

Phylogenetic Approaches to Molecular Epidemiology

Keith A. Crandall¹ and David Posada^{2,3}

¹Department of Integrative Biology & Department of Microbiology and
Molecular Biology, Brigham Young University Provo, UT 84602-5255, USA

²Variagenics, Inc. 60 Hampshire Street Cambridge, MA 02139-1548

³Center for Cancer Research Massachusetts Institute of Technology
Cambridge, MA 02139, USA

1. INTRODUCTION

Phylogenies, diagrams of branching patterns representing the estimated evolutionary histories among organisms or their parts (Crandall, 2001), have become essential tools in the study of the molecular epidemiology of disease agents. While the idea of using phylogenetic approaches to study epidemiology is not new (Harvey *et al.*, 1996; Harvey and Nee, 1994), this book is a testament to the extraordinary information that can be obtained through a phylogenetic analysis of the etiological agents of disease. A prime example of the troubles encountered when the phylogenetic approach is ignored comes from the outbreak of the West Nile Virus in New York City. This virus was responsible for multiple deaths in New York, yet the Centers for Disease Control and Prevention (CDC) initially misdiagnosed the causative agent as St. Louis encephalitis due to their lack of an appropriate phylogenetic comparison (Enserink, 1999). The study of origins, spread, and diversity of pathogens are clearly evolutionary questions. Only after the serological evidence was coupled with strong phylogenetic evidence was the etiological agent responsible for the encephalitis outbreak in New York correctly identified as the West Nile Virus (Lanciotti *et al.*, 1999). Likewise, other chapters in this book provide extensive examples of the insights obtained through phylogenetic thinking. Given the power of the phylogenetic approach, in this chapter we review the basic approaches and considerations in estimating phylogenies and phylogenetically based estimators of natural

selection. We refer the reader the Hillis (1999) for another view on the basics of phylogeny reconstruction, especially relative to HIV sequences, as well as to Posada *et al.* (2001) for a detailed account (including lots of equations!) on phylogeny reconstruction, ancestral state reconstruction, and hypothesis testing in a phylogenetic framework. For those who wish to delve even deeper, Swofford *et al.* (1996) provide the best and most extensive summary of phylogenetic methodology currently available.

2. SEQUENCE ALIGNMENT

A sequence alignment is a central part of any phylogenetic analysis. Indeed, ideally one would like to estimate a phylogeny and adjust the alignment simultaneously applying the same optimality criterion to both endeavors. However, the algorithms today allow this only on a limited basis with a limited range of optimality criteria (but see Giribet, 2001). Therefore, the standard approach to sequence alignment is to use generally available software, such as Clustal X (Thompson *et al.*, 1997) and hope for the best. Because the amino acid alphabet is made of 20 characters (Ala, Phe, Val, Iso, Leu, etc.) while the DNA alphabet is made of only 4 (A, C, G, T), the alignment of amino acids is easier and a more reliable procedure than the alignment of nucleotide sequences. Therefore, when working with coding sequences it is to our advantage to align the corresponding amino acids and then get back to the original nucleotides. Unfortunately, programs like Clustal do not automate this procedure, so the usual method is to align the nucleotides, translate them to amino acids, and check that the quality of the implied amino acid alignment. In this section, we will discuss approaches for assessing alignments before subsequent phylogenetic analyses.

2.1 What sequences to align?

The first question a researcher is faced with is what sequences should be included in the alignment. First of all there are the data collected for the analysis. These data, generally, are only as good as the shortest sequence. While some phylogeny reconstruction algorithms can deal with missing data, most ignore it and many give spurious results when missing data are included. Therefore, it is usually ideal to trim a data set down to exclude as much missing data as possible.

In addition to the sequences from the laboratory, in molecular epidemiology it is often desirable to include known lab strains in an analysis to guard against potential contamination (Korber *et al.*, 1995). Likewise, sequences resulting from a BLAST search (Altschul *et al.*, 1990; Altschul *et al.*, 1997; Zhang and Madden, 1997), a search for genetically similar

sequences in GenBank, can be included to verify that the sequences obtained in the lab are from the appropriate organism and gene region as well as to further guard against contamination. Often these sequences are more distant from the sequences under study and may force a more complicated alignment; therefore, we suggest that such sequences be used in a preliminary analysis to test for contamination etc. and then be deleted from subsequent analyses.

2.2 Adjusting alignments: What to look for?

Once a set of sequences is decided on and a preliminary alignment is obtained through ClustalX, one should always check this alignment for anomalies. For instance, the first check one should perform is a translation to proteins. There are at least two computer programs, Se-Al (Rambaut, 2002) and MacClade (Maddison and Maddison, 2000), that allow the user to toggle back and forth between nucleotide and amino acid alignment. One can then adjust the alignment to ensure proper translation to proteins (Fig. 1).

Furthermore, because there are 21 possible amino acid character states and only 4 possible nucleotide states and two-thirds fewer characters, it is much easier to align the amino acids rather than the nucleotides. However, the nucleotides often have greater information content for phylogeny amino acid alignments for refinement of alignments based on amino acid positions, but keeping the nucleotide information in tact for further analysis.

Once the final alignment is settled on, often there are still regions of ambiguity. These are not necessarily all regions with gaps. Some gaps are clear insertion or deletion events—real evolutionary events and should therefore be included in the evolutionary analysis if they can be unambiguously inferred. However, many times gaps are not unambiguously placed and therefore the positional homology (Swofford *et al.*, 1996) of the nucleotide characters is in question. Whenever the positional homology is in question, these characters should be excluded from subsequent phylogenetic analysis. Thus the common practice of “gap stripping” in virological studies is inappropriate (Posada *et al.*, 2001). Instead researchers should more carefully scrutinize their alignments to exclude individual columns of data with questionable homology (Hillis, 1994).

3. PHYLOGENY RECONSTRUCTION

Once a satisfactory alignment is obtained for the nucleotide sequence data, a phylogenetic (evolutionary) history can be estimated from these data. There are a variety of approaches to phylogeny estimation and there has been substantial debate on which approach represents the “best” method. Fundamental to the debate is the fact that there are different ways to optimize

character evolution on a tree, *i.e.*, different optimality criteria. Thus the first step in phylogeny estimation is to choose an optimality criterion.

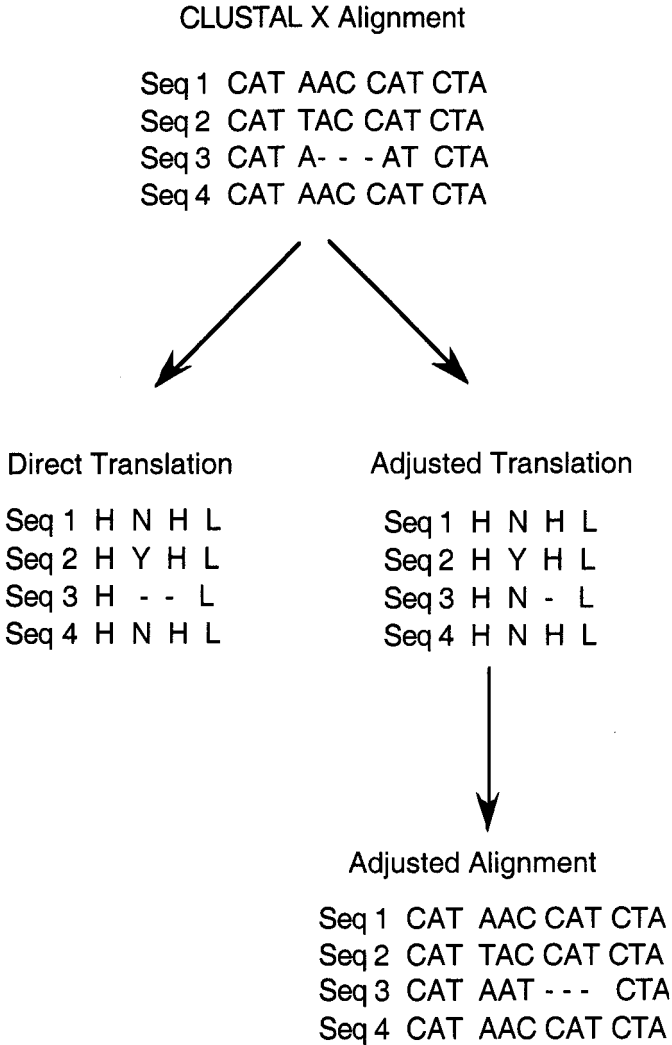


Fig. 1. A nucleotide alignment from Clustal X that disrupts the coding frame. By translating the nucleotides to amino acids a better alignment can be achieved and then used to refine the nucleotide alignment for subsequent analyses.

3.1 Optimality criteria

The optimality criteria are how one measures the goodness-of-fit of the data to a given hypothesis, where in the phylogenetic context, the

hypotheses are alternative tree topologies with associated branch lengths. The dominant criteria used in phylogenetics are maximum parsimony (Edwards, 1996; Edwards and Cavalli-Sforza, 1964), maximum likelihood (Cavalli-Sforza and Edwards, 1967; Felsenstein, 1981), and minimum evolution (Rzhetsky and Nei, 1992). The main object then, is to maximize or minimize a given statistic by assessing that statistic on all possible tree topologies. Thus for the principle of maximum parsimony, one tries to minimize the amount of character change along the phylogeny and therefore the phylogeny of choice is the one (or more) tree with the minimum overall tree length as a measure of character change. In contrast, maximum likelihood attempts to maximize the likelihood of the data given the tree and a model of evolution (Huelsenbeck and Crandall, 1997). Coupled with the choice of optimality criterion is the choice of search strategy, given that criterion.

3.2 Search strategies and speed

Ideally, one would like to optimize the tree statistic on all possible trees, thereby guaranteeing the best (or set of best) solution to the problem. This exhaustive search, however, is usually impractical for large numbers of sequences because the number of possible trees grows exponentially relative to the number of unique sequences added (Tab. 1). Therefore, exhaustive searches are prohibitive for more than 15 or so unique sequences. Branch and bound (Hendy and Penny, 1982) searches will guarantee the set of most optimal trees, but take short cuts to identify them so statistics for all trees do not need to be enumerated. This allows for a few more sequences to be added to the analysis compared to the exhaustive search. Since we are typically dealing with large numbers of sequences in molecular epidemiology, alternative strategies are desired to search the tree space.

Tab. 1. Number of possible unrooted bifurcating trees as a function of the number of terminal sequences.

Number of Unique Sequences	Number of Trees
10	2×10^6
50	3×10^{74}
100	2×10^{182}
1,000	$2 \times 10^{2,860}$
10,000	$8 \times 10^{38,658}$
100,000	$1 \times 10^{486,663}$
1,000,000	$1 \times 10^{5,866,723}$

The most generally used alternative strategy is the heuristic search. Because a heuristic search depends on the starting tree topology (Templeton, 1992) and there exist multiple islands of optimal trees (Maddison, 1991; Salter, 2001), it is essential to begin heuristic searches with randomly selected

tree topologies (in PAUP* through the RANDOM SEQUENCE ADDITION option). Repeating searches with different randomly selected starting trees allows one to explore the tree space and therefore have a greater chance of escaping local optima and finding the globally optimal solution (tree). Maximum likelihood searches are notoriously slow, since the calculation of the likelihood statistic is complex. However, alternative strategies for likelihood implementations have recently been developed that show great promise. The first is a genetic algorithm for exploring the tree space that uses “recombination” and “natural selection” in an algorithmic sense to selectively explore the tree space (Lewis, 1998). Here individuals in the population are defined by a tree, branch lengths, and parameter values in the model of evolution. Populations are then evolved to find the most fit individual. This method has recently been extended to allow processing of the genetic algorithm in parallel and shows great potential for increased search speeds for large data sets (Brauer *et al.*, 2002). Bayesian approaches also provide significant increases in efficiency in tree space exploration and therefore provide a faster approach to finding more optimal trees (Huelsenbeck *et al.*, 2001). The Bayesian approach has also been used to test molecular clocks, detect selection, select models of evolution, and to evaluate uncertainty in phylogenies.

3.3 Models of evolution

When calculating an optimality score for a tree given some criterion, a model of evolution is required to accomplish this calculation. A model of evolution is used to define the probability of substitution from one nucleotide to another (Fig. 2). In addition to the transition probabilities from one nucleotide to another, models can also take into account biases in nucleotide frequencies (Felsenstein, 1981), invariable sites, substitutional rate heterogeneity (Yang, 1996), and codon position (Muse and Gaut, 1994; Yang, 1994). Models of evolution have even been developed to take into account different reading frames (Pedersen and Jensen, 2001). Muse (1999) provides an excellent overview of models of evolution as well as an exploration of fitting models to HIV-1 sequences.

The model of evolution used in a phylogenetic analysis can have a significant effect on the resulting tree and therefore on conclusions made in a phylogenetic investigation (*e.g.*, Kelsey *et al.*, 1999). Nucleotide sequences used in molecular epidemiological studies often show biases associated with base frequencies, transition/transversion biases, and rate heterogeneity (Jenkins *et al.*, 2002; Posada and Crandall, 2001d). Therefore, it is critical to optimize a model to a given data set. A maximum likelihood framework provides a convenient approach to optimizing models to data through a series of hierarchical likelihood ratio tests that test assumptions about how

nucleotides evolve for a given data set (Huelsenbeck and Crandall, 1997). This approach has been formalized in a software implementation called ModelTest (Posada and Crandall, 1998). Posada *et al.* (2001) recently reviewed this approach in detail and therefore this discussion will not be repeated here. However, recently simulation studies have shown that this hierarchical likelihood ratio testing performs very well at recovering the true underlying model of evolution for simulated data sets (Posada, 2001; Posada and Crandall, 2001a).

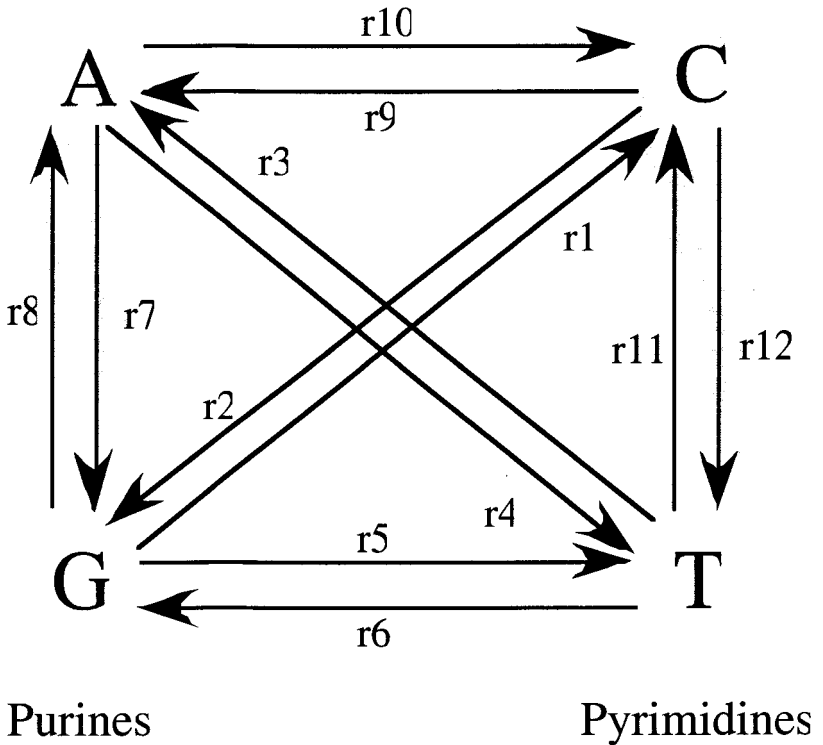


Fig. 2. Models of evolution specify different rates of evolution from one nucleotide to another. In this generalized model, there are 12 different rates (r_1 - r_{12}) associated with the different possible changes from one nucleotide to another. Models can be further complicated by incorporating nucleotide frequencies, codon position, rate heterogeneity, etc.

3.4 Confidence assessment

Once a model of evolution is selected and phylogenetic relationships are estimated, one then proceeds to assess the confidence of the estimated relationships. Typically, this is done using the bootstrap procedure

(Felsenstein, 1985). The bootstrap procedure creates a new data set by choosing columns of data from the original data set at random and with replacement until a new data set is created that has the same sequence length as the original. Note that because the bootstrap samples *with replacement*, some sites (or columns of data) will be represented multiple times whereas others will not be represented at all. Then a new tree is estimated from this resampled data set. This procedure is repeated multiple times (typically 100 to 1000) to achieve reasonable precision. Hillis and Bull (1993) evaluated the bootstrap approach to assessing confidence in phylogenetic analyses using computer simulations and a laboratory-generated known phylogeny. They showed that bootstrap proportions provide biased but highly conservative estimates of the probability of correctly inferring the corresponding clades, suggesting that bootstrap proportions of $\geq 70\%$ usually correspond to a probability of $\geq 95\%$ that the corresponding clade is real (Hillis and Bull, 1993). However, the bias associated with the bootstrap can become pronounced with large-scale phylogenies and thereby reduce the accuracy of the confidence assessment (Sanderson and Wojciechowski, 2000). Since most molecular epidemiological data sets are typically quite large, it is ideal to take into account this bias in confidence assessment. This can be accomplished through the use of an iterative bootstrap method (Zharkikh and Li, 1995), which eliminates the bias (reduction in accuracy) associated with increased sampling. Alternative approaches to confidence assessment can be carried out within a Bayesian framework, which allows for the estimation of the posterior probabilities for each node of the tree (Huelsenbeck and Ronquist, 2001; Huelsenbeck *et al.*, 2001).

3.5 Sampling considerations

An appropriate sampling strategy becomes a key consideration for both the accuracy of phylogeny reconstruction (Hillis, 1998), as well as parameter estimates associated with models of evolution (Sullivan *et al.*, 1999). Sampling considerations typically entail two components. First is the number of "taxa" or sequences needed for a given study. Second is the number of "characters" or nucleotides required per sequence. There has been substantial debate in the systematic literature on the relative importance of increasing the taxon sampling versus the character sampling (Greybeal, 1998; Kim, 1998; Poe, 1998; Poe and Swofford, 1999). This debate shows little sign of slowing down (Pollock *et al.*, 2002; Rosenberg and Kumar, 2001). In addition, epidemiological studies typically require geographic sampling considerations. Inferences of population structure and history will depend critically on an appropriate geographic sampling strategy that incorporates random sampling throughout the geographic distribution of the population of inference (Templeton *et al.*, 1995). Therefore, in designing molecular

epidemiological studies, careful consideration is warranted for the justification of sampling strategy in terms of numbers of sequences, length of sequences, and geographic distribution of samples relative to the hypotheses being tested.

4. HYPOTHESIS TESTING IN A PHYLOGENETIC FRAMEWORK

After a phylogenetic hypothesis of evolutionary relationships has been estimated, the tendency is to use this as the final evidence for or against a given hypothesis. However, there are more formal statistical frameworks for testing alternative phylogenetic hypotheses that should be incorporated in phylogenetic studies of molecular epidemiology (Crandall *et al.*, 1999b; Posada *et al.*, 2001).

The first of these frameworks was developed to test two *a priori* hypotheses. Templeton (1983) developed the first of these tests. His was a nonparametric test that simply asked if one hypothesis had a statistically significantly shorter tree length (in a parsimony framework) relative to the alternative hypothesis. A similar parametric test was developed by Kishino and Hasegawa (1989) within a likelihood framework. Both of these approaches were recently reviewed with associated examples from HIV-1 concerning the legitimacy of the group N strain as a new group of HIV-1 (Posada *et al.*, 2001). When these approaches are inappropriately applied to situations where the alternative hypotheses are not *a priori* (e.g., when one compares the best estimated phylogeny to an alternative) or when comparing multiple topologies, then these tests can be biased and lead to overconfidence in the wrong tree (Goldman *et al.*, 2000; Shimodaira and Hasegawa, 1999). Goldman *et al.* (2000) recommend a number of parametric and nonparametric alternative tests that do not suffer from these concerns. These tests are implemented in the phylogenetic software package PAUP* (Swofford, 2000). Again, Bayesian approaches are very promising in comparing different hypotheses while taking into account phylogenetic uncertainty (Huelsenbeck *et al.*, 2000).

5. RECOMBINATION

Recombination can play a dominant role in the evolution of infectious diseases (e.g., Gibbs *et al.*, 2001; Posada, 2002). The relative contribution of recombination versus mutation to the genetic diversity of a population can be a key component to designing effective drug and/or vaccine strategies. The quantification of this relative contribution to genetic diversity has rarely been attempted for infectious diseases. But it is clear that recombination can play a

significant role in the generation of diversity (e.g., Falush *et al.*, 2001; Feil *et al.*, 2001; Feil *et al.*, 1999; Guttman and Dykhuizen, 1994; Posada *et al.*, 2000; Rich *et al.*, 2001; Robertson *et al.*, 1995). Recombination can also affect our ability to accurately reconstruct evolutionary relationships (Posada and Crandall, 2002) and adversely affect our ability to accurately estimate parameters associated with molecular evolution (Schierup and Hein, 2000). The amount of recombination relative to mutation will also determine the clonality of an infectious agent (e.g., Bart *et al.*, 2001). Therefore it is desirable to test for recombination in a set of aligned sequences before a phylogenetic analysis is performed. Unfortunately, there are a great number of methods to choose from for detecting recombination (reviewed in Crandall and Templeton, 1999) with new methods being developed continuously (e.g., Dorman *et al.*, 2002). The central question then becomes, which method should be used to detect recombination. The answer, unfortunately, is not trivial.

Three different research groups have recently explored the ability of various methods to detect recombination. The first group studied the statistical power (the probability that a statistical test will reject the null hypothesis) of four distinct methods using simulated sequences under a coalescent model with recombination. The simulation results showed clear differences in statistical power among these four methods with the incompatibility approaches having the highest power and the phylogenetic approaches have lower power (Brown *et al.*, 2001). The next group also investigated the statistical power of four methods to detect recombination, but added variation in the mutation rate as well as the recombination rate. This is of interest because some methods may perform differentially well at different divergences. Again, incompatibility approaches performed better than phylogenetic methods and all methods detected fewer recombination events than theoretically possible (Wiuf *et al.*, 2001). These papers set the foundation for the third paper which capitalized on the theoretical contributions of this earlier work to perform more extensive simulation studies that examined the ability of fourteen different methods to detect recombination while varying recombination rate, mutation rate, and rate variation across sites. Again, there was no clearly superior method with different methods performing best at different levels of diversity (mutation rates), with incompatibility methods outperforming phylogenetic methods (Posada and Crandall, 2001b). All studies showed that the use of multiple techniques is a reasonable approach and that these techniques can be chosen relative to the amount of genetic diversity in the data set. Methods to detect recombination, methods to estimate recombination rates, and the impact of recombination on phylogenies in bacterial and viral settings were recently reviewed in detail (Posada *et al.*, 2002).

6. NETWORK APPROACHES FOR ESTIMATING GENE GENEALOGIES

As we have seen, often when performing genealogical analyses of sequence data in molecular epidemiological studies, recombination is a potential complicating factor. There are clearly methods for detecting recombination and some indication of their relative performance. Given the presence of recombination, can we still estimate genealogical relationships among sequences? Clearly the standard bifurcating tree approach to phylogeny reconstruction will not suffice. Not only does recombination affect our ability to reconstruct such trees (Posada and Crandall, 2002), but a bifurcating tree is an incorrect representation of reticulate evolutionary histories on first principles. Thus we must look to alternative more realistic representations of genealogical relationships in the presence of recombination. An effective alternative representation for such relationships is as genealogical networks.

Again, there are a host of methods that have been developed to represent genealogical relationships as networks (*e.g.*, Bandelt and Dress, 1992; Excoffier and Smouse, 1994; Strimmer and Moulton, 2000; Templeton *et al.*, 1992). They all have the advantage of being able to take into account population genetic phenomena such as recombination, nonbifurcating trees, and ancestral sequences still in the population. These phenomena are typically ignored by traditional methods of reconstructing phylogenetic relationships. Unlike the recombination methods, there have been no studies to examine the relative abilities of these methods to accurately reconstruct gene genealogies. However, these methods and the general ideas behind network approaches to estimating genealogical relationships have recently been reviewed (Posada and Crandall, 2001c). This review also provides a list of software (and associated websites) available to implement these methods.

7. DETECTING SELECTION

The standard approach to estimating the effects of natural selection in molecular sequence data is to estimate the ratio of nonsynonymous (dn - substitutions changing the amino acid) to synonymous substitution (ds) rate ratio (dn/ds). According to population genetic theory, if this ratio is greater than one, this is evidence of positive selection. If the ratio is less than one, it is evidence of purifying selection and if it equals one, this is evidence of neutral evolution (Sharp, 1997; Yang and Bielawski, 2000). The standard estimator of this ratio is the Nei-Gojobori method (Nei and Gojobori, 1986). However, this method has been shown to be a biased estimator of this ratio due to the pairwise comparisons (Crandall *et al.*, 1999a) and lack of an appropriate model of evolution (Yang and Nielsen, 1998). Alternative approaches allow

for the explicit incorporation of codon-based models of evolution (Goldman and Yang, 1994; Muse and Gaut, 1994). These models were used to develop more robust approaches to estimating selection through the dn/ds ratio. These approaches came in two varieties. First were the lineage-specific models that assumed constant selection pressure across sites but allowed the dn/ds rate variation across lineages (Yang, 1998; Yang and Nielsen, 1998). The alternative was the site-specific model that assumed constant selection pressure across lineages but allowed variation over sites (Nielsen and Yang, 1998; Yang *et al.*, 2000). While these methods were successful in identifying selection in some cases (*e.g.*, Zanotto *et al.*, 1999), they still suffer from the averaging effects across either sites or lineages. Ideally, one would prefer to test for selection at individual sites without averaging effects across sites (or lineages).

Two approaches have recently been developed to accommodate this desire. The first approach extends the models developed above in the maximum-likelihood framework to allow for both variation across sites and across lineages (Yang and Nielsen, 2002). This approach thereby allows for the reality of selection at individual sites with regions of functional constraint. This approach is implemented in the software package PAML (Yang, 2001).

An alternative approach is to reconstruct the evolutionary changes on a phylogeny and explore the magnitude of these changes relative to changes in biochemical properties (McClellan and McCracken, 2001). This approach identifies all the amino acid replacements in an evolutionary context (or compared to a reference sequence). It then classifies these changes relative to a suite of 31 biochemical properties and ranks each change on a scale from 1 to 8 in terms of the magnitude of the change relative to these properties. The magnitude of change is then used to infer the mode of natural selection (positive versus purifying selection, etc.) in a hypothesis testing framework. Using this approach, one can identify those particular amino acid replacements that have significant effects on the biochemical properties of protein evolution and are therefore likely candidates for testing their impact relative to protein structure. This approach is implemented in the computer software package TreeSAAP (Woolley *et al.*, 2002).

8. SUMMARY

Phylogenetic methods are essential tools for the study of molecular epidemiology. Many of the hypotheses associated with molecular epidemiology are historical in nature and therefore answered most straightforwardly by phylogenetic analyses. Some have even used phylogeny to predict future outcomes of infectious outbreaks (*e.g.*, Bush *et al.*, 1999). While some workers in infectious disease continue to ignore phylogenetics, as we

have seen in the case of the West Nile Virus this can be problematic and limiting in the interpretation and analysis of data (Fitch *et al.*, 2001). Instead, most exciting work in molecular epidemiology is now embracing population biology and evolutionary theory for a productive synthesis of ideas and approaches concerning infectious diseases and the design and evaluation of interventions for their treatment and prevention (Levin *et al.*, 1999).

9. ACKNOWLEDGEMENTS

We would like to thank the editor for inviting our chapter and for his patience. This work was supported by NIH grant R01-HD34350 and NSF DEB-0073154.

10. REFERENCES

- Altschul S.F., Gish W., Miller W., Myers E., and Lipman D.J. 1990. Basic local alignment search tool. *J Mol Biol* 215:403-410.
- Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z. *et al.* 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25:3389-3402.
- Bandelt H.-J. and Dress A.W.M. 1992. Split decomposition: A new and useful approach to phylogenetic analysis of distance data. *Mol Phylogen Evol* 1:242-252.
- Bart A., Barnabe C., Achtman M., Dankert J., van der Ende A. *et al.* 2001. The population structure of *Neisseria meningitidis* serogroup A fits the predictions for clonality. *Infect Gen Evol* 1:117-122.
- Brauer M.J., Holder M.T., Dries L.A., Zwickl D.J., Lewis P.O. *et al.* 2002. Genetic algorithms and parallel processing in maximum-likelihood phylogeny inference. *Mol Biol Evol*: in press.
- Brown C.J., Garner E.C., Dunker A.K., and Joyce P. 2001. The power to detect recombination using the coalescent. *Mol Biol Evol* 18:1421-1424.
- Bush R.M., Bender C.A., Subbarao K., Cox N.J., and Fitch W.M. 1999. Predicting the evolution of human influenza A. *Science* 286:1921-1925.
- Cavalli-Sforza L.L. and Edwards A.W.F. 1967. Phylogenetic analysis: models and estimation procedures. *Evolution* 32:550-570.
- Crandall K.A. 2001. Phylogeny. In *Encyclopedia of Genetics*, p. 1465-1466, Brenner S. and Miller J.H., eds. Academic Press, London.
- Crandall K.A., Kelsey C.R., Imamichi H., and Salzman N.P. 1999a. Parallel evolution of drug resistance in HIV: failure of nonsynonymous/synonymous substitution rate ratio to detect selection. *Mol Biol Evol* 16:372-382.
- Crandall K.A. and Templeton A.R. 1999. Statistical methods for detecting recombination. In *The Evolution of HIV*, p. 153-176, Crandall K.A., ed. The Johns Hopkins University Press, Baltimore, MD.
- Crandall K.A., Vasco D., Posada D., and Imamichi H. 1999b. Advances in understanding the evolution of HIV. *AIDS* 13:S39-S47.
- Dorman K.S., Kaplan A.H., and Sinsheimer J.S. 2002. Bootstrap confidence levels for HIV-1 recombination. *J Mol Evol* 54:200-209.
- Edwards A.W.F. 1996. The origin and early development of the method of minimum evolution for the reconstruction of phylogenetic trees. *Syst Biol* 45:79-91.
- Edwards A.W.F. and Cavalli-Sforza L.L. 1964. Reconstruction of evolutionary trees. In *Phenetic and phylogenetic classification*, p. 67-76, McNeill J. ed. Systematics Association Publication, London.
- Enserink M. 1999. Groups race to sequence and identify New York virus. *Science* 286:206-207.
- Excoffier L. and Smouse P.E. 1994. Using allele frequencies and geographic subdivision to reconstruct gene trees within a species: Molecular variance parsimony. *Genetics* 136:343-359.

- Falush D., Kraft C., Taylor N.S., Correa P., and Fox J.G. *et al.* 2001. Recombination and mutation during long-term gastric colonization by *Helicobacter pylori*: Estimates of clock rates, recombination size, and minimal age. *Proc Natl Acad Sci USA* 98:15056-15061.
- Feil E.J., Holmes E.C., Bessen D.E., Chan M.-S., Day N.P.J. *et al.* 2001. Recombination within natural populations of pathogenic bacteria: Short-term empirical estimates and long-term phylogenetic consequences. *Proc Natl Acad Sci USA* 98:182-187.
- Feil E.J., Maiden M.C.J., Achtman M., and Spratt B.G. 1999. The relative contributions of recombination and mutation to the divergence of clones of *Neisseria meningitidis*. *Mol Biol Evol* 16:1496-1502.
- Felsenstein J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. *J Mol Evol* 17:368-376.
- Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39:783-791.
- Fitch W., Brisse S., Stevens J., and Tibayrenc M. 2001. Infectious diseases and the golden age of phylogenetics: An E-debate. *Infect Gen Evol* 1:69-74.
- Gibbs M.J., Armstrong J.S., and Gibbs A.J., 2001. Recombination in the hemagglutinin gene of the 1918 "Spanish Flu". *Science* 293:1842-1845.
- Giribet G. 2001. Exploring the behavior of POY, a program for direct optimization of molecular data. *Cladistics* 17:S60-S70.
- Goldman N., Anderson J.P. and Rodrigo A.G. 2000. Likelihood-based tests of topologies in phylogenetics. *Syst Biol* 49:652-670.
- Goldman N. and Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol* 11:725-736.
- Greybeal A. 1998. Is it better to add taxa or characters to a difficult phylogenetic problem? *Syst Biol* 47:9-17.
- Guttman D.S. and Dykhuizen D.E. 1994. Clonal divergence in *Escherichia coli* as a result of recombination, not mutation. *Science* 266:1380-1383.
- Harvey P.H., Leigh Brown A.J., Maynard Smith J., and Nee S., eds. 1996. *New Uses for New Phylogenies*. Oxford University Press, Oxford, England.
- Harvey P.H. and Nee S. 1994. Phylogenetic epidemiology lives. *Trends Ecol Evol* 9:361-363.
- Hendy M.D. and Penny D. 1982. Branch and bound algorithms to determine minimal evolutionary trees. *Math Biosci* 59:277-290.
- Hillis D.M. 1994. Homology in molecular biology. In *Homology: The Hierarchical Basis of Comparative Biology*, p. 339-368, Hall B.K., ed. Academic Press, Inc., New York.
- Hillis D.M. 1998. Taxonomic sampling, phylogenetic accuracy, and investigator bias. *Syst Biol* 47:3-8.
- Hillis D.M. 1999. Phylogenetics and the study of HIV. In *The Evolution of HIV*, Crandall K.A., ed. Johns Hopkins University Press, Baltimore, MD.
- Hillis D.M. and Bull J.J. 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Syst Biol* 42:182-192.
- Huelsenbeck J.P. and Crandall K.A. 1997. Phylogeny estimation and hypothesis testing using maximum likelihood. *Annu Rev Ecol Syst* 28:437-466.
- Huelsenbeck J.P., Rannala B., and Masly J.P. 2000. Accommodating phylogenetic uncertainty in evolutionary studies. *Science* 288:2349-2350.
- Huelsenbeck J.P. and Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17:754-755.
- Huelsenbeck J.P., Ronquist F., Nielsen R., and Bollback J.P. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* 294:2310-2314.
- Jenkins G.M., Rambaut A., Pybus O.G., and Holmes E.C. 2002. Rates of molecular evolution in RNA viruses: A quantitative phylogenetic analysis. *J Mol Evol* 54:156-165.
- Kelsey C.R., Crandall K.A. and Voevodin A.F. 1999. Different models, different trees: The geographic origin of PTLV-I. *Mol Phylogeny Evol* 13:336-347.
- Kim J. 1998. Large-scale phylogenies and measuring the performance of phylogenetic estimators. *Syst Biol* 47:43-60.
- Kishino H. and Hasegawa M. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. *J Mol Evol* 29:170-179.
- Korber B.T.M., Learn G., Mullins J.I., Hahn B.H., and Wolinsky S. 1995. Protecting HIV databases. *Nature* 378:242-243.

- Lanciotti R.S., Roehrig J.T., Deubel V., Smith J., Parker M. *et al.* 1999. Origin of the West Nile Virus responsible for an outbreak of encephalitis in the Northeastern United States. *Science* 286:2333-2337.
- Levin B.R., Lipsitch M., and Bonhoeffer S. 1999. Population biology, evolution, and infectious disease: convergence and synthesis. *Science* 283:806-809.
- Lewis P.O. 1998. A genetic algorithm for maximum-likelihood phylogeny inference using nucleotide sequence data. *Mol Biol Evol* 15:277-283.
- Maddison D.R. 1991. The discovery and importance of multiple islands of most-parsimonious trees. *Syst Zool* 40:315-328.
- Maddison D.R. and Maddison W.P. 2000 *MacClade 4: Analysis of Phylogeny and Character Evolution*. Sinauer Associates, Sunderland, MA.
- McClellan D.A. and McCracken K.G. 2001. Estimating the influence of selection on the variable amino acid sites of the cytochrome B protein functional domain. *Mol Biol Evol* 18:917-925.
- Muse S. 1999. Modeling the molecular evolution of HIV sequences. In *The Evolution of HIV*, in press, Crandall K.A., ed. Johns Hopkins University Press, Baltimore, MD.
- 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* 11:715-724.
- Nei M. and Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* 3:418-426.
- Nielsen R. and Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* 148:929-936.
- Pedersen A.-M. K. and Jensen J.L. 2001. A dependent-rates model and an MCMC-based methodology for the maximum-likelihood analysis of sequences with overlapping reading frames. *Mol Biol Evol* 18:691-699.
- Poe S. 1998. Sensitivity of phylogeny estimation to taxonomic sampling. *Syst Biol* 47:18-31.
- Poe S. and Swofford D.L. 1999. Taxon sampling revisited. *Nature* 398:299-300.
- Pollock D.D., Zwickl D.J., McGuire J.A., and Hillis D.M. 2002. Increased taxon sampling is advantageous for phylogenetic inference. *Syst Biol*: in press.
- Posada D. 2001. The effect of branch length variation on the selection of models of molecular evolution. *J Mol Evol* 52:434-444.
- Posada D. 2002. Evaluation of methods for detecting recombination from DNA sequences: Empirical data. *Mol Biol Evol* 19: in press.
- Posada D. and Crandall K.A. 1998. Modeltest: Testing the model of DNA substitution. *Bioinformatics* 14:817-818.
- Posada D. and Crandall K.A. 2001a. A comparison of different strategies for selecting models of DNA substitution. *Syst Biol* 50:580-601.
- Posada D. and Crandall K.A. 2001b. Evaluation of methods for detecting recombination from DNA sequences: Computer simulations. *Proc Natl Acad Sci USA* 98:13757-13762.
- Posada D. and Crandall K.A. 2001c. Intraspecific gene genealogies: trees grafting into networks. *Trends Ecol Evol* 16:37-45.
- Posada D. and Crandall K.A. 2001d. Selecting models of nucleotide substitution: An application to Human Immunodeficiency Virus 1 (HIV-1). *Mol Biol Evol* 18:897-906.
- Posada D. and Crandall K.A. 2002. The effect of recombination on the accuracy of phylogeny estimation. *J Mol Evol* 54:396-402.
- Posada D., Crandall K.A., and Hillis D.M. 2001. Phylogenetics of HIV. In *Computational and Evolutionary Analysis of HIV Molecular Sequences*, p. 121-160, Rodrigo A.G. and Learn G.H. Jr., eds. Kluwer Academic Publishers, Dordrecht, The Netherlands.
- Posada D., Crandall K.A., and Holmes E.C. 2002. Recombination in evolutionary genomics. *Annu Rev Genet*: in press.
- Posada D., Crandall K.A., Nguyen M., Demma J.C., and Viscidi R.P. 2000. Population genetics of the *porB* gene of *Neisseria gonorrhoeae*. *Mol Biol Evol*:423-436.
- Rambaut A. 2002 *Se-Al: Sequence Alignment Editor*, Department of Zoology, University of Oxford (<http://evolve.zoo.ox.ac.uk>).
- Rich S.M., Sawyer S.A., and Barbour A.G. 2001. Antigen polymorphism in *Borrelia hermsii*, a clonal pathogenic bacterium. *Proc Natl Acad Sci USA* 98:15038-15043.
- Robertson D.L., Hahn B.H., and Sharp P.M. 1995. Recombination in AIDS viruses. *J Mol Evol* 40:249-259.
- Rosenberg M.S. and Kumar S. 2001. Incomplete taxon sampling is not a problem for phylogenetic inference. *Proc Natl Acad Sci USA* 98:10751-10756.

- Rzhetsky A. and Nei M. 1992. A simple method for estimating and testing minimum-evolution trees. *Mol Biol Evol* 9:945-967.
- Salter L.A. 2001. Complexity of the likelihood surface for a large DNA dataset. *Syst Biol* 50:970-978.
- Sanderson M.J. and Wojcicichowski M.F. 2000. Improved bootstrap confidence limits in large-scale phylogenies, with an example from Neo-Astragalus (Leguminosae). *Syst Biol* 49:671-685.
- Schierup M.H. and Hein J. 2000. Consequences of recombination on traditional phylogenetic analysis. *Genetics* 156:879-891.
- Sharp P.M. 1997. In search of molecular Darwinism. *Nature* 385:111-112.
- Shimodaira H. and Hasegawa M. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol Biol Evol* 16:1114-1116.
- Strimmer K. and Moulton V. 2000. Likelihood analysis of phylogenetic networks using directed graphical methods. *Mol Biol Evol* 17:875-881.
- Sullivan J., Swofford D.L., and Naylor G.J.P. 1999. The effect of taxon sampling on estimating rate heterogeneity parameters of maximum-likelihood models. *Mol Biol Evol* 16:1347-1356.
- Swofford D.L. 2000 PAUP*. *Phylogenetic Analysis Using Parsimony (*and Other Methods)*. Sinauer Associates, Sunderland, PA.
- Swofford D.L., Olsen G.J., Waddell P.J., and Hillis D.M. 1996. *Phylogenetic Inference*. In *Molecular Systematics*, p. 407-514, Hillis D.M., Moritz C., and Mable B.K., eds. Sinauer Associates, Inc., Sunderland, MA.
- Templeton A.R. 1983. Phylogenetic inference from restriction endonuclease cleavage site maps with particular reference to the evolution of humans and the apes. *Evolution* 37:221-244.
- Templeton A.R. 1992. Human origins and analysis of mitochondrial DNA sequences. *Science* 255:737.
- Templeton A.R., Crandall K.A., and Sing C.F. 1992. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. *Genetics* 132:619-633.
- Templeton A.R., Routman E., and Phillips C.A. 1995. Separating population structure from population history: a cladistic analysis of geographical distribution of mitochondrial DNA haplotypes in the tiger salamander, *Ambystoma tigrinum*. *Genetics* 140:767-782.
- Thompson J.D., Gibson T.J., Plewniak F., Jeanmougin F., and Higgins D.G. 1997. The clustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* 24:4876-4882.
- Wiuf C., Christensen T., and Hein J. 2001. A simulation study of the reliability of recombination detection methods. *Mol Biol Evol*: in press.
- Woolley S., Johnson J., Smith M.J., Crandall K.A., and McClellan D.A. 2002. TreeSAAP: A phylogenetic approach to identifying selective influences on amino acid properties. *Bioinformatics*: submitted.
- Yang Z. 1994. Estimating the pattern of nucleotide substitution. *J Mol Evol* 39:105-111.
- Yang Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol Evol* 11:367-372.
- Yang Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol Biol Evol* 15:568-573.
- Yang Z. 2001 PAML: *Phylogenetic Analysis by Maximum Likelihood*. University College London, London.
- Yang Z. and Bielawski J.P. 2000. Statistical methods for detecting molecular adaptation. *Trends Ecol Evol* 15:496-503.
- Yang Z. and Nielsen R. 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *J Mol Evol* 46:409-418.
- Yang Z. and Nielsen R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol Biol Evol*: in press.
- Yang Z., Nielsen R., Goldman N., and Pedersen A.-M. K. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 155:431-449.
- Zanotto P.M., Kallas E.G., Souza R.F., and Holmes E.C. 1999. Genealogical evidence for positive selection in the *nef* gene of HIV-1. *Genetics* 153:1077-1089.
- Zhang J. and Madden T.L. 1997. PowerBLAST: A new network BLAST application for interactive or automated sequence analysis and annotation. *Genome Research* 7:649-656.
- Zharkikh A. and Li W.-H. 1995. Estimation of confidence in phylogeny: The complete-and-partial bootstrap technique. *Mol Phylog Evol* 4:44-63.