecular phylogeny is the study of the history of a portion of the genome. The most frequent use of molecular phylogenies is to extrapolate the gene relationships onto the organisms themselves. For this to be reliable there cannot be gene transfer between organisms and it is essential to avoid confusion among the members of a multigene family. When several genes give the same answer, the extrapolation to the organism is probably valid. Molecular phylogeny is responsible for our understanding of the historical relationships among microorganisms and of the origins of plants, animals and fungi (Woese [8]). One of the most profound discoveries is that prokaryotic life (cellular organisms without true membrane-bound nuclei) is of two kinds: the Bacteria (or Eubacteria), which includes most of the commonly studied "bacteria", and the Archaea. The eukaryotes (called Eucarya) are specific relatives of the Archaea (see Woese *et al.* [9]).

When one is interested in the relationships of genes per se, then the above restrictions on gene transfer and gene families are unnecessary. Gene trees can be used to establish the correspondence of genes in different organisms (and guessing the function of a new sequence), studying the histories of multigene families (thereby studying the origins of the genetic complexity seen in organisms today), and studying the dynamics of sequence change. The dynamics of sequence change can in turn be used to study the functional constraints on sequences and to refine the models used in phylogenetic inference.

The analysis of organismal relationships is not just an exercise in taxonomy. From a history of organisms can we infer the history of important physiologies and biochemical activities. For example, the distribution of methanogenesis (production of methane, a significant green-house gas) suggests that it was invented once, and then lost at least four times. Subsequence analyses of the genes responsible for methanogenesis were consistent with this view. Thus it seems that ancient methanogens (and perhaps the present-day ones as well) had sufficient genetic flexibility to successfully abandon methanogenesis as a source of energy and adopt a different life-style. Other properties also have interesting histories. The distribution of extreme thermophily among Bacteria and Archaea suggests that most living forms represent adaptations of thermophilic ancestors to more mild conditions. The distribution of chlorophyll-based photosynthesis is consistent with a single origin and numerous losses. In contrast, oxygenic photosynthesis (which is responsible for nearly all the oxygen in our atmosphere) seems to be an irreversible specialization; there are no know cases in which it has been successfully lost. Oxygen respiration has a very complex distribution, and might involve multiple invention and/or gene transfers.

If a hypothesis about history is to be more than an empty exercise, then it must make predictions. For example, if the Archaea and Eucarya are specific relatives, then there should be innovations shared by these two groups, but absent in the Bacteria. When this was explored by sequencing random pieces from the genome of an archaeon, we found a sequence with close relationship to the TATA-binding protein (TBP) of eukaryotes, the key element in the first step of gene expression (Marsh *et al.*, 1994). As a consistency check, the molecular phylogeny of the TBP genes was shown to match the organismal relationships inferred from other molecules. Because TBP is comprised of two copies of a domain that was duplicated in the distant past, the phylogeny of the half molecules could be used to study the gene duplication. The data were most consistent with a single duplication, before the divergence of the archaeal and eukaryotic lineages. To date, all characterized components of the archaeal transcription apparatus (the machinery that copies a portion of the DNA into RNA) are qualitatively and quantitatively more similar to those of Eucarya than to those of Bacteria.

Given the multiplicity of applications of molecular phylogeny, we have been interested in reliable methods for inferring molecular relationships. The first step for (essentially) all methods is to define a correspondence of residues in the various molecules. This process, called sequence alignment, has a very strict biological definition; we require that all residues that are aligned (placed in the same column) be descended from the same residue in the corresponding gene in their common ancestor. The identity of the residues may have changed one or more times, but the key element is that present-day residues trace back to a single ancestral residue. Several caveats follow from this definition:

(*i*) an "optimal" alignment may not be historically correct;

(*ii*) multiple sequence alignment programs frequently estimate a phylogenetic tree using an inferior method and then base the alignment on this tree; and

(*iii*) if a region of alignment is wrong, it will contribute nonsense (possibly systematic noise) to subsequent analyses.

The best protection against these problems is to limit phylogenetic analyses to those regions of the alignment that are most certain.

Given an alignment, it is necessary to pick a tree inference method. Sorting through the available methods is a challenge. This is aided by categorizing them in two dimensions. The first dimension distinguishes clustering methods from objective-based methods. The former are definitions of how to find a tree, while the latter are definitions of how to evaluate trees. Only objective-based methods allow the comparison of the relative merits of any two trees. The biological information is entirely contained in the definition of the objective criterion, while the strategy for finding the best tree according to the objective is left as an exercise. The second dimension distinguishes methods that impose equal divergence on all lineages from those that do not. The former methods are frequently said to impose a molecular clock; they result in an ultrametric tree. The latter methods yield a more general additive tree (or do not assign branch lengths at all). When distant relationships are examined, the data are often not consistent with equal divergence in all lineages, so we routinely use methods that do not impose a clock.

Even after limiting methods to objective-based methods that do not impose a clock, many methods remain. Among these are:

(*i*) maximum parsimony, seeking the history that requires the fewest events to produce the observed data;

(*ii*) least squares distance methods, seeking the history for which the expected number of sequence differences best fits the observed sequence differences; and

(*iii*) maximum likelihood, seeking the history with the highest probability of giving rise to the sequences that we observe today. The distance and maximum likelihood methods require a detailed model of how sequences change.

All of the biology is encapsulated in these models, the rest is arithmetic that is consistent (it will converge on the correct answer if the model is correct and infinite data are available), but which differs in how close it comes to optimal use of the data. We have come to favor maximum likelihood methods.

Joseph Felsenstein has long advocated maximum likelihood methods, and he was the first to write a program for maximum likelihood analysis of nucleotide sequences[2]. In maximum likelihood methods, it does not matter that the absolute probability of the best tree giving rise to the data is very low, since this is the probability of reproducing the process, not the probability that it happened (just as 1000 coin tosses will give some sequence of heads and tails, it is reproducing that sequence that has the a probability of about 10^{-300}). What does matter is that the data would be much more likely to result from some histories than from alternative histories, or from independent origins.

Felsenstein's program [3] was considered slow, and was only rarely used. We wrote a faster version (fastDNAml) that could also be run on parallel supercomputers (Olsen *et al.*, [7]). Using this program we were able to infer large (over 60 sequence) maximum likelihood trees. Partially overlapping trees were pieced together to make a 473 sequence tree of prokaryotic 16S ribosomal RNA sequences.

When considering any phylogenetic tree it is important to realize three things:

(*i*) it might not be the global optimum (it was not for some of our subtrees);

(*ii*) even if it were optimal, the simplifications made in the evolutionary model and the finite amount of data available mean that the optimal tree may not be the historically correct tree (in a large tree, the number of minor variants that fit the data almost equally well virtually ensures that the optimal tree is not the true tree); and

(*iii*) when the model is reasonable and there are sufficient data, the inferred tree is usually a very good approximation of the true tree.

That is, considering a molecular phylogeny to be reality is naïve, but using it as a working hypothesis is very powerful.In addition, a variety of methods can be used to assess which features a most certain, and which are least certain.

Most phylogenetic inferences in the literature are based on analyses that treat all positions in the sequences as evolving according to the same rules—in particular, at the same rate. This is not consistent with the tendency to refer to conserved regions, variable regions, etc. Limited treatments of site heterogeneity have considered molecules composed of sites drawn from a population with a distribution of rates, but have not assigned specific rates to specific sites (Olsen [6]; Jin & Nei [3]). One reason for this is that assigning a specific rate to each site adds about 2000 adjustable parameters to the evolutionary model, an anathema to statisticians.

We have been exploring the value and justification of estimating and using site-specific rates of change (S.C. Pracht, R. Overbeek, and G.J.O., unpublished). We have written a program (DNArates) to give a maximum likelihood estimate of the rate of change at each position in an RNA or DNA sequence alignment and a phylogenetic tree. We are also seeking to defined criteria that should be met if site-specific rates of change are to be usefully employed. Current results suggest that the criteria proposed are met with many molecular data sets. It seems that as few as 10 to 20 related sequences are sufficient to infer rates of change that are significantly better than the default assumption that sites are homogeneous.

Bibliography

[1] Barns, S. M., Fundyga, R. E., Jeffries, M. W., and Pace, N. R. 1994. Remarkable archaeal diversity detected in a Yellowstone National Park hot spring environment. *Proc. Natl. Acad. Sci. USA* **91:** 1609–1613.

[2] Felsenstein, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.* **17:** 368–376.

[3] Felsenstein, J. 1990. *PHYLIP Manual Version 3.3.* University Herbarium, University of California, Berkeley, CA.

[4] Jin, L., and Nei, M. 1990. Limitations of the evolutionary parsimony method of phylogenetic analysis. *Mol. Biol. Evol.* **7:** 82–102.

[5] Marsh, T. L., Reich, C. I., Whitelock, R. B., and Olsen, G. J. 1994. Transcription factor IID in the Archaea: Sequences in the Thermococcus celer genome would encode a product closely related to the TATA-binding protein of eukaryotes. *Proc. Natl. Acad. Sci. USA* **81:** 4180–4184.

[6] Olsen, G. J. 1987. The earliest phylogenetic branchings: Comparing rRNA-based evolutionary trees inferred with various techniques. *Cold Spring Harbor Symp. Quant. Biol.* **52:** 825–838.

[7] Olsen, G. J., Matsuda, H., Hagstrom, R., and Overbeek, R. 1994. fastDNAml: A tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. *Comput. Appl. Biosci.* **10:** 41–48.

[8] Woese, C. R. 1987. Bacterial evolution. *Microbiol. Rev.* **51:** 221–271.

[9] Woese, C. R., Kandler, O., and Wheelis, M. L. 1990. Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci. USA* **87:** 4576–4579.

Accepted-mutation Parsimony: A Method for Reconstructing Trees from Highly Divergent Amino Acid Sequences

Kenneth A. Rice
Department of Organismic and Evolutional Biology
Harvard University
Cambridge, Massachusetts
rice@green.harvard.edu

Lee F. Kolakowski
Children's Hospital
Boston, Massachusetts
lfk@receptor.mgh.harvard.edu

Increased awareness that modern proteins are derived from ancient multigene families has motivated a search for phylogenetic reconstruction methods that work on large data sets composed of hundreds of highly divergent proteins, and that allow differentiating identity by homoplasy from identity by descent. Current implementations of maximum likelihood approaches are so slow that they work only on fewer than 100 sequences, leaving two broad classes of phylogenetic reconstruction methods which can be applied to large, highly divergent data sets. These are transformed distance methods and heuristic parsimony methods. Transformed-distance methods are found to perform well in a wide variety of variable-rate cases [8], but they do not allow discrimination between hypotheses of identity by descent and identity by homoplasy.

Differentiating homoplasy from identity by descent turns out to be very important in elucidating the history of multigene families; homoplasy is ubiquitous in the evolution of protein families [2]. This motivates use of a parsimony method for phylogenetic reconstruction, since parsimony methods allow discriminating identity by homoplasy from identity by descent if the data are relatively free of noise. Unfortunately, parsimony has been shown to be inconsistent when the rates of evolution differ greatly across lineages [1]. Character weighting and character-state transition weighting are used for enlarging the region of rate-variation space in which parsimony performs consistently. A recent study comparing the efficacy of several methods of tree reconstruction at recovering known trees found weighted maximum parsimony to perform well in a wide variety of four-taxon, unequal-rate cases [8]. Swofford's parsimony program PAUP incorporates a stepmatrix, a character-state-transition matrix which allows arbitrary costs to be assigned to state changes. Entries in the stepmatrix are larger if the change is rarer or more difficult [4].

We have developed an accepted-mutation stepmatrix to be used in combination with heuristic parsimony to classify G-protein-coupled receptors and other ancient protein families. We have named this technique *accepted-mutation parsimony* (AMP). A simple form of accepted-mutation stepmatrix can be constructed as follows. Given a collection of amino acid sequences, perform all pairwise comparisons and place the substitution probabilities into a 20 by 20 symmetric matrix M, where M_{ij} is the probability of a substitution of amino acid i by amino acid j. Next replace each entry M_{ij} by $1 - M_{ij}$. Finally, scale each stepmatrix entry M_{ij} so that it falls on the interval $[1, n]$, where n is a positive integer (e.g. 100). The principal diagonal contains zeroes by convention, implying that there is no cost for a continuation. The stepmatrix is used in an otherwise ordinary maximum parsimony analysis to weight character state transitions. Stepmatrix entries increase with the rarity of the substitution, which has the effect of downweighting common (and presumably noisy) replacements.

The foregoing stepmatrix is somewhat unrealistic because a rare amino acid appears to participate in substitutions less often than a common one, other things being equal. This motivates another form of stepmatrix in which entry M_{ij} is conditioned on the prevalence of amino acid i in the following fashion. Let M_{ij} be the probability of a substitution of amino acid i by amino acid j, as before. Next divide each entry in the i'th row of M by the frequency of the i'th amino acid. Then replace each M_{ij} by $1 - M_{ij}$ and scale as before. Note that this leads to an asymmetric stepmatrix. Another refinement of the basic technique is to coerce each triple of elements from the 20 by 20 stepmatrix to satisfy triangle inquality. Finally, an iterative version of the method has been developed which obtains substitution probabilites from a tree produced by unweighted parsimony.

Amino acid sequence alignment programs historically have used Dayhoff's PAM matrix or a similar matrix based on the substitution probabilities of several proteins, and one such matrix is used to align all proteins.[5, 6] It is therefore tempting to pick a representative collection of soluble proteins and calculate an all-purpose protein stepmatrix. This is a mistake, however, because different protein data sets yield very different matrices.

For example, the matrices derived from G-protein-coupled receptors and from globins are quite different. This can be shown by the correlation coefficient for the corresponding elements of the upper triangles of the two matrices. Pearson's r = 0.436, t = 6.64, df = 190-21=169. Most of the differences in the receptor and globin matrices appear to be due to differences in hydropathicity of the two protein families. Globins, like other soluble, globular proteins, have a concentration of hydrophobic residues such as leucine and isoleucine in the molecule's interior. Thus a change from an isoleucine to a leucine is likely to change the steric conformation of the molecule, and there is strong selective pressure against such changes. G-protein-coupled receptors, on the other hand, are membrane-spanning proteins which are formed as a bundle of seven predominately hydrophobic αhelices. Hydrophilic residues and critical ligand-binding sites tend to be found on the internal faces of the helices forming the α-helical bundle, whereas hydrophobic residues tend to be found on the outside. Thus a replacement of leucine by isoleucine is not likely to change the receptor sterically, nor to disrupt the ligand-binding site.

Accepted-mutation parsimony has been used to study the evolution of G-protein-coupled receptors and to classify orphan receptors [2]. A more definite appraisal of the merits of accepted-mutation stepmatrices must await the creation of a simulator which can produce data sets from a given tree and a given set of amino acid functional constraints; this is the direction in which the research is proceeding. In the meantime, the following procedure is suggested: The stepmatrix must be calculated from a data set large enough to ameliorate the effects of sampling error. Therefore, calculate the stepmatrix from the data set itself if there are more than about 3000 amino acids in the data set. Otherwise, calculate it from an alignment of proteins which are believed to have similar amino acid functional constraints. Simple and frequency-adjusted accepted-mutation stepmatrices are calculated by the program *amp*, which is available as C source code from the corresponding author at rice@oeb.harvard.edu. *Amp* accepts a collection of aligned amino acid sequences in Stanford-Intelligenetics format, and it yields a stepmatrix with appropriate PAUP type declarations.

Bibliography

[1] Huelsenbeck, J. P., and Hillis, D. M., "Success of phylogenetic methods in the four-taxon case," *Systematic Biology*, **42** (1993), 247–264.

[2] Rice, K. A., "The Origin, Evolution, and Classification of G-protein-coupled Receptors," *Doctoral Thesis, Harvard University, Department of Organismic and Evolutional Biology*, Cambridge, Massachusetts (1994).

[3] "Cases in which parsimony and compatibility methods will be positively misleading," *Systematic Zoology*, **27** (1978), 401–410.

[4] Swofford, D. L., "PAUP: Plylogenetic Analysis Using Parsimony," Version 3.0s+3, pp. 1–179, Illinois Natural History Survey, Champaign, IL (1991).

[5] Pearson, W. R., "Rapid and sensitive sequence comparision with FASTP and FASTA," *Methods in Enzymology*, **183** (1990), 63–98.

[6] Hein, J., "Unified approach to alignment and phylogenies," *Methods in Enzymology*, **183** (1990), 626–645.

A Biologist's Perspective on Interesting Questions and Practical Difficulties in Phylogeny Reconstruction

David S. Roos
Department of Biology
University of Pennsylvania
Philadelphia, PA 19104-6018
droos@sas.upenn.edu

Background – Drug sensitivity and resistance a protozoan parasite.

The protozoan parasite Toxoplasma gondii is a ubiquitous pathogen thought to affect 30% of the US population. Infections rarely produce clinical symptoms, however, because the host's cellular immune response is unusually effective in controlling the acute (tachyzoite) form. T. gondii parasites are also capable of differentiation into latent 'tissue cysts' (bradyzoites), which are not eliminated by the immune response and probably remain viable within the tissues of infected individuals for life, reemerging occasionally to provide a natural boost to host immunity. In this context, it is not surprising that acute toxoplasmosis has classically been known from immunosuppressed individuals (primarily transplant or cancer patients) and from congenital cases[1]. More recently, toxoplasmic encephalitis has emerged as a leading opportunistic infection associated with AIDS[23]. Traditional chemotherapeutic management of toxoplasmosis involves a combination of sulfonamides (such as sulfadiazine) and antifolates (such as pyrimethamine). Blocking sequential steps in the critical folate metabolic pathway[3], these compounds are synergistic in their action against Toxoplasma, and have proven highly effective in otherwise healthly individuals, who require only a temporary boost to their innate ability to fight off the infection. The chief difficulty in controlling congenital toxoplasmosis arises from the lack of detection: as infection is generally subclinical, the pregnant woman may not realize she has become infected[4]. Toxoplasmosis associated with AIDS poses a much more difficult problem: because T. gondii-infected individuals harbor latent bradyzoite cysts within their tissues, treatment of HIV-positive patients must be maintained for life, to guard against reemergence of a parasite infection which the diminished host immunity would be unable to combat. Unfortunately, chronic pyrimethamine/sulfa therapy is often impossible, due to the development of hypersensitive (allergic) reactions to sulfonamides. Pyrimethamine alone is insufficient to prevent relapse. Long-term antifolate therapy may also be associated with the emergence of drug resistance. These unfortunate conditions lend considerable urgency to the development of novel drugs and treatment strategies for AIDS-toxoplasmosis. Beyond its direct medical importance, T. gondii is an appealing organism for study from a biological perspective. The mechanisms involved in regulating its developmental life cycle – from tachyzoite, to bradyzoite, to sexual stage gametes, to oocyst, to sporozoite, and back to tachyzoite – are completely unknown. Toxoplasma also provides a revealing window into the evolutionary origins and phylogenetic classification of eukaryotes, as it is sufficiently similar to more familiar eukaryotes (such as ourselves) to be easily understood from a cell biological and molecular genetic perspective, but sufficiently different to be suitable as a roadmap for understanding certain more highly derived taxa (such as malaria parasites)[5, 6].The recent development of molecular tools suitable for the genetic analysis of Toxoplasma renders this parasite among the most accessible systems for experimental manipulation[6, 7]

Interesting Questions – Applications of phylogenetic analysis to drug design.

How might phylogenetic analysis facilitate the development of new (or improved) drugs effective against Toxoplasma gondii?

1. Phylogenetic trees can suggest effective drugs by identifying sister taxa. Among the characters likely to be shared by related taxa are similarities in drug sensitivity. Indeed, pharmacological markers have commonly been used as traits for classical phylogenetic analysis. Reversing this process, identification of related groups may provide insights into suitable chemotherapeutic approaches. For example, although Pneumocystis carinii (another opportunistic pathogen associated with AIDS) has sometimes been classified a sporozoan protist similar to Toxoplasma, molecular analysis unequivocally groups this organism with the fungi[8]. This provides a rationale for the greater effectiveness of antifolates such trimethoprim (rather than pyrimethamine), and suggests new courses of therapy based on long experience with antifungal agents. Phylogenetic analysis argues that the coccidian parasites (including T. gondii, malaria and many parasites of veterinary importance) are monophyletic, and closely related to dinoflagellates and ciliates5.

2. Phylogenetic analysis can determine how different a parasite target is from the host cell homolog, and hence how suitable for drug development. Where appropriate treatment strategies cannot be devised simply by analogy with therapies proven for related taxa, comparison of specific traits (enzymes, behavioral patterns, etc) can often be of use in ascertaining the likelihood of effective intervention. For example, the enzymes dihydropteroate synthase (DHPS) and uracil phosphoribosyl transferase (UPRT) are central to pyrimidine nucleoside metabolism in many pathogens (including Toxoplasma), but are completely lacking in mammals, and therefore potential drug targets. Evolutionary conservation of the thymidylate synthase (TS) protein sequence argues that pathogen-specific TS inhibitors may be difficult to develop, but analysis of dihydrofolate reductase (DHFR[9]) and hypoxanthine-xanthine- guanine phosphoribosyl transferase (HXGPRT[10]) genes indicates significant divergence from the human homologs. Recently, examination of ribosomal genes has identified an entirely unexpected category of protein synthesis machinery – more similar in organization to chloroplasts than nuclear ribosomes – in malaria, Toxoplasma, and related parasites, providing another promising target for drug design[11, 12].

3. Multiple sequence alignments facilitate the identification of functional homologies, leading to better drugs through structural analysis. Because of the intense pharmacological and biochemical interest in chemotherapeutic targets and rational drug design, the genes encoding many key metabolic enzymes have been cloned and sequenced from pathogenic organisms spanning a broad phylogenetic spectrum. In combination with structural studies, multiple sequence alignments can play a major role in structure-based drug design. In the case of T. gondii, for example, such studies are currently underway for all of the enzymes noted above. Coupled with site-directed mutagenesis and functional testing in recombinant expression systems and transgenic

parasites[13],modeling studies on DHFR have provided explanations for the unusual sensitivity of malaria to cycloguanil, and for the unfortunate ease with which malaria parasites develop antifolate resistance, while drug resistance in Toxoplasma is less common (M. Reynolds and D.S. Roos, unpublished). Studies on HXGPRT suggest why the parasite enzyme – but not the human homolog – can phosphoribosylate xanthine in addition to hypoxanthine and guanine, providing a window for structure-based drug design (D. Carter, B. Ullman & D.S. Roos, unpublished). Similarly, examination of the plastid-like ribosomal genes of Toxoplasma may explain observed patterns of drug sensitivity and identify new targets for intervention[12].

In addition to the practical examples relevant to drug design noted above, phylogenetic analysis of protozoan parasites such as T. gondii is central to many important biological concerns beyond the scope of this brief discussion: What does it mean to be a parasite? How frequently, and in what evolutionary contexts, has this strategy been adopted? What are the evolutionary relationships among protozoa?[14] What is/are the origin(s) of 'higher' eukaryotes?[15] What are the origins and functions of endosymbiotic organelles?[11, 12, 16]

Practical difficulties – What we need (and what we don't).

What sorts of computational tools would facilitate phylogenetic analysis? Given the daunting number of tree topologies possible when considering even a tiny fraction of known biological species, improved computing power and the impressive recent advances in efficient algorithm design[17] will certainly be put to good use. It is useful to keep in mind, however, that with few exceptions (such as ribosomal sequences), the numbers of taxa available in a given dataset is likely to be relatively small: dozens rather than hundreds or thousands. DHFR, for example, is an important drug target in many pathogenic organisms and one of the most intensively studied of all enzymes, yet fewer than 50 sequences are known to date, with new entries appearing at the rate of only 5/yr for the last several years. Approximately 20 sequences are available for HXGPRT, 10 for DHPS, and only 4 for UPRT.

Even more importantly, many boundaries on the topology of a complete 'tree of life' have been established unequivocally through classical morphometric and developmental systematics, and confirmed by molecular analyses. For example, we no longer need to consider all possible topologies involving snakes, chickens, bats, and mice; it has been abundantly demonstrated that the former two and the latter two each form sister pairings, with the evolutionary root lying between these two groups. If we seek to add crocodiles to the tree, we need consider only three possible topologies (those in which crocodilians are sisters of snakes, chickens, or deep to the snake/chicken pairing) – not all of the 90 possible rooted, bifurcating trees for five taxa. Thus, it is not necessarily the case that the number of plausible tree topologies must explode with increasing numbers of taxa. In most cases, biologically interesting phylogeny reconstruction problems can safely be broken down into questions which are sufficiently restricted to permit consideration of all possible tree topologies.

Perhaps the greatest problem facing molecular phylogenetics at present is not the tree construction itself, but the quality of available datasets. Multiple sequence alignment is something of a black art, where the results of computational alignment algorithms[17] are manipulated by hand to provide a pleasing result. The need for such manipulation reflects the complex nature of the underlying sequence, and our poor understanding of both biological function and evolutionary mechanisms and constraints. To cite just a few difficulties applicable to amino acid sequences (analogous problems apply to nucleic acid data):

- Important features relevant to the conservation of individual loci (characters) are defined in terms of the various physical parameters which characterize amino acids: charge, shape, size, sensitivity to pH, etc. Moreover, relevant features for cross-species comparison often involve not just single loci, but complex strings of (not necessarily contiguous) amino acids which make up secondary or tertiary structural elements. Because the functional significance of specific amino acids differs between loci within a given protein (and sometimes even at the same locus, when comparing across taxa), certain loci are likely to be absolutely conserved, while others may be relatively free to diverge. It is therefore impossible to encode a universally applicable set of prior constraints for sequence alignment. Amino-acid substitution tables[19] provide, at best, only an approximation of the relevant substitution probabilities.
- Different regions within the protein under consideration may be structurally interchangable; thus a primary sequence alignment which is correct with respect to structure and function may not accurately reflect evolutionary homology.
- Many different biological mechanisms are capable of generating mutational change, but even aside from the difficulty in assessing the functional implications of such mutations (see above), we know little of the relative frequencies for the various different types of mutations possible (with a few exceptions, such as transition/transversion ratios[20]). This makes it difficult to assign penalties for, say, deletions vs. point mutations, in generating multiple sequence alignments.

These complications make it clear that there will always be significant obstacles to inferring a historically correct evolutionary tree from biological datasets. To minimize these difficulties, perhaps the greatest need is for the development of a practical user interface which recognizes the reality of multiple sequence alignment. Such an interface would provide visual displays of characters (in terms of charge, size, secondary structure predictions, hydrophobicity, conservation in other taxa, Dayhoff substitution probabilities, etc) based on simple toggle keys, while also permitting statistically-based, computer-assisted alignment (at both the protein level and for the underlying DNA sequence) in regions where desirable. Some progress has been made in this direction[21], but much work still remains to be done.

One particularly appealing aspect of such an interface would be the possibility of coupling the editor with machine learning software. This approach could yield improve automated alignment strategies, and might even serve to identify interesting patterns reflecting the underlying biological mechanisms responsible for gene/protein design.

Returning to the computational problems of trees themselves (as opposed to multiple sequence alignments), perhaps the greatest practical need is statistical rather than algorithmic. How confident can we be of the trees generated by molecular phylogenetic analysis? Although plausible trees can be produced using any of several algorithms, it is difficult to determine the reliability of such trees[4] – what do values obtained from bootstrap analyses, decay indices, log-likelihood measurements, etc really mean in biological terms?

A related question: when are the various different trees obtained using different datasets (or different algorithms) significantly different? This is not necessarily a matter of identifying the one true tree based on analysis of inherently noisy datasets – in some cases two (or more) distinct trees may both be correct. For example, while the cytoplasmic ribosomes of maize are more closely related to human cytoplasmic ribosomes than either is to their bacterial counterparts, maize chloroplast ribosomes are most closely related to bacteria, indicating lateral transfer of genetic information (in support of an endosymbiotic origin of plastids[16]). Analysis of cytoskeletal (and several other) nuclear genes from eukaryotic cells routinely yields significantly different tree topologies than phylogenies based on ribosomal genes. Is this a matter of noisy data (and if so, do the similar trees produced using different datasets reflect important biological algorithms), or are both classes of trees correct, indicating lateral gene transfer[23]?

The explosion of information now being deposited in biological databases shows no sign of abating, bringing phylogenetic questions to the fore in many research laboratories which have not previously dealt with such issues. To maximally exploit this environment, it is imperative that researchers at the interface between computer science and biology make every effort at communication, and that tools developed for this research (computer software, biological databases, nucleic acid sequencing strategies, etc) remain accessible even to occasional users – not just those immersed in their development. While challenging, the potential payoffs, in areas ranging from improved antimicrobial drug design to deepening our understanding of the origins of life, are tremendous.

Bibliography

[1] McLeod, R. and J.S. Remington (1987) In Harrison's Principles of Internal Medicine, 11th ed. E. Braunwald et al., editors. McGraw-Hill, New York. pp. 791.

[2] Luft, B.J. and J.S. Remington (1992) Clin. Infect. Dis. 15:211-222.

[3] While humans require folic acid as a vitamin, most protozoa synthesize their own folate, in a complex reaction catalyzed by dihydropteroate synthase (DHPS), the target of sulfa-drugs. Pyrimethamine acts as a competitive inhibitor of dihydrofolate reductase (DHFR), an enzyme whose product is essential for the synthesis of pyrimidine nucleotides and many other essential metabolic building blocks. In contrast to methotrexate (a common anti-cancer drug), pyrimethamine is selective for the parasite enzyme.

[4] Because of the effectiveness of the immune response against Toxoplasma, secondary infection during pregnancy is not a cause for concern; transplacental transfer has never been documented in a previously seropositive healthy woman. However, women who are seronegative prior to pregnancy are well advised to avoid eating rare meat (a possible source of T. gondii tissue cysts, as well as other parasites); avoid contact with kitty litter (a possible source of oocysts, as the parasite sexual cycle occurs only in felines); and wear gloves while farming or gardening, as soil may be contaminated with cat fecal matter.

[5] Gajadhar, A.A., W.C. Marquardt, R. Hall, J. Gunderson, E.V. Ariztia- Carmona and M.L. Sogin (1991) Molec. Biochem. Parasitol. 45:147-154. Gagnon, S., R.C. Levesque, M.L. Sogin, and A.A. Gajadhar (1993) Molec. Biochem. Parasitol. 60:145-148.

[6] Roos, D.S, R.G.K. Donald, N.S. Morrissette, and A.L.C. Moulton (1994) Methods Cell Biol. 45:23-65.

[7] Boothroyd, J.C., K. Kim, E.R. Pfefferkorn, L.D. Sibley, and D. Soldati (1995) Methods Molec. Genet. 6:1-29. Donald, R.G.K., and D.R. Roos (1995) Proc. Nat'l Acad. Sci. USA 92:in press.

[8] Edman, J.C., J.A. Kovacs, H. Masur, D.V. Santi, H.J. Elwood and M.L. Sogin (1988) Nature 334:519-522.

[9] Roos, D.S. (1993) J. Biol. Chem. 268:6269-6280.

[10] Ullman, B., and T.E. Allen (1995) In Molecular Approaches to Parasitology. J.C. Boothroyd and R. Komuniecki, eds. Wiley-Liss, New York. pp. 123-141.

[11] Biochem. Parasitol. 44:115-124. Williamson, D.H., M.J. Gardner, P. Preiser, D.J. Moore, K. Rangachari and R.J.M. Wilson (1994) Molec. Gen. Genet. 243:249-252.

[12] Beckers, C.J.M, D.S. Roos, R.G.K. Donald, B.J. Luft, J.C. Schwab, Y. Cao and K.A. Joiner (1995) J. Clin. Invest. 95:367-376.

[13] Donald, R.G.K. and D.S. Roos (1993) Proc. Nat'l Acad. Sci. USA 90:11703-11707. Donald, R.G.K. and D.S. Roos (1994) Molec. Biochem. Parasitol. 63:243-253.

[14] Cavalier-Smith, T. (1993) Microbiol. Rev. 57:953-994.

[15] Wainright, P.O., G. Hinkle, M.L. Sogin and S.K. Stickel (1993) Science 260:340-342. Hasegawa, M., T. Hashimoto, J. Adachi, N. Iwabe and T. Miyata. (1993) J. Molec. Evol. 36:380-388. Baldauf, S.L., and J.D. Palmer Proc. Nat'l Acad. Sci. USA 90:11558-11562.

[16] Gray, M.W. (1992) Int. Rev. Cytol. 141:233-357. Morden, C.W., C.F. Delwiche, M. Kuhsel, and J.D. Palmer (1992) BioSystems 28:75-90. Cavalier-Smith, T. (1992) BioSystems 28:91-106. Margulis, L. (1992) BioSystems 28:107-108. Whatley, J.M. (1993) Int. Rev. Cytol. 144:259- 299. Delwiche, C.F., M. Kuhsel and J.D. Palmer (1995) Molec. Phylogenetics Evol. 4:in press.

[17] Farach, M., S. Kannan and T. Warnow (1995) Algorithmica 13:155-179.

[18] Feng, D.F., and R.F. Doolittle (1987) J. Molec. Evol. 35:351-360. Higgins, D.G., and P.M. Sharp (1989) CABIOS 5:151-153. Thorne, J.L., H. Kishino and J. Felsenstein (1991) J.Molec.Evol. 33:114-124.

[19] Dayhoff, M.O., R.M. Schwartz and B.C. Orcutt (1979) In Atlas of Protein Sequence and Structure. M.O. Dayhoff, editor. National Biomedical Research Foundation, Washington, DC.

[20] Kimura, M. (1980) J. Molec. Evol. 16:111-120.

[21] Schuler, G.D., S.F. Altschul and D.J. Lipman (1991) Proteins 9:180- 190. Maddison, W.P., and D.R. Maddison (1992) MacClade 3.0. Sinauer Associates, Sunderland, MA.

[22] Felsenstein, J. (1981) J. Molec. Evol. 17:368-376. Felsenstein, J. (1985) Evolution 39:783-791. Kishino, H., and M. Hasegawa (1989) J. Molec. Evol. 29:170. Faith, D., and P. Cranston (1991) Cladistics 7:1-28. Hillis, D.M., and J.J. Bull (1993) Syst. Biol. 42:182-192. Felsenstein, J. and H.Kishino (1993) Syst. Biol. 42:193-200. Hillis, D.M. (1995) Syst. Biol. 44:3-16.

[23] Lake, J.A., and M.C. Rivera (1994) Proc. Nat'l Acad. Sci. USA 91:2880-2881.

On the models of nucleotide substitution applicable to ribosomal RNA genes

Andrey Rzhetsky
Institute of Molecular Evolutionary Genetics and
Department of Biology
Pennsylvania State University
University Park, Pennsylvania 16802
AUR1@PSUVM.PSU.EDU

A molecular systematist who is interested in the inference of phylogenetic relationships within particular set of taxa typically goes through the following steps.

1. Selecting appropriate genes. That is, one should choose genes that change fast if the taxa of interest are expected to be closely related, and choose slowly evolving genes to infer phylogeny from distantly related species.

2. Sequencing new genes.

3. Aligning newly obtained gene sequences with those already published.

4. Selecting a model of nucleotide substitution that fits the data.

5. Reconstructing a tree and evaluating reliability of this tree.

Each of these steps is associated with a number of practically important problems. The initial plan of this talk included discussion of the last two steps of phylogenetic analysis with respect to the distance-matrix methods, but later on it was changed to give more room to the issues related to the models of nucleotide substitution using ribosomal RNA (rRNA) data as an example. There are many models of nucleotide substitution (for review see Tavaré[12]; Nei[6]; Rodrigues 1991) that can be applied to analysis of rRNA genes. However, none of these models was designed specifically to match peculiar properties of these genes. The most prominent characteristic of all rRNAs is their highly conserved secondary structure defined by complementary regions within the same molecule (Gutell[2]). A mature rRNA molecule includes both double-stranded and single-stranded regions. Some sites of rRNA genes are functionally constrained and are virtually invariable, whereas other sites accept substitutions and their functional constraints seem to be less demanding ([2]).The model considered here describes nucleotide substitution in the double-stranded and the single-stranded regions by Markov chains in continuous time with sixteen and four discrete states, respectively (see Schoniger and von Haeseler[11]; Tillier[13]; Muse[5]). Heterogeneity of substitution rate among different nucleotide sites is taken into account by assuming that the rate of the Markov process in a site in a single-stranded region (or in a pair of sites in a double-stranded region) is a random variable following a gamma distribution (e.g., see [1]; [7]; [3]; Yang 1993). In the simplest case when the intensity matrix for the sixteen-state Markov process includes just two parameters, it is relatively easy to obtain explicit transition probabilities and an estimator for "evolutionary distance" between double-stranded regions of present-day rRNAs[9].

The next part of the talk concerned testing of various components of a model of nucleotide substitution with regard to particular data set. Some of such tests can be conducted in the absence of knowledge about true phylogenetic tree underlying the present-day sequences. The first example is a test of stationarity of base composition in a set of present-day genes (see [12] for discussion of tests for two sequences). Stationarity of a Markov process implies that the probability to find nucleotide site in a particular state does not change with time, and the expected frequencies of nucleotides in the present-day sequences are the same. One difficulty in testing equality of the observed base frequencies is that they are phylogenetically correlated and the true tree is unknown. It is nevertheless possible to compute dispersion matrix for the observed frequencies without explicit information about the tree[10]. Using this dispersion matrix one can construct a statistic following a chi-square distribution with (3 m) degrees of freedom under hypothesis of stationarity of the Markov process, where m is the number of sequences. Application of this test to real rRNA data has shown that the stationarity assumption is acceptable when genes from very closely related taxa are considered but is almost always violated when genes from distant groups are compared.

For a certain set of simple nucleotide substitution models, one can also construct statistics for testing the fit of a model to a particular data set before phylogenetic tree is reconstructed[10]. Corresponding tests can be constructed by analogy with Wald's large sample test for composite hypothesis (see Rao[8]; pp. 417-419). The main advantage of this test is that in comparing two nested models this test requires only the maximum likelihood estimates of parameters under the more general model. If the general model is formulated in terms of the expected frequencies of nucleotide configurations occupying homologous sites in the present-day sequences (rather than in terms of the entries of the intensity matrix) estimation of the parameters clearly does not require maximization of the likelihood function. In addition, if it is feasible to find explicit relationships among parameters of the general model under the simpler model, a test statistic based on relationships of these parameters can be constructed easily. For example, to construct a statistic for testing Jukes and Cantor's[4] model against Kimura's (1980) model for a pair of nucleotide sequences, we can reformulate Kimura's model in terms of parameters P and Q (the expected proportions of transitional and transversional differences between the sequences, respectively). We can then find that under Jukes and Cantor's model $2P = Q$ and the required Wald's statistic for a pair of sequences is $x = (2p - q)/Var(2p - q)$ where p and q are estimates of P and Q, respectively. Statistic x approximately follows a chi-square distribution with one degree of freedom. This approach can be generalized for a larger number of sequences in the following way. It is impractical to consider restrictions on m-nucleotide patterns (for a set of m sequences, m ¿ 3) because (i) the expected frequencies of these m-patterns depend on the true tree topology, and (ii) the real genes are too short to provide sufficient information for testing restrictions on the frequencies of these patterns. It seems to be more reasonable to combine statistics for pairs of sequences into a composite statistic (examples are given in Rzhetsky and Nei[10]). Although the power of this test should be somewhat lower than power of a more rigorous likelihood-ratio test, it tends to reject simple models of nucleotide substitution (e.g., Jukes and Cantor's[4] and Kimura's models) in most cases when applied to real data. Some features of nucleotide substitution in rRNA genes that are not taken into account in any of the existing models were also discussed.

Bibliography

[1] Golding, G.B., 1983 Estimates of DNA and protein sequence divergence: an examination of some assumptions. Mol. Biol. Evol. 1: 125-142.

[2] Gutell, R.R., 1992 Evolutionary characteristics of 16S and 23S rRNA structures. In: H. Hartman and K. Matsuno (eds). The origin and evolution of the cell. World Scientific, New Jersey, pp. 243-309.

[3] Jin, L., and M. Nei, 1990 Limitation of the evolutionary parsimony method of phylogenetic analysis. Mol. Biol. Evol. 7: 82-102.

[4] Jukes, T.H., and C.R. Cantor, 1969 Evolution of protein molecules. pp. 21-132 In: Mammalian Protein Metabolism, vol. 3. Edited by H.N. Munro. Academic Press, New York.

[5] Muse, S.V., 1995 Evolutionary analysis of DNA sequences subject to constraints on secondary structure. Genetics 139: 1429-1439.

[6] Nei, M., 1987 Molecular evolutionary genetics. Columbia University Press. New York.

[7] Nei, M., and T. Gojobori, 1986 Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. 3:418-426.

[8] Rao, C.R., 1973 Linear statistical inference and its applications. Second edition. Wiley, New York.

[9] Rzhetsky, A., 1995 Estimating evolutionary distance between rRNA genes. (In preparation).

[10] Rzhetsky, A., and M. Nei, 1994 Testing applicability of several substitution models for DNA sequence data. Mol. Biol. Evol. 12: 131-151.

[11] Schoniger, M., and A. von Haeseler, 1994 A stochastic model for the evolution of autocorrelated DNA sequences. Mol. Phyl. Evol. 3: 240-247.

[12] Tavaré, S. 1986 Some probabilistic and statistical problems in the analysis of DNA sequences. Lectures on mathematics in the life sciences. 17: 57-86.

[13] Tillier, E.R.M., 1994 Maximum likelihood with multiparameter models of substitution. J. Mol. Evol. 39: 409-417.

Reconstructing evolutionary trees under a variety of Markov-style models

Mike Steel

Department of Mathematics and Statistics
University of Canterbury
Christchurch, New Zealand
M.Steel@math.canterbury.ac.nz

Statistically-based methods for reconstructing phylogenetic trees from DNA sequence data are, typically, based on very simple and often inadequate models of nucleotide mutation. It is usually assumed that sites evolve independently and identically according to a Markov process on a tree. Furthermore, this Markov process is usually highly constrained, for instance it is often taken to be stationary, and reversible, with restrictions on the form of the transition matrices (Rodriguez *et al.* 1990[11]). Although this is a nice foundation for developing methods, it is clear that many real sequences do not evolve according to such simple schemes. For instance, sequences of reasonable length that exhibit between them markedly different biases in their nucleotide frequencies are unlikely to evolve according to **any** stationary Markov process, and thus tree reconstruction methods based on such models, are liable to mislead (Lockhart *et al.* 1994 [8]). Such considerations motivated the development of methods based on more general (but still iid-based) models (Steel, 1994[12]; Lake 1994[7]).

Independence between sites may be compromised by numerous factors, including functional constraints, the genetic code, secondary/tertiary structure, the biochemical details of repair mechanisms. A simple, though computationally demanding, approach to incorporate dependency between small numbers of sites is to enlarge the state space on which the Markov process operates - for instance, modelling of codon evolution using 256 states (Muse and Gaut, 1994[10]; Goldman and Yang, 1994 [4]). Sites in most sequences also do not evolve identically - there may be sites which cannot change, due to functional constraints, and other sites that mutate quickly; thus some recent studies (Yang, 1993[15]; Steel *et al.* 1994[12]) have incorporated a distribution of rates across sites (where a rate parameter for each site is chosen from some distribution). We discuss this "mixed model" shortly.

Introducing more realistic models, however, raises the possibility that the tree might become "lost" in the data in the sense that by incorporating enough parameters, one might fit the data perfectly to **every** tree. More precisely, for sites evolving independently according to a Markov tree, let $\mathbf{f} = \mathbf{f}(T, M)$ denote the vector of probabilities of generating each of the possible 4^n patterns possible at a site in n aligned sequences, a function of the underlying evolutionary tree T and the stochastic model M (encompasing the parameters associated with the edges of T and the distribution of rates across sites, etc). It is desirable to determine under what class of models $\mathbf{f}(T, M)$ determines T - that is, identifies the tree T, but not necessarily the position of the root or the parameters in M. In that case, sufficient data (independently evolving sites) generated under this class of models will allow, in principle, the unique recovery of T. However if different trees can produce the same vector $\mathbf{f}$ for suitable choices of M from some class, then this class cannot be used to develop a statistically consistent, well-defined tree reconstruction method.

In the case of iid site evolution, it turns out that, except for certain degenerate cases (where the transition matrices on the edges of T are either permutation matrices or singular), T is always determined uniquely from $\mathbf{f}$ (up to the position of the root). This fundamental result was first stated by Chang and Hartigan (1991)[2]; for a proof see Steel (1994)[12] or Lake (1994)[7]. The proof is quite constructive, and based on forming the logarithm of the determinant of the joint distribution on patterns for all pairs of sequences, which turns out to be additive on the true tree. The resulting "LogDet" method (also called "Paralinear distances" by Lake) has been useful in studying sequences for which traditional methods appear to be mislead - see for instance Lockhart *et al.*[8] (1994) and Lake (1994)[7].

However this pleasant uniqueness property is destroyed when there is an unknown distribution of rates across sites (the mixture model). For example, suppose each site evolves according to a simple process, such as the symmetric 2-state model due to Cavender and Farris, or the Kimura 3ST model. Further, suppose that each site has a rate parameter λ which scales the tree "geometrically" - that is, the expected number of changes on each edge at that site is multiplied by λ. For convenience assume that λ is drawn independently from a distribution ϑ. Then, if nothing is known about ϑ it can be impossible to uniquely recover T from $\mathbf{f}$ - indeed one can say more - for each tree one can construct an associated distribution of rates across sites, and choice of edge parameters so as to induce exactly the same $\mathbf{f}$ vector. These results, which are described in more detail, along with proofs, in Steel *et al.* (1994)[12] complement recent results of Chang (1995)[1] who has shown that many standard methods can become inconsistent under this type of mixture model.

However if ϑ is known exactly, or if other restrictions are applied, for instance a molecular clock, or the condition that a (unknown) proportion of sites cannot change state, but the remainder evolve iid, then T can be uniquely recovered from $\mathbf{f}$.

The proof of these and the previous result follows from a remarkable description of $\mathbf{f}$ which applies when sequence sites evolve according to a random walk on an Abelian group. In the biologically-relevant cases where this group is Z_2 (the Cavender-Farris model) or $Z_2 \times Z_2$ (a generalisation of the Kimura 3ST model) this representation takes the form $\mathbf{f} = H^{-1}M\,[H\gamma]$, where M is the moment generating function for ϑ, H is a Hadamard matrix which is independent of T, and γ is a vector which encodes the tree and its edge parameters (for further details see Steel *et al.* 1994)[13]. One of the attractive features of this representation is that it is invertible ($\gamma = H^{-1}M^{-1}[H\mathbf{f}]$) providing a technique for reconstructing trees by applying a tree-independent transformation (viz. $\mathbf{x} \rightarrow H^{-1}M^{-1}[H\mathbf{x}]$) to the observed frequencies of patterns in the data (for applications of this technique to real data, see Hendy *et al.* 1994). However this representation is also useful as a tool for exploring, analytically, properties of other tree-reconstruction methods and the space of $\mathbf{f}$ vectors that can arise under these models. For example, the fact that most entries in γ must be zero provides phylogenetic invariants, that essentially characterise this space (see Steel *et al.* 1993)[14].

A recent application of this Hadamard representation is a characterization of the linear phylogenetic invariants arising under the Jukes-Cantor model. In particular, the dimension of the space of linear invariants for **any** binary tree with n leaves is precisely $4^{n-1} - F_{2n-2}$, where F_k is the k-th Fibonacci number, and a basis can be compactly described. Furthermore, one can construct, for each binary tree T an associated phylogenetic invariant which is strictly positive for the $\mathbf{f}$ vector given by any other tree - a result which should allow for one-sided rather than two-sided statistical tests. Details and proofs appear in Steel and Fu, 1995[3]. Mike Hendy (1995)[5] has recently extended this approach to analyse linear invariants under other submodels of the Kimura 3ST model, both with and without a molecular clock.

One can also consider the related question of when the collection $\mathbf{d} = [d_{ij}]$ of expected pairwise dissimilarities between sequences suffice to determine a tree. If, for example, the underlying Markov process is reversible, and sites evolve with a known distribution of rates across sites, then the tree is determined by $\mathbf{d}$, since d_{ij} as a monotone increasing function of the expected number E_{ij} of substitution separating i and j, and $E = [E_{ij}]$ is additive on the true tree. We have:

$$d_{ij} = d_\infty \; [1 - \sum_{k=1}^{3} \; \pi_k \; M \, (\beta_k \; E_{ij})]$$

where β_k are (proportional to) the strictly negative eigenvalues of the rate matrix (of the reversible Markov process), the π_i are non-negative and sum to 1, $d_\infty = 1 - \sum_{i=1}^{4} v_i^2$ where $v = [v_i]$ is the stationary distribution of the Markov process, and, as before, M is the mgf of the distribution of rates across sites. For certain covarion-style models (Miyamoto and Fitch, 1995)[9] one can obtain an analogous formula. Details will appear elsewhere (Steel, in preparation).

Bibliography

[1] Chang, J.T. (1995), Inconsistency of evolutionary tree topology reconstruction methods when susbstitution rates vary across characters, manuscript.

[2] Chang, J. and Hartigan, B. (1991), Reconstruction of evolutionary trees from pairwise distributions on current species, *Interface*, 254-257.

[3] Fu, Y.X. and Steel, M.A. (1995), Classifying and counting linear phylogenetic invariants for the Jukes Cantor model, *J.Comp. Biol.* (in press).

[4] Goldman, N. and Yang, Z. (1994), A codon-based model of nucleotide substitution for protein-coding DNA sequences, *Mol. Biol. Evol.* **11(5):** 725-736.

[5] Hendy, M.D. (1995), Complete families of invariants for some stochastic models of sequence evolution, with and without the molecular clock assumption (manuscript).

[6] Hendy, M.D., Penny D. and Steel, M.A. (1994), A discrete Fourier analysis for evolutionary trees, *Proc. Natl. Acad. Sci. USA,* **91:** 3339-3343.

[7] Lake, J.A. (1994), Reconstructing evolutionary trees from DNA and protein sequences: Paralinear distances. *Proc. Natl. Acad. Sci. USA* **91:** 1455-1459.

[8] Lockhart, P.J., Steel, M.A., Hendy, M.D. and Penny, D (1994), Recovering evolutionary trees under a more realistic model of sequence evolution. *Mol.Biol.Evol* **11(4)**: 605-612

[9] Miyamoto, M. M. and Fitch, W.M. (1995), Testing the covarion hypothesis of molecular evolution. *Mol. Biol. Evol.* (in press).

[10] Muse, S.V. and Gaut. B.S. (1994), A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates with application to the chloroplast genome, *Mol. Biol. Evol.* **11(5):** 715-724.

[11] Rodriguez, F., Oliver, J.L., Marin, A. and Medina, J.R. (1990), The general stochastic model of nucleotide substitution. *J. Theor. Biol.* **142:** 485-501.

[12] Steel, M.A. (1994), Recovering a tree from the leaf colorations it generates under a Markov model, *Appl. Math. Lett.* **7(2):** 19-24.

[13] Steel, M.A. Székely, L. and Hendy, M.D. (1994), Reconstructing trees from sequences whose sites evolve at variable rates. *J. Comp. Biol.* **1(2):** 153-163.

[14] Steel, M.A. Székely, L. Erdös, P. and Waddell, P. (1993), A complet e family of phylogenetic invariants for any number of taxa under Kimura's 3ST model. *NZ Journal of Botany* (conference proceedings), **13(3):** 289 - 296.

[15] Yang, Z. (1993), Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.* **10(6):** 1396-1401.

Reconstructing phylogenetic trees when sites are dependent

Simon Tavaré and Yinsuo Feng
Department of Mathematics
University of Southern California
Los Angeles, CA 90089-1113
stavare@gnome.usc.edu

1. 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.

2. 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 $R = (r(i, j))$.

3. 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 $\alpha \rightarrow 1$, 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. citeFel1), 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.

4. 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.

Bibliography

[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.

Inference of the conditional independence structure of dependent binary data

E. A. Thompson
Department of Statistics
University of Washington
Seattle, WA 98195
thompson@stat.washington.edu

The autologistic model for dependent binary data (Besag, 1994) has recently been used in several genetic problems to illuminate the pattern of dependence. These include:

- the detection of interference, or dependence in the presence/absence of crossover events occurring in different chromosome segments within a gamete (Zhao et al., 1990),
- the estimation of genealogical relationships, or dependence in the presence/absence of multilocus DNA fingerprint bands in different California condors (Geyer et al. 1994), and
- the inference of relationshops among populations, or dependence in the presence/absence of variant alleles in different Chibcha-speaking tribes of Panama and Costa Rica (Thompson at al., 1992).

Here we describe two additional genetic applications of the model, in which the objective is to infer patterns of conditional independence.

The data are

$$\mathbf{X} = \{X_{ij};\ i = 1, ..., m;\ j = 1, ..., n\} \tag{1}$$

and the objective is to infer dependence between m units i on the basis of n independent binary factors j. In these examples we take each X_{ij} to be ± 1. If complete data, $\mathbf{X}$, are available, the autologistic model likelihood takes the exponential family form

$$\begin{aligned} P(\mathbf{X}; \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) &= (c(\theta))^{-1} \exp(\sum_i \theta_i T_i(\mathbf{X})) \\ &= (c(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}))^{-1} \exp(\sum_i \alpha_i U_i + \sum_j \beta_j V_j + \sum \sum_{i<k} \gamma_{ik} S_{ik}) \end{aligned} \tag{2}$$

where the natural sufficient statistics are

$$\begin{aligned} \mathbf{T} &= (U_1, ..., U_m, V_1, ..., V_n, \qquad S_{1,2}, ..., S_{m-1,m}) \\ U_i &= \sum_j X_{ij},\ V_j = \sum_i X_{ij},\ S_{ik} = \sum \sum_{i<k} X_{ij} X_{kj} \end{aligned} \tag{3}$$

with corresponding natural parameters

$$\theta = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$$

and the normalizing constant

$$c(\theta) = \sum_{\mathbf{X}} \exp(\sum_i \theta_i T_i(\mathbf{X})) \tag{4}$$

If only data $\mathbf{Y}$, a subset of $\mathbf{X}$, are available, the likelihood is of the form

$$L(\theta) = P(\mathbf{Y}; \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{\mathbf{X} \mapsto \mathbf{Y}} P(\mathbf{X}; \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = c_{\mathbf{Y}}(\theta)(c(\theta))^{-1} \tag{5}$$

where $c_{\mathbf{Y}}(\theta)$ is analogous to $c(\theta)$, being the sum of the same exponential expression over the smaller set of $\mathbf{X}$-values corresponding to $\mathbf{Y}$.

The name "autologistic" derives from the fact that the probability for each X_{ij} conditionally on all the remainder has a logistic form. If $\gamma_{ik} = 0$, X_{ij} is conditionally independent of X_{kj}, given the remaining X_{lj}. In our applications we are interested only in positive dependence, and thus likelihood maximisation is carried out subject to the constraints $\gamma_{ik} \geq 0$. Due to the exponential form, maximisation is straightforward, in principle. The score vector is

$$\frac{\partial \log L(\theta)}{\partial \theta} = \mathrm{E}_\theta(\mathbf{T}(\mathbf{X}) \mid \mathbf{Y}) \; - \; \mathrm{E}_\theta(\mathbf{T}(\mathbf{X}))$$

and the matrix of log-likelihood second derivatives is

$$\frac{\partial^2 \log L(\theta)}{\partial \theta^2} \; = \; var_\theta(\mathbf{T}(\mathbf{X}) \mid \mathbf{Y}) \; - \; var_\theta(\mathbf{T}(\mathbf{X}))$$

while the log-likelihood difference between any two hypotheses θ and ψ is easily computed as

$$\log\left(\frac{L(\theta)}{L(\psi)}\right) \; = \; \log\left(\frac{c_{\mathbf{Y}}(\theta)}{c_{\mathbf{Y}}(\psi)}\right) \; - \; \log\left(\frac{c(\theta)}{c(\psi)}\right) \tag{6}$$

The only difficulty in likelihood analysis is in the evaluation of the normalising function (4); this requires summation of $n2^m$ terms, and is impractical if $m \geq 15$. However, in this case Monte Carlo likelihood analysis using a Metropolis Markov chain Monte Carlo sampler is possible: this approach was developed by Geyer and Thompson (1992).

Our first application is the ordering of multiple loci, on the basis of grandparental origins of genes in independent gametes. Gametes are scored with $X_{ij} = +1$ if the gene at locus i in segregation j is grandpaternal, and $X_{ij} = -1$ if it is grandmaternal. If there are few missing data, minimisation of obligatory recombinants is feasible, and for closely linked loci provides the maximum likelihood solution (Thompson, 1987). However, with more missing data, there is no easy way to avoid considering all alternative locus orders. The problem fits neatly into the autologistic model framework (5). Due to symmetry between the two grandparental origins, we may take $\boldsymbol{\alpha} = \boldsymbol{\beta} = 0$. Linkage creates only positive dependence; $\gamma_{ik} \geq 0$. In the absence of interference, grandparental origins at different loci are independent, conditional on the origins at any intervening locus; thus only γ_{ik} between adjacent loci should be positive. Another problem in ordering multiple tightly linked loci is that there may be insufficient recombination events occurring for all pairs of loci, (i, k), to be resolved; with no missing data, such complete dependence in grandparental origins will be obvious ($X_{ij} = X_{kj}$ for all j), but with missing data conditions leading to infinite maximum likelihood estimates of γ_{ik} are less clear. Alternative methods for resolving this "Phase I" problem were considered by Geyer and Thompson (1992), but here we use the simple trick of adding a single null gamete j_0 with $X_{ij_0} = 0$ for all i. This is equivalent to an $n/(n+1)$ shrinkage factor in the likelihood estimation.

The method outlined was applied to a set of simulated data on 8 linked loci, with 18 informative gametes, in which 25% of the observations (chosen at random) were missing. Two of the loci were unresolved by the data. It was indeed found that only γ_{ik} values between adjacent loci were substantially non-zero. All these were at least 0.35, and most other estimates were 0, with the 6 non-zero values all less than 0.1. The order of the loci (with the exception of the unresolved pair) could thus be easily written down from the maximum likelihood estimates of the γ_{ik}. Maximisation was repeated, constraining all γ-values between non-adjacent loci to zero. The log-likelihood difference between the two hypotheses was small (0.5). The "wrong" non-zero estimates occurred primarily at the ends of the ordering; they are "edge effects". One solution is to add a locus at which no observations are made to each end of an inferred order, and to re-estimate the dependence parameters.

A second illustrative example more directly related to phylogenies is to mtDNA D-loop sequence variants in a sample of 63 Nuu-Chah-Nulth Amerindians (Griffiths and Tavaré, 1994). There are 14 mitochondrial sequences, characterised by 18 variable sites. Since there are no transversions, the data can be scored as binary; A and T were scored as $+1$, G and T as -1, with 5 sites having A or G, and 13 having T or C. In this example there are no missing data. For convenience, we again considered the model with $\boldsymbol{\alpha} = \boldsymbol{\beta} = 0$. Again dependence will be non-negative, and, conditional on an extant ancestral type, descendant variant sequences should be independent. Estimating the γ_{ik}, the tree structure was clearly illuminated. The 13 estimates corresponding to branches of the tree given by Griffiths and Tavare (1994) ranged from 0.6 to 0.9. Only 16 of the remaining 78 were non-zero, and all were less than 0.3. In this example data set, ancestral types appear to be extant, but if the parameter estimates indicate possible missing ancestral types, these may be added, without data, once the preliminary tree structure is estimated, as with the addition of the flanking loci in the locus ordering problem.

Likelihood analysis in inference of locus order, in phylogeny estimation, and in other problems of relationship estimation, encounters difficulties due to the large discrete space of alternative hypotheses. The autologistic model provides a tool for the analysis of these problems, in which the structure is determined from the maximum likelihood parameter estimates, without the need to consider the discrete alternatives *a priori*.

Acknowledgement

I am grateful to Charles Geyer and Simon Tavaré for discussions. Work supported in part by NSF grant BIR 9305835.

Bibliography

[1] Besag J. (1974) Spatial interaction and the statistical analysis of lattice systems (with Discussion). *J. R. Statist. Soc. B* **36:** 192–236.

[2] Geyer, C. J. and E. A. Thompson.
(1992) Constrained Monte Carlo maximum likelihood for dependent data, (with discussion). *J. Roy. Statist. Soc. Ser. B* **54:** 657–699.

[3] Geyer, C. J., O. A. Ryder, L. G. Chemnick and E. A. Thompson. (1993) Analysis of Relatedness in the California Condors from DNA fingerprints. *Mol. Biol. Evol.* **10:** 571–589.

[4] Griffiths, R. C. and Tavaré, S. (1994). Ancestral inference in population genetics. *Statistical Science* **9:** 307–319.

[5] Thompson, E. A. (1987). Crossover counts and likelihood in multipoint linkage analysis. *I.M.A J. of Math. Appl. in Med. & Biol.* **4:** 93–108.

[6] Thompson, E. A., J. V. Neel, P. E. Smouse. and R. Barrantes. (1992) Microevolution of the Chibcha-speaking peoples of lower Central America: rare genes in an Amerindian Complex. *Amer. J. Hum. Genet.* **51:** 609–626.

[7] Zhao, L P., Thompson, E. A. and Prentice, R. (1990). Joint estimation of recombination fractions and interference coefficients in multilocus linkage analysis. *Amer. J. Hum. Genet.* **47:** 255–265.

Combining protein secondary structure and evolution

Jeffrey L. Thorne

Program in Statistical Genetics,
Statistics Department, Box 8203,
North Carolina State University,
Raleigh, NC 27695-8203 U.S.A.
thorne@stat.ncsu.edu

Nick Goldman

David T. Jones

Division of Mathematical Biology,
National Institute for Medical Research,
The Ridgeway, Mill Hill,
London NW7 1AA, U.K.

There is great interest in determination of protein structures. The interest stems both from the information that structure can provide about protein function and from applications such as intelligent drug design that are assisted by knowledge of structure. Some structures have been determined but, due to expense and technical challenges, the structures of most proteins are unknown. A step toward understanding the three-dimensional structure of a protein is to first infer the secondary structure of the protein.

Many methods have been proposed for inference of protein secondary structures from groups of aligned protein sequences. These methods are based on the idea that sequences with a common biological function probably have very similar secondary structures. Methods for inferring protein secondary structure seem to have hit a barrier. Several different methods reach a certain rough level of accuracy but none can exceed this level. The accuracy barrier may exist because previously proposed methods treat protein sequences as independent realizations of an underlying process. In reality, sequences with a common biological function tend to be correlated due to a common evolutionary history.

We are developing a secondary structure prediction method that can account for this evolutionary correlation. The method analyzes aligned sequences with unknown secondary structure in a likelihood framework. Our data sets consist of groups of related protein sequences from which a common secondary structure is to be inferred. The model on which we base our likelihood calculations has three components.

Component 1: The amino acid replacement processes

The first component, the replacement processes, specifies rates of change from one amino acid type to another for each of several categories of secondary structure. There have been remarkably few attempts to construct evolutionary models of amino acid replacement. The lack of attempts may be attributable to the complexity of the replacement process. For example, it is known that most replacement events involve chemically similar amino acids but there is no theory that allows replacement patterns to be accurately predicted solely on the basis of chemical similarity. The most widely-used model of replacement (see Dayhoff et al., 1978) is empirically derived. Parameters in the Dayhoff model were estimated by examining replacement patterns in groups of closely related sequences. The empirical basis of the Dayhoff model means that it possesses the realistic feature of favoring replacement events involving chemically similar types of amino acids.

A disadvantage of the Dayhoff and other previously proposed replacement models is they ignore protein secondary structure. They effectively model the evolution of the "average" site in the "average" protein. This is problematic because the physical environment of a protein site may greatly influence the replacement process at the site. In other words, there may be variation among sites in the replacement process. We believe much of this variation can be explained by our model because it allows different secondary structure categories to experience different replacement processes.

Our approach is predicated on the biologically plausible assumption that the replacement process at a sequence site largely depends on the underlying type of secondary structure. For our pilot study, we model three different replacement processes. Each process corresponds to one of three categories of secondary structure: α-helix, β-sheet, and coil ("other"). Later, additional features such as whether a site is on the surface or interior of the protein will be included.

We use many sets of closely related sequences to infer relative rates of replacement for each category of secondary structure. We can do this because each of our sets of sequences includes one sequence with known secondary structure. The fact that the sequences in each set are closely related and therefore easily aligned – coupled with our assumption that evolution never causes a sequence site to switch to a new category of secondary structure – allows the number of replacement events since divergence of each pair of sequences to be accurately counted for each category of secondary structure. The replacement process parameters for each category of secondary structure are estimated from these counts.

Component 2: Organization of secondary structure

The second component of our model describes the organization of secondary structures along a protein sequence. The secondary structure at a site is not independent of secondary structure at surrounding sites; secondary structure is a "regional" property of proteins. We believe the organization of secondary structure along a protein sequence can be well described by hidden Markov models. The first use of hidden Markov models in sequence evolution is by Felsenstein and Churchill. A description of their model has not yet been published (Felsenstein, pers. comm.) but their method is implemented and briefly described in documentation accompanying the versions 3.5 and later of the Phylip software package (Felsenstein 1989). The Felsenstein–Churchill framework was developed to allow regional heterogeneity of nucleotide substitution rates. Our modification of this framework permits regional heterogeneity of protein secondary structure instead. The states of our hidden Markov models are categories of secondary structure and the organization of these states along the sequence is described by a first–order Markov chain. The states are not directly observable but can be inferred through the evolutionary pattern of replacements. Parameters of the hidden Markov models are estimated from the aforementioned sets of closely related sequences.

Component 3: The phylogeny

The third and final component of our model is a description of the evolutionary tree that relates the sequences. This description includes both the topology of the tree and the lengths of its branches. This is the component that both explains the correlation among the sequences and is ignored by other approaches for secondary structure prediction. To calculate the probability of observing the residues in a particular column given the model parameters and the underlying secondary structure category, we adopt the algorithm of Felsenstein (1981). The parameters for the first two components of our model can be estimated from the sets of closely related sequences discussed above. In contrast, the evolutionary tree will vary among data sets and maximum likelihood estimates of parameters that describe the tree need to be obtained for each particular data set.

Example and Future Directions

When a set of four aligned β–amylase sequences is analyzed with our model, the analysis yields a maximum log–likelihood of -3334.77. The corresponding maximum log–likelihood for the Dayhoff model is -3366.06. The large difference in log–likelihoods hints that our model is superior to the Dayhoff model for this data set but a rigorous demonstration that this difference is sufficient to reject the Dayhoff model in favor of our own still needs to be performed. Because the secondary structure underlying one of the four β–amylase sequences is known, we can compare the true structure to our prediction. Using the category of secondary structure with highest probability at a site to predict the secondary structure at a site, our method is correct for 331 out of 496 positions.

We envision many further improvements to our model. These improvements include adding additional secondary structure categories, refining the model of secondary structure organization along the sequence, and allowing heterogeneity of replacement rates among sites in the same category of secondary structure. The promising results of the analysis of the β–amylase data set and anticipated improvements to our model lend credence to our belief that a combined treatment of molecular evolution and protein secondary structure will be more fruitful than consideration of either topic in isolation.

Bibliography

[1] Dayhoff MO, Schwartz RM, Orcutt BC (1978) A model of evolutionary change in proteins. Pp. 345–352 in M.O. Dayhoff, ed. Atlas of protein sequence structure, vol. 5, suppl. 3. National Biomedical Research Foundation, Washington D.C.

[2] Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17:368–376.

[3] Felsenstein J (1989) Phylip – Phylogeny Inference Package (Version 3.2). Cladistics 5:164–166.

Reliability of tree reconstruction methods if sites are not evolving independently

Arndt von Haeseler
Zoologisches Institut
Universität München
Postfach 202136
D-80021 München, Germany
arndt@zi.biologie.uni-muenchen.de

Michael Schöniger
Theoretische Chemie
TU-München
Lichtenbergstr. 4
D-85747 Garching, Germany
schoenig@theochem.tu-muenchen.de

We present a model to simulate the evolution of DNA sequences taking into account dependencies between neighboring sites. More precisely, the sequence is modeled as a Markov chain of order one. The probability c_{xy} to observe nucleotide y at position i depends on the nucleotide x present at position $i-1$. These probabilities are summarized in the so-called correlation matrix $\mathbf{C} = (c_{xy})_{x,y \in \{A,G,C,T\}}$.

Evolution of such sequences of length l is modeled in the framework of time-homogeneous, time-continuous Markov chain methodology. To this end a rate matrix $\mathbf{R} = (r_{ij})_{i,j \in \mathcal{S}^l}$ is defined on the enlarged state space $\mathcal{S}^l = \{\underbrace{AA \cdots A}_{l}, \ldots, TT \ldots T\}$, where

$$r_{ij} = \begin{cases} \pi_j & \text{if} \quad HD(i,j) = 1 \\ 0 & \text{if} \quad HD(i,j) > 1 \\ -\sum_{i \neq k} r_{ik} & \text{if} \quad HD(i,j) = 0 \end{cases}$$

and π_j is the probability to observe sequence j under correlation. Given $\mathbf{R}$ it is – at least theoretically – straightforward to compute $\mathbf{P}(t) = \exp(\mathbf{R}t)$ the probability transition matrix for time t.

In practice, however, this approach is infeasible as soon as l exceeds 5 or 6. For some evolutionary tree reconstruction methods (e.g. the neighbor-joining) it suffices to compute not $\mathbf{P}(t)$ but rather the expected number of observed differences per position (a.k.a. Hamming distance (HD)). We present a method to approximate the Hamming-distance for any size of l and any time t. Figure 1 displays the accumulation of sequence differences, if sequences evolved with independent sites ($l = 1$) or with dependencies that stretch over $l = 50$ nucleotides.

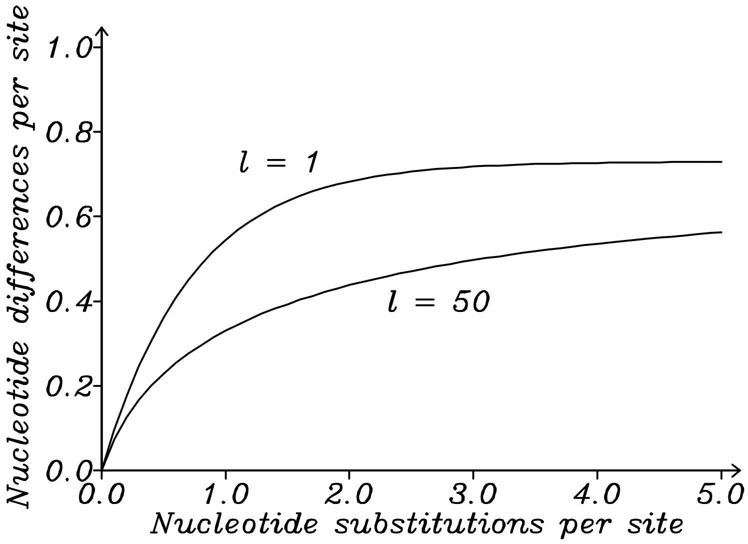

Figure 0.1: Accumulation of sequence differences for the $l = 1$ model (upper curve) and adependency model, assuming $l = 50$ (lower curve).

One readily sees from Figure 1 that the actual number of substitutions is underestimated if one assumes a model of independently evolving sequence sites if in reality sequences show dependencies.

The second part of the talk illustrates the effect of the reliability of various tree reconstruction methods if a wrong model of sequence evolution (in our case: independent sites) is assumed. We show that in the case of the famous unrooted four species tree with two long edges, i.e. a high substitution rate, that are opposite to each other in the tree, and a small substitution rate for the remaining three edges, tree reconstruction methods will in fact fail to reconstruct the true tree, even if one assumes no sampling error in the computation of distances. The following table summarizes our results.

Table 0.1:

reconstruction method	evolution according to: HKY	$l = 2$
MP	ic	ic
NJ+uc	ic	ic
NJ+jc	ic	ic
NJ+km	ic	ic
NJ+fel	ic	ic
NJ+hky	$\emptyset$	$\approx \emptyset$
ML	$\emptyset$	$\approx \emptyset$

The tree reconstruction methods were maximum-parsimony (MP), neighbor-joining (NJ) with either uncorrected (uc), Jukes & Cantor corrected (jc), Kimura corrected (km), Felsenstein (fel) corrected, or Hasegawa, Kishino, and Yano (hky) corrected distances. ML shows the results for the maximum-likelihood method.

ic (inconsistent) indicates that the wrong tree is reconstructed for a substantial portion of choices of parameters. $\emptyset$ implies that for any choice of parameters the method reconstructs the true tree. $\approx \emptyset$ reflects the fact that for certain choices of parameters the method fails to find the model tree. On the other hand, this region (in an appropriately scaled environment) is so small that one can ignore it for all practical purposes.

Finally, we show that the phylogenetic tree for 30 metazoan ribosomal RNA sequences reflects the generally accepted phylogenetic relationship among Metazoa much better, if one is mindful of the dependencies between base-paired nucleotides in the stem regions of these molecules.

Bibliography

[1] Felsenstein, J. (1978) Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. **27**: 401-410.

[2] Huelsenbeck, J.P. & Hillis, D.M. (1993) Success of phylogenetic methods in the four taxon case. Syst. Biol. **42**: 247-264.

[3] Schöniger, M. & von Haeseler, A. (1994) A stochastic model for the evolution of autocorrelated DNA sequences. Mol. Phyl. Evol **3**:240-247.

[4] Schöniger, M. & von Haeseler, A. (1995) Performance of the maximum-likelihood, neighbor-joining, and maximum-parsimony method when sequence sites are not independent. submitted.

[5] von Haeseler, A. & Schöniger, M. (1995) Ribosomal RNA phylogeny derived from a correlation model of sequence evolution. In: From data to knowledge. W. Gaul & D. Pfeiffer (eds.). Springer, Berlin (Germany), pages 000-000.

[6] von Haeseler, A. & Schöniger, M. (1995) Evolution of sequences with dependent sites. Manuscript in progress.

New Algorithms for Constructing Perfect Phylogenies

Tandy J. Warnow
Department of Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104-6018
tandy@central.cis.upenn.edu

Abstract The *Character Compatibility Problem* is a classical problem in computational biology concerned with constructing phylogenetic trees of minimum possible evolution from qualitative character sets. This problem arose in the 1970's, and until recently the only cases for which efficient algorithms were found were for binary (i.e. two-state) characters and for two characters at a time, while the complexity of the general problem remained open. In this paper we will discuss the remarkable progress on this problem since 1990.

Keywords evolution, phylogenetic tree, algorithms.

1 Introduction

The Character Compatibility Problem is a classical problem in computational biology which seeks to construct a phylogenetic tree of minimum possible evolution for an input species set, described by qualitative characters. Qualitative characters can be described mathematically as functions from the species set S to the integers. If there are n species and k characters, we can represent our data by a $n \times k$ integer matrix M, where the entry M_{ij} is the state of species s_i for the j^{th} character. The i^{th} row in this matrix is a vector over Z^k (where Z denotes the integers) which is identified with the species s_i. In this model, in a phylogeny T for the species set S we replace the name of the species at the leaves by the vector for that species, and we also assign to each internal node of the tree T a vector in Z^k. Thus, we are hypothesizing the character states of the ancestral species in our phylogeny. The desired property for this phylogeny is that, for each state of each character, the set of nodes in the phylogeny having that state should form a subtree. That is, the subgraph induced by those nodes should be a connected graph. Such a phylogeny is called a *perfect phylogeny.* We can now state the following:

Perfect Phylogeny Problem: Given a set of species defined by a set of characters, determine whether a perfect phylogeny exists.

This problem is also referred to in the literature as the *Character Compatibility Problem.*

1.1 History of the character compatibility problem

The character compatibility approach to phylogenetic tree construction is related to parsimony and compatibility criteria. These criteria give ways of measuring the fit of a phylogenetic tree to the data. Each approach has its proponents, but essentially they agree in their philosophy, which is that *the tree of minimum evolution is the right tree.*

The concept of characters being perfectly compatible with a given tree arose naturally from these approaches, in that it represents the best *possible* tree (i.e. having minimum possible evolution on the tree). This objective was championed by LeQuesne in a series of papers [17, 18, 19, 20], and later given its firm mathematical foundation in the literature by a series of papers [6, 7].

Since construction of phylogenies using compatibility criteria or any of a number of variations of the parsimony criterion result in *NP-hard* problems, the Character Compatibility Problem was assumed to be *NP-Complete* as well. This was finally proved in [23] and [3]. These results thus imply that efficient algorithms to solve these problems probably can never be found.

For further information about this and other models for computing phylogenetic trees, please see Felsenstein's review paper in [9].

2 Efficient Algorithms

In this section we consider the problem of determining whether a set of characters is perfectly compatible, and if it is, to construct a perfect phylogeny for the data set. There are three natural parameters to the character compatibility problem: n, the number of species (or sequences); k, the number of characters (or length of the sequences); and r, the maximum number of states per character. For this problem, we are interested in algorithms which have running times bounded by polynomials in n, r, and k.

Until 1990, the only progress on algorithms for the character compatibility problem was to show that binary character compatibility [21] and compatibility of two characters at a time [8, 11] could be solved in polynomial time.

In the last few years, there have been several breakthroughs, so that now it is known that if *any* one of these parameters is bounded, then the character compatibility problem can be solved in polynomial time. It is not hard to see that if we bound the number of species by a constant s, then the number of binary phylogenetic trees on s species is also only a constant, so that checking each possible binary tree is a polynomial time algorithm (using the algorithm of Fitch [10]). The other results are more surprising. Algorithms for fixed number of characters first transform the input into a graph and then solve a corresponding graph-theoretic problem. See [22, 14, 2, 4, 13] for these algorithms. We will describe algorithms for the biologically more relevant case where the number of states per character can be bounded.

2.1 Efficient algorithms for fixed numbers of states

Characters derived from molecular data (such as DNA sequences) are associated with the columns of a multiple alignment, and thus have the restriction that the number of states achieved by any particular character is limited to a small constant. In the case of DNA or RNA sequences, this number is four, while for amino-acid sequences, the number is 20. As we will see, by restricting the number of states, we are able to obtain efficient algorithms for constructing perfect phylogenies (if they exist, of course!).

The first case of this type that was handled was for binary characters. Polynomial time algorithms for binary character compatibility were found by several authors with the best being the $O(nk)$ algorithm of Gusfield [12] for testing compatibility of k characters on n species. This is optimal, since the input requires $O(nk)$ space, and each entry must be checked. Algorithms for determining compatibility of three-state characters were found in 1990 by Dress and Steel [5] and Kannan and

Warnow [15]; these algorithms had complexity $O(nk^2)$ and $O(n^2k)$ respectively, where n is the number of species and k the number of characters. Kannan and Warnow then extended the techniques of their algorithm for three-state characters, deriving an $O(n^2k)$ algorithm for four-state character compatibility [15]; they conjectured that it is possible to determine compatibility of k r-state characters on n species in $O(f(r)n^2k)$, where $f(r)$ is a function which depends only upon r and not upon n or k.

A recent result [1] of Agarwala and Fernandez-Baca is an algorithm for character compatibility which has running time $O(2^{3r}(nk^3 + k^4))$, so that the character compatibility problem can be solved in polynomial time when the number of states per character is bounded. More recently, Kannan and Warnow [16] have obtained an running time of $O(2^{2r}nk^2)$ for the perfect phylogeny problem.

Applications

As stated before, the character compatibility problem seeks to find a tree which is *perfect*, and therefore likely not to exist. Despite this, one of the tools that biologists frequently avail themselves of is to find large sets of characters which are pairwise compatible, in their search for *optimal* trees. Since pairwise compatibility does not ensure set-wise compatibility, the algorithms referred to in this paper could assist the biologists in the design of heuristics for tree constructions.

The main benefit of algorithms for the character compatibility problem, besides their use in the development of heuristics for tree construction, is that they provably find the optimal tree when the data supports a perfect phylogeny.

3 Acknowledgements

The author was supported by an NSF National Young Investigator Award CCR 9457800.

Bibliography

[1] R. Agarwala and D. Fernández-Baca, *A Polynomial-time Algorithm for the Perfect Phylogeny Problem when the Number of Character States is Fixed*, SIAM J. on Computing, Vol. 23, No. 6, pp. 1216-1224.

[2] R. Agarwala and D. Fernández-Baca, *Simple and Efficient algorithms for the Triangulating Colored Graphs and Perfect Phylogeny Problems*, submitted.

[3] H. Bodlaender, M. Fellows, and T. Warnow, *Two strikes against perfect phylogeny*, In Proceedings of the 19th International Colloquium on Automata, Languages, and Programming, Springer Verlag, Lecture Notes in Computer Science (1992), pp. 273–283.

[4] H. Bodlaender and T. Kloks, *A simple linear time algorithm for triangulating three-colored graphs*, In Proceedings of the 9th Annual Symposium on Theoretical Aspects of Computer Science (1992), pp. 415–423. To appear, Journal of Algorithms.

[5] A. Dress and M. Steel, *Convex tree realizations of partitions*, Appl. Math. Letters, Vol. 5, No. 3 (1992), pp. 3–6.

[6] G.F. Estabrook, *Cladistic methodology: a discussion of the theoretical basis for the induction of evolutionary history*, Annu. Rev. Ecol. Syst., 3 (1972), pp. 427-456.

[7] G. F. Estabrook, C. S. Johnson Jr., and F. R. McMorris, *An idealized concept of the true cladistic character*, Mathematical Biosciences, 23 (1975), pp. 263–272.

[8] Estabrook, G.F. and Landrum, L. 1975: A simple test for the possible simultaneous evolutionary divergence of two aminoacid positions, *Taxon.* 24: 609-613.

[9] Felsenstein, J. 1982: Numerical methods for inferring evolutionary trees, *The quarterly review of biology*, Vol. 57, No. 4.

[10] Fitch, Wm. 1971: Toward defining the course of evolution: minimum change for a specified tree topology, *Syst. Zool.*, 20:406-416.

[11] Fitch, W.M. 1975: Toward finding the tree of maximum parsimony, Proc. Eighth International Conference on Numerical Taxonomy, G.F. Estabrook, ed., pp. 189-230, W.H. Freeman, San Francisco.

[12] D. Gusfield, *Efficient algorithms for inferring evolutionary trees*, Networks, 21 (1991), pp. 19–28.

[13] Idury, R. and Schaffer, A. 1993: Triangulating three-colored graphs in linear time and linear space, to appear, *SIAM J. Discrete Mathematics.*

[14] S. Kannan and T. Warnow, *Triangulating three-colored graphs*, SIAM J. on Disc. Math., 5 (1992), pp. 249–258.

[15] S. Kannan and T. Warnow, *Inferring evolutionary history from DNA sequences*, SIAM J. on Computing, Volume 23, No. 3, (1994) pp. 713-737.

[16] S. Kannan and T. Warnow, *A fast algorithm for the computation and enumeration of perfect phylogenies*, Proceedings, ACM/SIAM Symposium on Discrete Algorithms (SODA), San Francisco, California, January 1995.

[17] LeQuesne, W.J. 1969: A method of selection of characters in numerical taxonomy, *Syst. Zool.*, 18: 201-205.

[18] LeQuesne, W.J. 1972: Further studies on the uniquely derived character concept, *Syst. Zool.*, 21: 281-288.

[19] LeQuesne, W.J. 1974: The uniquely evolved character concept and its cladistic application, *Syst. Zool.*, 23: 513-517.

[20] LeQuesne, W.J. 1977: The uniquely evolved character concept, *Syst. Zool.*, 26, pp. 218-223.

[21] McMorris, F.R. 1977: On the compatibility of binary qualitative taxonomic characters, *Bull. Math. Biol.* 39: 133-138.

[22] F. R. McMorris, T. Warnow, and T. Wimer, *Triangulating vertex colored graphs*, SIAM J. on Discrete Mathematics, Vol. 7. No. 2, (1994), pp. 296-306.

[23] M. A. Steel, *The complexity of reconstructing trees from qualitative characters and subtrees*, Journal of Classification, 9 (1992), pp. 91–116.