Models of protein sequence evolution and their applications

Jeffrey L Thorne

Homologous sequences are correlated due to their common ancestry. Probabilistic models of sequence evolution are employed routinely to properly account for these phylogenetic correlations. These increasingly realistic models provide a basis for studying evolution and for exploiting it to better understand protein structure and function. Notable recent advances have been made in the treatment of insertion and deletion events, the estimation of amino-acid replacement rates, and the detection of positive selection.

Addresses
Program in Statistical Genetics, Statistics Department, Box 8203, North Carolina State University, Raleigh, North Carolina 27695-8203, USA; e-mail: thorne@statgen.ncsu.edu

Current Opinion in Genetics & Development 2000, **10**:602–605

0959-437X/00/$ – see front matter

Introduction

Many areas of biological research concern the state of a biological system at one instant or during a short period of time; however, evolution is largely an historical science. In addition to characterizing the state of a biological system at a particular instant, the goal of evolutionary studies is to follow the state of the system through long periods of time. Historically important evolutionary events tend to be either difficult or impossible to observe or repeat in the laboratory. Although exciting research regarding the evolutionary process can sometimes be performed via well-designed experiments on organisms with short generation times (e.g. see [1]), many aspects of the evolutionary process are not easily studied with controlled experimentation. Typically, an evolutionary biologist is faced with the task of extracting as much information as possible from a data set that was generated under an uncontrolled natural process. In these cases, the method of data analysis can be crucial to success of the research project. Therefore, statistical methodology has a central role in molecular evolutionary research.

In recent years, newly proposed techniques for studying molecular evolution have been based mainly on probabilistic models. Models of molecular evolution are highly simplified descriptions of the history and process of sequence change. The history is represented by the phylogenetic tree topology and the expected amount of evolution on each branch (i.e. the branch lengths). The observed sequences are at the tips of this tree, whereas the sequences that correspond to other positions on the tree are almost always unobservable. The fact that only sequences representing tips of the tree can be observed makes molecular evolution both interesting and difficult to study.

To represent the process of evolution, probabilistic models are based on explicit assumptions that govern the instantaneous rates of change from each possible sequence to each other possible sequence. Specification of the rates of sequence changes determines the probability of change, after some period of evolution, from one sequence to each other possible sequence. The ability to calculate these transition probabilities permits data sets to be analyzed in the maximum likelihood and Bayesian frameworks. A detailed introduction to these statistical frameworks will not be given here. Swofford *et al.* [2] provide an excellent coverage of maximum likelihood techniques for studying molecular evolution. An overview of Bayesian techniques for genetic data analysis is also available [3].

Model-based data analysis tools are powerful because the assumptions regarding how data were generated are built directly into them. When more biological knowledge is incorporated into models in the form of more realistic assumptions, more biological information can be extracted from a data set. Comparative genomic and proteomic data are generated by complicated evolutionary processes. Over time, models of these processes are becoming more realistic and are thereby yielding better data analysis tools.

In this review, I focus on recent progress in the development of models for sequence evolution. Although protein evolution is emphasized, many of the approaches described here are also pertinent to the study of non-coding sequence evolution.

Insertions and deletions

Most effort for modeling sequence evolution has been concentrated on the processes of nucleotide substitution and amino-acid replacement. As insertion and deletion events have proven difficult to model, conventional methods for phylogeny reconstruction require a known sequence alignment and neglect alignment uncertainty. Alignment columns with gaps are either removed from the analysis or are treated in an ad hoc fashion. As a result, evolutionary information from insertions and deletions is typically ignored during phylogeny reconstruction. The difficulty with insertions and deletions may be the most vexing problem for model-based approaches to studying sequence evolution.

Attempts have been made to deal with alignment uncertainty during phylogeny reconstruction [4,5]. These approaches are attractive but their probabilistic treatments of gaps in alignments have not been reconciled with an explicit stochastic evolutionary model of insertion and deletion. Although a satisfying and computationally tractable way to deal with alignment uncertainty during phylogeny reconstruction does not appear imminent, progress is being made. Steel and Hein [6••] have successfully combined sequence alignment and phylogeny

reconstruction for the special case of a star phylogeny (i.e. the special case where all branches on an evolutionary tree intersect at a single point). In principle, their method can be extended to a more complicated tree structure (J Hein, personal communication) but advances must be made before this extension is computationally feasible.

The progress with treatment of insertions and deletions is also likely to impact other areas of biology. For example, one issue that frequently arises is whether two sequences are evolutionarily homologous. This is inherently interesting and also pertinent because homologous sequences often share biological function. Hein *et al.* [7] have demonstrated that incorporating insertions and deletions into an evolutionary model can yield a statistically rigorous, computationally tractable, and potentially sensitive test of sequence homology.

Amino-acid replacement

In contrast to the insertion and deletion processes, models of amino-acid replacement and nucleotide substitution are better developed. Dayhoff *et al.* [8,9] proposed the most influential model of amino-acid replacement. Their simple model assumes that all sites in a protein evolve independently of one another. At each site, the process of amino-acid replacement is defined by a matrix of replacement rates. For each possible change from one of the twenty amino acid types to another, there is a corresponding matrix entry. With the Dayhoff model, all sites evolve according to the same rate matrix.

The Dayhoff procedure for estimating replacement rates relies on comparisons between sequences that are very similar. Comparisons involving sequences that are somewhat dissimilar, and are therefore likely to be less closely related, are discarded. Instead of relying on only closely related sequences, some studies have employed maximum likelihood techniques to estimate replacement rates from large data sets (e.g. [10–12]). Because of the relatively high number of parameters needed to estimate rates of replacement from each amino acid type to each other type, the maximum likelihood approaches are computationally demanding and are sometimes intractable. Recently, a new approach for empirically estimating amino-acid replacement rates has been introduced [13••]. As with the Dayhoff technique, this approach is computationally undemanding. In contrast to the Dayhoff approach, the new approach is not restricted to comparisons of closely related sequences.

Variation of evolutionary processes among sequence sites

A limitation of the Dayhoff model is the assumption that all sites in a protein evolve according to the same rate matrix. This assumption has been relaxed in various ways. Most significantly, Yang [14] devised a clever and computationally tractable strategy for allowing different sites in a sequence to evolve at different rates. An appealing feature of this technique is that both quickly and slowly evolving sites do not need to be identified prior to data analysis.

Almost invariably, evolutionary models that incorporate rate heterogeneity among sites are found to fit data better than the corresponding models with rate homogeneity. In addition, phylogeny reconstruction seems to be more successful when heterogeneity is permitted. Unfortunately, the biological underpinnings of rate heterogeneity are not well understood.

The Yang strategy assumes that, except for rate heterogeneity, all sites evolve according to the same process. In reality, the variation of evolutionary processes among sites is not solely attributable to variation in the absolute rates of evolution among sites. Bruno [15] has incorporated variation of preferred residue types among sites into models of amino-acid replacement. Variation of preferred residues among sites leads to rate variation among sites. For example, changes might be infrequent at a site where there is strong selection pressure favoring cysteine but changes might be more frequent at a site where natural selection does not favor any particular amino acid. This modeling approach is applicable to data sets with many diverse sequences but may be overly parameter-rich for other data sets.

Another strategy for permitting variation of preferred residues among sites is to define several categories of sites. A recently introduced method [16•] starts with the assumption that each site belongs to exactly one category and that each category has certain preferred physicochemical attributes of bulk and hydrophobicity. The replacement rate matrix for each category is structured so that amino acid types with attributes that are similar to those favored by the category will more frequently occupy sites within that category. Although the category to which each site belongs is treated as unknown, the preferred attributes of the categories can be estimated.

Protein evolution and protein structure

Some of the heterogeneity of amino-acid replacement rates is associated with variation of structural environments among sites. Although protein secondary structure has an influence on amino acid replacement rates, solvent accessibility has a stronger correlation with them. Replacement rates at sites on the surface of globular proteins are about twice the rates at sites that are less accessible to solvent [17]. Separate amino-acid replacement models can be constructed for each of a variety of different structural environments [18–20]. Most effort has been expended on relating patterns of amino acid replacement to structural environment in globular proteins. Recently, models have been developed for describing the evolution and organization of transmembrane proteins [21].

Protein secondary structure and solvent accessibility are regional properties of a protein. For instance, if one protein site is in an α-helical conformation, then adjacent sites are also prone to being in α-helical environments. By coupling models of amino-acid replacement for different structural environments with models for the organization of structural environments along a protein sequence, a coherent statistical framework for protein secondary structure can be

developed. This framework has the advantage of being able to properly account for the phylogenetic correlations among a group of aligned protein sequences that are assumed to share a common but unknown underlying secondary structure [22]. Unfortunately, this explicit evolutionary approach for protein secondary structure prediction is not yet competitive with the best approaches (e.g. [23]). One explanation for the reduced success of the evolutionary approach is its inability to satisfactorily deal with insertions and deletions. Evolutionary lineages experience insertions and deletions at higher rates in coil regions of proteins than in α-helices or β-sheets [24,25] but this tendency has not yet been incorporated into evolutionary models.

Codon-based models

A limitation of all the aforementioned models of amino-acid replacement is that they are based on changes among 20 states, where each state represents an amino acid type. In reality, evolution occurs at the level of DNA sequences. Therefore, it is preferable to frame models of sequence evolution in terms of codons rather than in terms of amino acids [26–28]. To date, most codon-based models have employed the assumption that all changes to a codon involve only one of the three codon positions.

In the evolution literature, some changes to a codon are termed 'synonymous' because they do not alter the amino acid specified by the codon. Other changes are termed 'non-synonymous' because they do result in a different amino acid being specified by the codon. By contrasting synonymous and non-synonymous rates of change, natural selection can be studied. For example, if all amino acid replacements are selectively neutral, then rates of codon substitution would not depend on whether changes are synonymous or non-synonymous. Codon-based models provide a natural framework for estimating synonymous and non-synonymous rates [27–29].

Codons that undergo substantially more non-synonymous than synonymous changes are of particular interest. A surplus of non-synonymous changes relative to synonymous changes is known as 'positive selection' and seems to be characteristic of sites with interesting functional roles. Codon-based models for detecting positive selection are becoming increasingly refined [30,31•]. Examples of their applications include the study of viral evolution [30,31•] and plant–pathogen interactions [32,33].

Recently, codon-based models have become quite sophisticated. For example, they have been designed to reflect the depressed CpG levels that are sometimes observed [34]. In another study, the form of the codon substitution rate matrix was reconciled with a simple traditional population genetic model [35].

Evolutionary dependence among sites

Models of sequence change typically assume that different sequence units (e.g. nucleotides, amino acids, codons) evolve independently of one another. Hidden Markov models [36,37] and stochastic grammars [38] have been adapted to permit simple correlations of evolutionary processes among sites but progress with more complicated forms of dependence among sites is just beginning [39•].

A first step toward modelling the dependence may be to identify pairs of sites that evolve in a dependent fashion. These may be sites that are functionally related. One way for determining whether sites evolve in a correlated fashion is to develop a measure for the amount of covariation between the amino acids in two columns of a sequence alignment [40]. Simulations can then be performed to determine if the observed value for this measure is extreme in relation to values that would be observed if the two sites actually evolved independently. Another approach for identifying correlated evolution between two sites is to build a simple model for correlated evolution and then compare its fit to that of a model that does not allow correlated evolution between the two sites [41•].

Conclusions

Advances in the modelling of protein evolution are ongoing, exciting, and occurring at an ever more rapid rate. Substantial effort is being made to understand the relationship between protein structure and protein evolution. The ability of codon-based models to link changes at the DNA level to changes at the protein level is increasingly being exploited. In addition, improved treatment of insertions and deletions is on the horizon.

The relationship between evolutionary models and evolutionary knowledge is synergistic. Increased evolutionary knowledge leads to better models. Better models provide better tools for data analysis, and better models thereby lead to an enhanced understanding of evolution.

Acknowledgements

I thank N Goldman for his insightful suggestions. This work was supported by National Institutes of Health grant GM45344.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Bull JJ, Badgett MR, Wichman HA: **Big-benefit mutations in a bacteriophage inhibited with heat.** *Mol Biol Evol* 2000, **17**:942-950.
2. Swofford DL, Olsen GJ, Waddell PJ, Hillis DM: **Phylogenetic inference.** In *Molecular Systematics*, edn 2. Edited by Hillis DM, Moritz C, Mable BK. Sunderland, Massachusetts: Sinauer Associates; 1996:407-514.
3. Shoemaker JS, Painter IS, Weir BS: **Bayesian statistics in genetics — a guide for the uninitiated.** *Trends Genet* 1999, **15**:354-358.
4. Allison L, Wallace CS: **The posterior probability distribution of alignments and its application to parameter estimation of evolutionary trees and to optimization of multiple alignments.** *J Mol Evol* 1994, **39**:418-430.
5. Mitchison GJ: **A probabilistic treatment of phylogeny and sequence alignment.** *J Mol Evol* 1999, **49**:11-22.

6. Steel M, Hein JJ: **A generalisation of the Thorne-Kishino-**
•• **Felsenstein model of statistical alignment to k sequences related by a star tree.** *Appl Math Lett* 2000, in press.
Sequence alignment and phylogeny reconstruction are two of the central problems in computational biology. Although it is based on an oversimplified evolutionary model of insertion and deletion, this paper represents substantial progress toward the joint solution of both problems.

7. Hein J, Wiuf C, Knudsen B, Moller MB, Wibling G: **Statistical alignment: computational properties, homology testing and goodness-of-fit.** *J Mol Biol* 2000, **302**:265-279.

8. Dayhoff MO, Eck RV, Park CM: **A model of evolutionary change in proteins.** In *Atlas of Protein Sequence and Structure*, vol 5. Edited by Dayhoff MO. Washington DC: National Biomedical Research Foundation; 1972:89-99.

9. Dayhoff MO, Schwartz RM, Orcutt BC: **A model of evolutionary change in proteins.** In *Atlas of Protein Sequence and Structure*, vol 5, suppl 3. Edited by Dayhoff MO. Washington DC: National Biomedical Research Foundation; 1978:345-352.

10. Adachi J, Hasegawa M: **Model of amino acid substitution in proteins encoded by mitochondrial DNA.** *J Mol Evol* 1996, **42**:459-468.

11. Yang Z, Nielsen R, Hasegawa M: **Models of amino acid substitution and applications to mitochondrial protein evolution.** *Mol Biol Evol* 1998, **15**:1600-1611.

12. Adachi J, Waddell PJ, Martin W, Hasegawa M: **Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA.** *J Mol Evol* 2000, **50**:348-358.

13. Müller T, Vingron M: **Modeling amino acid replacement.** *J Comp Biol*
•• 2000, in press.
A novel approach for estimating replacement-rate matrices from sets of pairs of protein sequences. This approach is computationally feasible and seems to have desirable statistical properties. It is likely to be frequently applied.

14. Yang Z: **Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods.** *J Mol Evol* 1994, **39**:306-314.

15. Bruno WJ: **Modeling residue usage in aligned protein sequences via maximum likelihood.** *Mol Biol Evol* 1996, **13**:1368-1374.

16. Koshi JM, Mindell DP, Goldstein RA: **Using physical-chemistry**
• **based mutation models in phylogenetic analyses of HIV-1 subtypes.** *Mol Biol Evol* 1999, **16**:173-179.
Early indications are that this physicochemical-based strategy of modelling amino acid replacement may fit data quite well but a more extensive evaluation is needed. The authors illustrate their modelling strategy by reconstructing the evolutionary relationships among HIV-1 subtypes and they highlight disparities between their inferred HIV-1 phylogeny and earlier results that were based on less-realistic evolutionary models.

17. Goldman N, Thorne JL, Jones DT: **Assessing the impact of secondary structure and solvent accessibility on protein evolution.** *Genetics* 1998, **149**:445-458.

18. Koshi JM, Goldstein, RA: **Context-dependent optimal substitution matrices.** *Protein Eng* 1995, **8**:641-645.

19. Koshi JM, Goldstein RA: **Mutation matrices and physical-chemical properties: correlations and implications.** *Proteins* 1996, **27**:336-344.

20. Thorne JL, Goldman N, Jones DT: **Combining protein evolution and secondary structure.** *Mol Biol Evol* 1996, **13**:666-673.

21. Liò P, Goldman N: **Using protein structural information in evolutionary inference: transmembrane proteins.** *Mol Biol Evol* 1999, **16**:1696-1710.

22. Goldman N, Thorne JL, Jones DT: **Using evolutionary trees in protein secondary structure prediction and other comparative sequence analyses.** *J Mol Biol* 1996, **263**:196-208.

23. McGuffin LJ, Bryson K, Jones DT: **The PSIPRED protein structure prediction server.** *Bioinformatics* 2000, **16**:404-405.

24. Benner SA, Gerloff D: **Patterns of divergence in homologous proteins as indicators of secondary structure and tertiary structure: a prediction of the structure of the catalytic domain of protein kinases.** *Adv Enz Reg* 1991, **31**:121-181.

25. Thornton JM, Flores TP, Jones DT, Swindells MB: **Prediction of progress at last.** *Nature* 1991, **354**:105-106.

26. Schöniger M, Hofacker GL, Borstnik B: **Stochastic traits of molecular evolution—acceptance of point mutations in native actin genes.** *J Theor Biol* 1990, **143**:287-306.

27. Goldman N, Yang Z: **A codon-based model of nucleotide substitution for protein-coding DNA sequences.** *Mol Biol Evol* 1994, **11**:725-736.

28. Muse SV, Gaut BS: **A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with applications to the chloroplast genome.** *Mol Biol Evol* 1994, **11**:715-724.

29. Muse SV: **Estimating synonymous and nonsynonymous substitution rates.** *Mol Biol Evol* 1996, **13**:105-114.

30. Nielsen R, Yang Z: **Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene.** *Genetics* 1998, **148**:929-936.

31. Yang Z, Nielsen R, Goldman N, Pedersen AMK: **Codon-substitution**
• **models for heterogeneous selection pressure.** *Genetics* 2000, **155**:431-449.
A thorough examination of several novel schemes for incorporating positive selection into codon-based models of protein evolution. By analyzing a wide range of data sets, the authors convincingly demonstrate that existence of codons undergoing positive selection can be statistically detected even when, for the vast majority of codons, amino-acid replacements are either neutral or deleterious.

32. Bishop JG, Dean AM, Mitchell-Olds T: **Rapid evolution in plant chitinases: molecular targets of selection in plant–pathogen coevolution.** *Proc Natl Acad Sci USA* 2000, **97**:5322-5327.

33. Stotz HU, Bishop JG, Bergmann CW, Koch M, Albersheim P, Darvill AG, Labavitch JM: **Identification of target amino acids that affect interactions of fungal polygalacturonases and their plant inhibitors.** *Phys and Mol Plant Path* 2000, **56**:117-130.

34. Pedersen AMK, Wiuf C, Christiansen FB: **A codon-based model designed to describe lentiviral evolution.** *Mol Biol Evol* 1998, **15**:1069-1081.

35. Halpern A, Bruno WJ: **Evolutionary distances for protein-coding sequences: Modeling site-specific residue frequencies.** *Mol Biol Evol* 1998, **15**:910-917.

36. Yang Z: **A space-time process model for the evolution of DNA sequences.** *Genetics* 1995, **139**:993-1005.

37. Felsenstein J, Churchill GA: **A hidden Markov model approach to variation among sites in rate of evolution.** *Mol Biol Evol* 1996, **13**:93-104.

38. Knudsen B, Hein J: **RNA secondary structure prediction using stochastic context-free grammars and evolutionary history.** *Bioinformatics* 1999, **15**:446-454.

39. Jensen JL, Pedersen AMK: **Probabilistic models of DNA sequence**
• **evolution with context dependent rates of substitution.** *Adv in Appl Prob* 2000, **32**:499-517.
This is probably the most advanced treatment of evolutionary dependence among sequence sites. Its value will increase when the approach is extended to the analysis of data sets consisting of more than two sequences.

40. Wollenberg KR, Atchley WR: **Separation of phylogenetic and functional associations in biological sequences by using the parametric bootstrap.** *Proc Natl Acad Sci USA* 2000, **97**:3288-3291.

41. Pollock DD, Taylor WR, Goldman N: **Coevolving protein residues:**
• **maximum likelihood identification and relationship to structure.** *J Mol Biol* 1999, **287**:187-198.
In this paper, explicit models for the evolutionary covariation among protein sites form the basis for methods to detect sites that covary. Advantages of this model-based approach include the ability to carefully incorporate biological knowledge, the ability to estimate the amount of covariation, and enhanced statistical power for the detection of covarying sites. A disadvantage of explicitly modelling evolutionary covariation to detect covarying sites is that the possible nature of the covariation between sites must be specified *a priori*.