Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Biomedical Informatics 39 (2006) 51–61
Journal of Biomedical Informatics
www.elsevier.com/locate/yjbin

Methodological Review

Phylogenetics by likelihood: Evolutionary modeling as a tool for understanding the genome

Carolin Kosiol, Lee Bofkin, Simon Whelan*

EMBL—European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

Received 28 February 2005
Available online 15 September 2005

Abstract

Molecular evolutionary studies provide a means of investigating how cells function and how organisms adapt to their environment. The products of evolutionary studies provide medically important insights to the source of major diseases, such as HIV, and hold the key to understand the developing immunity of pathogenic bacteria to antibiotics. They have also helped mankind understand its place in nature, casting light on the selective forces and environmental conditions that resulted in modern humans. The use of likelihood as a framework for statistical modeling in phylogenetics has played a fundamental role in studying molecular evolution, enabling rigorous and robust conclusions to be drawn from sequence data. The first half of this article is a general introduction to the likelihood method for inferring phylogenies, the properties of the models used, and how it can be used for statistical testing. The latter half of the article focuses on the emerging new generation of phylogenetic models that describe heterogeneity in the evolutionary process along sequences, including the recoding of protein coding sequence data to amino acids and codons, and various approaches for describing dependencies between sites in a sequence. We conclude with a detailed case study examining how modern modeling approaches have been successfully employed to identify adaptive evolution in proteins.

Keywords: Phylogenetics; Evolution; Selection; Likelihood; Markov model; Context dependency; HMM

1. Introduction

The Human Genome Project has resulted in revolutionary changes to biology and medicine [1]. The international human genome sequencing consortium finished the draft sequence in June 2000, and finally completed the human genome sequence in April 2003, the 50th anniversary of the discovery of the DNA [2]. Together with a variety of molecular sequences of non-human species, this very large quantity of data is publicly available for comparative studies and subject to biomedical interpretation. The computational analysis of molecular sequence data is playing an increasingly important role in biomedical science [3–9]. The comparison of evolutionarily related sequences has proved an effective tool in numerous research areas, including finding novel functional structures in genomes [1,5,7–9], the detection of homologues within and between genomes [1,5,7,8,10], protein structure prediction [11,12], and the elucidation of how biochemical molecules function [13–15]. The statistical modeling of evolution is a powerful approach for studying how genomes function and how they evolve. Recent improvements in models have improved the estimation of evolutionary relationships [16–18], and enabled the elucidation of complex, biomedically important phenomena through sophisticated model construction and comparison [13,19,20]. The application of models to sequence data requires a general framework for inference, usually either by the likelihood approach, discussed here, or through Bayesian analysis.

Maximum likelihood (ML) is a long established method for statistical inference [21,22], extensively tested for many years and successfully applied to a wide variety of problems, ranging from classical population genetics to the modeling of world economies [22]. The likelihood value,

* Corresponding author. Fax: +44 1223 494468.
E-mail address: simon@ebi.ac.uk (S. Whelan).

1532-0464/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2005.08.003

L, used in phylogenetic inference is the probability of observing the data (e.g., a set of aligned nucleotide sequences) under a given phylogenetic tree and a specified model of evolution: $L = \Pr(\text{data}|\text{tree, model})$. ML is used in phylogenetics to find the optimal set of parameters contained within the tree and model that best describes the observed data [6]. The tree describes the topology of the evolutionary relationships between the sequences and a set of branch lengths describing how much evolution has occurred in different regions of the tree. The model contains a set of parameters explicitly describing the evolutionary process; for example, the rate of transition mutation ($A \leftrightarrow G$ and $C \leftrightarrow T$) in DNA [23].

This first half of this review aims to provide a general overview of likelihood approaches in phylogenetics, including tree topology estimation, the properties of evolutionary models, and the use of statistical testing to investigate interesting biological questions. These topics are more comprehensively discussed in other recently published review articles and books [4,6,24]. The latter half focuses on some of the latest innovations for modeling variation in the evolutionary process along sequences, demonstrating how advances in statistical modeling are allowing progressively more biomedically important information to be extracted from sequence data. We conclude by discussing the methodology used for the detection of molecular adaptation, where new approaches have already proved valuable.

2. Tree topology estimation

A bifurcating tree is usually used to describe the statistical dependencies present in biological sequence data resulting from evolutionary relatedness, and the estimation of this topology remains the primary objective for many phylogenetic studies. A tree structure imposes a series of assumptions; most importantly, these include all sequences sharing a common ancestor, and branches of the evolutionary tree evolving independently of one another. Violations in the former occur when paralogous regions are included in a data set; for example, when only a subset of the domains present in a protein is shared between all of the sequences under consideration. The latter assumption is violated when large-scale mutational events occur, including recombination [25], gene conversion [20], or horizontal transfer [26]. Before performing any type of phylogenetic analysis one should try to ensure that the assumptions implicit through using a tree structure are not violated.

ML, in common with other statistical methods, offers an effective and robust way to obtain a topology estimate and to measure our confidence in that estimate [6,16,27]. The optimal topology is that with the highest likelihood, and finding it requires calculating the likelihood of all topologies. In practice, this is impractical for even relatively modest numbers of sequences, and tree estimation is performed using heuristic algorithms. For example, for 20 sequences there are approximately 2×10^{20} potential topologies, a number too great for even the most efficient computer program to work through in a reasonable time. Heuristic methods are not certain to find the globally optimal topology [6,28]. Instead, they rely on hill-climbing optimization techniques [6], such as Nearest Neighbor Interchange (NNI) and Tree Bisection-Reconnection (TBR), which may become stuck in local optima. These algorithms function by making small rearrangements from a given candidate tree to propose new topologies. The likelihood of each new topology is calculated and that with the highest value is the candidate tree for the next iteration. This process is repeated until no improvements in topology can be found and the final candidate is taken to be the optimal tree topology. The results from these heuristics vary depending on the original candidate tree and it is advisable to repeat the estimation procedure from different starting trees. Other heuristics based on hierarchical clustering approaches are also popular; some of these are based on full likelihood approaches (e.g., star-decomposition and stepwise addition [6]), whilst others are based solely on pair-wise distance estimates (e.g., neighbor joining and its derivatives [6,29]). There are numerous software available for performing heuristic searches and it is beyond the scope of this article to detail them; Joe Felsenstein maintains a large repository of popular programs at the URL: http://evolution.genetics.washington.edu/phylip/software.html

There is no uniformly best method or program for phylogeny estimation because performance is highly dependent on the data under consideration. Once a particular evolutionary model is chosen for tree estimation, it is advisable to build a list of candidate tree topologies using as many different heuristics, with as many different starting trees, as feasible. In practice, some software may not contain the model chosen for tree estimation. In these cases, it is advisable to choose a set of models available in the program that closely resemble the chosen tree estimation criteria and add all of their results to the candidate list. This list should then be assessed using a single, reliable program implementing the chosen tree estimation criteria and the topology with the highest likelihood is the final estimate. The use of a single program to assess all candidate topologies is necessary because likelihood computation can differ between software; for example, some remove all columns in an alignment with gaps, or optimize to different degrees of rigor.

The effect of using an incorrect topology estimate depends on the purpose of the study. Where the primary aim is to investigate the evolutionary relationship between a group of sequences or organisms, the effect of an error is obvious. When the quantity of interest is a parameter of an evolutionary model, the effect of a topological error is harder to interpret, often introducing a poorly characterized bias that can lead to inaccurate inferences. Recently, for example, small errors in tree topology have been the subject for contested claims of positive natural selection in proteins [30,31]. A general rule of thumb that has been successfully employed in many studies is to assume that reasonable topology estimates lead to reasonable parame-

ter estimates. Providing the errors are restricted to relatively few short branches, where only a small number of evolutionary changes have occurred, this argument is likely to hold [16,32]. As a note of caution, however, small errors in estimates will occasionally lead to seriously misleading inferences and every effort should be made to obtain the best possible tree estimate whenever performing a phylogenetic study.

3. Evolutionary modeling

The statistical modeling of the evolutionary process is of great importance when performing phylogenetic studies [6]. When comparing reasonably divergent sequences, counting the raw sequence identity (percentage of sites with observed changes) underestimates the amount of evolution that has occurred because, by chance alone, some sites will have incurred multiple substitutions. The probabilistic models used in ML provide more accurate evolutionary distance estimates by accounting for these unobserved changes, becoming more important as sequence divergence increases and the disparity between sequence identity and evolutionary distance grows. In the extreme case, when all sites have undergone numerous changes, the sequences are effectively random and no longer contain any evolutionary information; an event often referred to as saturation. In phylogenetics, models describe evolution as a series of random mutational events, and contain an explicit description of the rate that individual characters (such as A, C, G, and T in DNA) replace each other. Given a branch length, these relative rates are used to calculate the probabilities of characters either remaining the same or replacing each other, and, using the pruning algorithm of Felsenstein [33], are used to calculate the likelihood, L. The parameters comprising the tree and model are estimated using numerical optimization procedures to find the highest likelihood, which represents the combination of parameter values that best describes the observed data.

The models used in phylogenetics often make biologically relevant assumptions about the evolutionary process [4,6,24]. They come from a special class of statistical models called Markov processes, which assume that the rate that a site changes depends only on its current state and not on previous ancestral states. This assumption is reasonable because, during evolution, mutation and natural selection can only act upon the molecules present in an organism and have no knowledge of what came previously. Additional explicit assumptions about sequence evolution are often imposed upon this Markov process, some reflecting beliefs about the underlying processes influencing molecular evolution, whilst others are mathematical conveniences that enable efficient likelihood calculation. The sites in a sequence are often assumed to change to the same evolutionary process and, in simple models, at the same overall rate. This implies that all sites are independent of one another and have the same evolutionary constraints. There is strong evidence that the independence of sites assumption is frequently violated, but it is still commonly accepted because of its computational benefit. Modern approaches to dealing with different types of evolutionary dependencies within sequences are discussed in more detail in later sections. Other widely used assumptions are that the evolutionary process is the same through time (time homogeneity), that the relative frequencies of characters in the data do not change over time (stationarity), and that the evolutionary process looks the same going forward and backwards (reversibility). These assumptions broadly hold, although there are notable exceptions, and are useful in ensuring models remain relatively simple and biologically interpretable.

The remaining set of common assumptions are model dependent and define the parameters describing the relative rates of change between different character states in the model (exchangeability parameters); for example, the rate transition mutations occur relative to transversion mutations in DNA [23,34]. In a phylogenetic analysis, parameters are either estimated using ML for each dataset (mechanistic parameters), or set to previously estimated values from very large, representative datasets (empirical parameters). Mechanistic parameters are useful for describing factors in the evolutionary process that vary greatly between datasets, such as selective pressures [13] and the frequency of character states in the data (e.g., the relative occurrence of amino acids) [35,36]; their estimated values provide insights about the evolution of specific sequences. On the other hand, empirical parameters are useful when there are large numbers of parameters and/or when specific factors in the evolutionary process are expected to be similar between data sets. Many evolutionary models contain a mix of mechanistic and empirical parameters. For example, in models of protein evolution, the 190 amino acid exchangeability parameters are empirical [32,37], because it is impractical to estimate this number of parameters from the majority of protein data sets. Models consisting purely of empirical parameters have proved very useful for other topics related to phylogenetics, including homology detection (e.g., the PAM [37] and BLOSUM [38] matrices used by BLAST [39]), and sequence alignment (e.g., CLUSTAL contains implicit descriptions of character replacement [40]).

4. Statistical testing

Likelihood inference can be used to address many biologically important questions by examining parameters estimated from the data or by comparing how well similar models explain sequence evolution. For example, parameter values may be used to identify those data that contain the most extreme transition mutation bias amongst a set of alignments, those that are the most conserved, and those that have the greatest selective constraints. One of the most appealing features of ML estimation is that it provides a long-established method for statistical inference [6,21,22]. In addition to providing accurate point estimates of

parameters, it also gives information about the uncertainty of our estimates through the calculation of confidence intervals (CIs), which allows the rigorous comparison of competing hypotheses. CIs are a simple measure of how much we trust parameters estimated from the data and are often provided with the output of phylogeny programs. A large CI suggests a parameter that is difficult to estimate, whilst a small CI is indicative of an accurate parameter estimate. The range of the CI can be used as a simple measure for testing hypotheses; for example, to test whether a parameter is not significantly different from 1.0, the 95% CI of the estimate can be examined and if it does not include 1.0, the hypothesis is rejected.

Likelihood also offers another very powerful way of comparing hypotheses, the likelihood ratio test (LRT) [6,21,22,41,42]. This requires the formation of two competing hypotheses, represented by models with different restraints on their parameters. For example, the relative frequency of transition mutations and transversion mutations in DNA evolution can be investigated through two competing hypotheses. The null hypothesis (H_0: likelihood L_0), describes the rate of transition and transversion mutation as equal, and the alternate hypothesis (H_1: likelihood L_1), has transitions occurring at a different rate to transversions. The ML values ($\hat{L}$) for the competing hypotheses are compared using the LRT statistic $2\Delta = 2\ln(\hat{L}_1/\hat{L}_0) = 2\{\ln(\hat{L}_1) - \ln(\hat{L}_0)\}$. This statistic has very useful properties for significance testing when certain conditions are met. Particularly, when H_0 can be formed by placing restrictions on the parameters in H_1, the hypotheses are said to be nested and for significance testing 2Δ can be compared to the 95% point of a χ^2_n distribution (where n is the number of parameters by which H_0 and H_1 differ). Many complex biological problems about the evolutionary process have been investigated using carefully constructing nested hypotheses, and the approach now plays a crucial role in many phylogenetic studies [4,13,43].

It is not possible to use χ^2 distributions for assessing the significance of LRTs under certain conditions, the most common occurring when comparing non-nested models. The rigorous comparison of hypotheses in this situation necessitates simulation methods for obtaining the required distribution for significance testing [4,6,44,45]. When a specific parameter (e.g., tree topology) is the focus for the study and not the choice of model, then information theoretic approaches can be used to choose between candidate models that estimate the parameter of interest. In general, these methods balance the complexity of the model (described as the number of parameters it contains, K) against the quality of description it provides through a given criterion (usually $\hat{L}$). A popular choice for performing such comparisons is An Information Criterion (AIC [46]), where under each model $\mathrm{AIC} = -2\ln\hat{L} + 2K$, and the model with the lowest AIC is chosen for subsequent analysis. Other popular information theoretic approaches include a corrected version of the AIC ($\mathrm{AIC_c}$ [47]) and the Bayesian Information Criterion (BIC [48]). These model choice methods have found wide usage in the model choice schema used in ModelTest [49] and ProtTest [50].

To this point, we have discussed statistical testing only in terms of the parameters in an evolutionary model. There are also established procedures for measuring confidence and comparing tree topologies [6,51,52]. It has been proved that ML is a consistent estimator of tree topology if the model that generated the data are used for analysis [27]. In other words, under the true model of evolution, $\hat{L}$ will converge towards the correct tree topology as the length of the sequences examined increases. In most studies, there will not be sufficient data for tight convergence and the tree estimate, like other statistical estimators, contains a degree of error. Currently, the four most popular approaches to quantifying this error are simple bootstrapping [53]; the Shimodaira–Hasegawa (SH) test [54,55]; the Approximately Unbiased (AU) test [55,56]; and the Swofford–Olsen–Waddell–Hillis (SOWH) test [51]. These approaches all address subtly different aspects of the same question. Bootstrapping, the simplest of all the tests, is performed on a per branch basis and is usually most appropriate when one wishes to assess whether certain partitions of the data, represented by branches in a tree, truly exist. Bootstrapping is not as useful for examining all of the branches of a tree, because it becomes difficult to interpret what all the individual bootstrap values mean. This form of bootstrapping to assess confidence is demonstrably biased [57–59], but remains a practical and useful tool in many studies. The other three tests use slightly different approaches to produce a confidence set of trees, which contains all trees from a pre-specified set that are not significantly different from the optimal tree. The SH- and AU-test both use a non-parametric bootstrapping approach that re-samples data from the observed sequences, which means the method is model independent but can suffer when there is limited sequence data. The SOWH-test is based on a parametric bootstrapping approach that samples data according to the parameters estimated in the model, resulting in a method that is model dependent, but still relatively effective when only limited sequence data are available. The Kishino–Hasegawa [60] test is purposely excluded from this list; the conditions required for its correct application to phylogenetic data require an a priori specification of an optimal tree. In the majority of studies, this condition is rarely met and the test is probably best avoided [4,52].

5. Heterogeneity in the evolutionary process

There is accumulating evidence that the evolutionary process varies between sites in biological sequences. Even in non-functional regions of the genome, there appears to be variability in the mutational process [5,61]. This variation is even more pronounced in active genomic segments. In protein coding sequences, changes that impede function are unlikely to be accepted by selection (e.g., mutation in the active site), whilst those altering less vital areas are under fewer selective constraints (e.g., mutation in non-func-

tional loop regions). Context-dependent mutation and changes in selection resulting from variable environmental pressures further complicate this variation. For example, methylation of CG dinucleotides in vertebrates leads to rapid deamination to CA and TG; a non-reversible, context-dependent mutational event [62]. In Fig. 1 we summarize two complementary approaches for describing amongst-site heterogeneity in sequence evolution: recoding the sequence data (wedges) and describing aspects of the heterogeneity in the evolutionary model (concentric circles). These modeling approaches accounting for spatial heterogeneity are useful for investigating the variable selective constraints that functional units exert on the genome, but neglect changes in the evolutionary process over time, such as between lineages in a tree. A general introduction to the approaches used to investigate temporal heterogeneity is beyond the scope of this review, although we briefly touch upon some aspects of this rapidly expanding research area [63–66] when discussing methods used to detect variation in selection acting on protein coding sequences.

5.1. Recoding sequence data

The presence of biological structures in biological sequences often makes it practical to use recoded versions of the data. The three wedges of the circle in Fig. 1 describe three commonly used ways of labeling coding sequence data: the raw DNA (applicable to all genomic sequence), the coding triplet, and the translated amino acid sequences. The chosen data labeling and the movement from one level of labeling to another are part of modeling procedure, and decisions made when recoding information in the data will affect the properties of the model. The simplest, homogeneous models for these different labels are contained in the inner most ring. It is natural to describe evolution at the nucleotide level because it represents the true source of genetic variability, mutation. The majority of parameters in DNA models are mechanistic, including biologically relevant factors, such as bias towards transition mutations and variability in nucleotide frequency [4,6]. Recoding nucleotide sequences to the 64 codons allows the dependencies resulting from the degeneracy of the genetic code to be explicitly incorporated into the evolutionary model. Codon based models [13,67,68] are usually mechanistic, and in addition to the parameters contained in DNA models, describe the tendency of mutations maintaining the encoded amino acid (synonymous) to be accepted by selection more frequently than those that change the amino acid (non-synonymous). A later portion of this article examines how advanced codon models detect adaptive evolution.

When the purpose of a study is to estimate an evolutionary relationship or to investigate protein structure and function, it is often useful to translate the coding sequence to amino acids. This allows fast and effective computation, but discards some potentially useful evolutionary information, as some amino acids are encoded by multiple codons and mutational changes at the DNA level do not always result in amino acid changes. The majority of models describing protein evolution are based on empirical parameters, describing the rates of amino acid replacement as constant between sites and between proteins; an approach typified by the first description of protein evolution by Dayhoff and co-workers [37,69]. Recently, more advanced estimation procedures have progressively led to more accurate descriptions of the evolutionary process [32]. There are also mechanistic additions to protein models that adapt the relative frequencies that the amino acids occur to reflect better the proteins under consideration [35,36].

5.2. Rate variation

All of the models discussed so far assume the same, homogeneous evolutionary process acting at all sites in a sequence. In the post-genomic era, the sequences used for phylogenetic and comparative genomic analyses are increasingly long and therefore more likely to contain

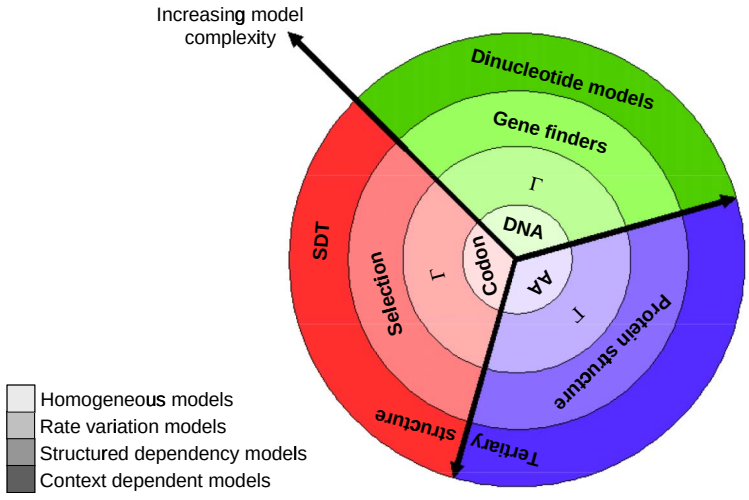

Fig. 1. A diagram demonstrating the relationship between evolutionary models that describe dependencies between sites.

regions evolving under different mutational processes and selective constraints. Proceeding outward from the center, the rings of Fig. 1 represent increasingly complex and realistic models of evolution. The many complexities of molecular evolution are primarily manifested as a notable difference in the rate that sites change. For example, the active units in a genome evolve under different and complex selective constraints, but techniques for identifying them from non-functional areas usually rely on the observation that they evolve more slowly [70]. Similarly, the degeneracy of the genetic code results in substantially different rates of evolution at the three codon positions.

The second concentric ring contains models that describe non-specific variation in evolutionary rate amongst sites. The standard approach to characterizing this variation is to describe each site's rate as a random draw from a statistical distribution, whilst maintaining all other aspects of the evolutionary process. In other words, each site has a defined probability of evolving at a given rate independent of its neighbors. It is also assumed that this rate is constant throughout evolution; a fast changing site has an elevated evolutionary rate throughout the phylogenetic tree. Uzzel and Corbin [71] first suggested that in DNA the rate variation observed in coding sequences could be described using a Γ-distribution. Yang implemented this description as a probabilistic model, using a continuous Γ-distribution containing a single, biologically interpretable, shape parameter that can accommodate varying degrees of rate heterogeneity [72]. The value of the shape parameter is inversely related to the degree of rate variation it describes: for values below 1.0 it describes the extensive rate variation characteristic of functional regions (e.g., protein coding sequences), with numerous sites evolving at a low rate and a few faster evolving sites; values greater than 1.0 convey limited rate variation, which occurs often in non-functional regions (e.g., pseudogenes). Later, Yang proposed breaking the distribution into a pre-specified number of categories to make the model computationally more efficient [73]. This approach has been successfully employed in many studies and under all levels of data. The inclusion of Γ-distributed rates has been demonstrated to affect, and usually improve, the estimation of other evolutionary parameters, including the tree [74].

5.3. Structured models

The next ring in Fig. 1 describes models where the whole process of evolution varies amongst sites, usually as the result of underlying genomic structures. This covers two categories of model: mixture models, which describe the evolution at sites independently of their neighbors; and hidden Markov models (HMMs), which allow a limited linkage between sites and their immediate neighbors [3,75]. Readers should note that these approaches, and the context-dependent models described below, represent some of the most innovative and advanced methodology available for phylogenetics and comparative genomics. Their general utility is, in some cases, unproven and their widespread application could be hindered by theoretical and computational problems. The increase in complexity in these models (number of parameters) may interfere with effective inference, whether using likelihood or Bayes. At best, this results in the loss in statistical power and/or computational problems. In more difficult circumstances, the parameters cannot be distinguished from one another, often referred to as a lack of identifiability, which can lead to inconsistent inferential methodology, where additional data does not improve the accuracy of parameter estimates [76]. With extensive research and continued application these obstacles will be overcome and the descendents of some of the models described below may become standard tools for future evolutionary biologists.

5.3.1. Mixture models

Two types of mixture model are commonly used in phylogenetics: random-effects models and fixed-effects models. Random-effects models assume no knowledge of the structures present in the sequence under consideration and each site has a defined probability of belonging to different types of evolutionary process. The models of Γ-distributed rate variation described above are special cases of these types of model, where each site has a probability of evolving at each given rate. More general random-effects mixture models allow better characterization of the causes of evolutionary heterogeneity and the identification of the sites affected. They have been employed with amino acid data labeling to investigate regional variation in the properties of protein evolution [12,77] and with DNA labeling to examine variation in the mutational process [78]. To date, their most successful application has been through the use of advanced codon models to detect adaptive evolution [13].

Fixed-effects models assume a full knowledge of the structures present in the data and use it to separate the data into biologically meaningful partitions that evolve to different evolutionary processes. These partitions can include different genes in a concatenated sequence, the three codon positions, or the exons and introns of a gene. Models can then be constructed to investigate similarities and differences between the evolutionary processes acting in these partitions. For example, an exon may be defined as having a similar mutational bias towards transitions as introns, but a different frequency of nucleotides and rate. This approach has been used to examine differences between genes when constructing 'genomic' phylogenies [18] and when deciding how best to combine individual genes for phylogenetic purposes [79]. Yang [80] described a general version of the method for nucleotide models, which is implemented in the PAML package [81].

5.3.2. HMMs

Mixture models enable us to describe variation in the evolutionary process, but still assume that sites evolve independently. HMMs are generalized mixture models that allow correlations between nearby sites. The hidden states

described in the model refer to different underlying evolutionary processes. Fig. 2A contains a schematic of a simple phylogenetic HMM containing two hidden states: GC rich (the square) and GC poor (the circle). The bubble-plot within each hidden state describes the type of evolution occurring, with bigger bubbles representing more frequent substitutions (see [4] for more details). The relative sizes of the arrows linking the hidden states describe how frequently they interchange: the thick, curved arrows leading from the circle (or square) back to itself indicate that when in a hidden state the process is likely to stay there; and the thin arrows connecting the circle and square indicate that the two hidden states rarely interchange. The infrequent change between hidden states encourages the model to describe evolution as blocks of GC-rich and GC-poor evolving regions. Fig. 2B shows how this model would look when expanded to cover 8 sites in a sequence, and the degree of shading of the hidden states is to convey that the model has no a priori knowledge of what state each site in the sequence belongs to. Fig. 2C demonstrates this model when applied to sequence data. There are established algorithms to find the flow through the HMM that best describes the observed data [3], demonstrated by the thickness of the arrows and the shading of the hidden states in the figure. For example, the first five states have strong red circles with thick arrows linking them, and weak blue squares with thin arrows leading from them, indicating great confidence that they belong to the GC-poor hidden-state. The seventh and eighth sites have light circles and strongly shaded squares, depicting confidence that these regions belong to a GC-rich region. The model is unsure of the hidden state at the sixth site because it falls at the junction of a GC-rich and GC-poor region. Fig. 2 also effectively demonstrates the advantage of HMMs over simple mixture models. If the model did not contain the underlying structure of the HMM, the third site in the sequence would be labeled as a GC-rich region, but the influence of neighboring sites correctly identifies it as a GC-poor region.

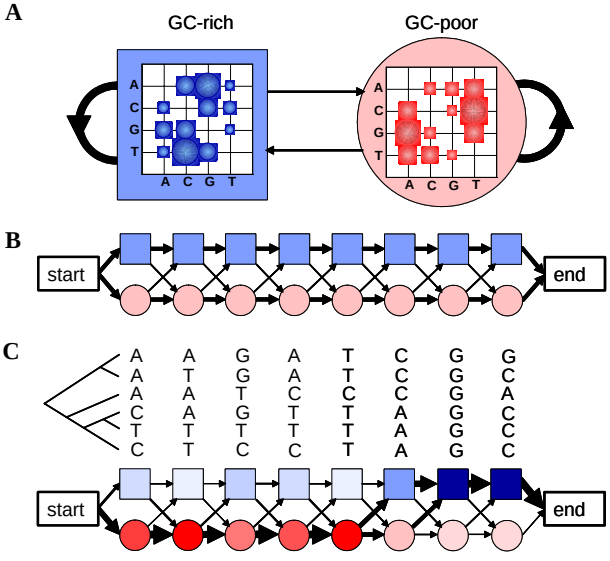

Fig. 2. A schematic description of a hidden Markov model describing variation in GC-content. (A) General description of the HMM: the square and circle represent the two hidden states of rich and poor GC-content; the bubble-plot within each hidden state describes the type of evolution occurring; and the thickness of the arrows describe the frequency of transition between the hidden states. (B) The HMM primed for analysis on an 8 nucleotide sequence alignment; the relatively equal shading of the hidden states indicate that the model has no a priori knowledge of the sequence, whilst the tendency of thicker arrows to keep the model in a single state reflect the model's inherent inclination to remain in a hidden state. (C) The HMM applied to data; the relative shading of the hidden states measure the model's confidence for sites in the alignment belonging to them. For more details see text.

HMMs were first introduced to phylogenetic inference by Felsenstein and Churchill [82], who used the approach to describe local similarities in evolutionary rate. They have since been applied at the amino acid level to estimate protein structure [11,83]. These models exploit the dependency of amino acid evolution on its solvent accessibility and its secondary structure environment. At the DNA level, some gene finders have improved their accuracy by exploiting variation in the evolutionary process of exons, introns, splice sites, and intergenic regions [84,85]. Phylogenetic HMMs are becoming increasingly popular and are starting to find mainstream use [86], with their results now routinely used to examine the degree of conservation shared between genomic sequences.

5.4. Context-dependent models

HMMs still assume a simple evolutionary process where the mutational and selective forces at a site are dependent only on the overall type of evolutionary process at neighboring sites. The outer-most ring in Fig. 1 describes models of context-dependent evolution, where the specific characters at neighboring sites affect a sites evolution. The earliest of these describe the selective forces acting on nucleotides in the stem regions of RNA molecules. Observing a mutation that changes a single base in these regions is rare because they are strongly selected to maintain complementary base pairing. Regular changes, however do occur at a low rate when both complementary bases mutate in rapid succession, negating the selective disadvantage. This problem has been addressed by recoding the data in stem regions as pairs of nucleotides (e.g., $AA, AC, \ldots, TG, TT$) and allowing these rapid mutations to occur as simultaneous events in the model, demonstrating the effectiveness of data labeling to address difficult mathematical problems. There are, however, problems that recoding the data cannot fix, for example, the unidirectional hyper-mutation of CG in vertebrates. In this case, the mutation at any one site is only dependent on the state of its immediate neighbors, but these neighbors are affected by their neighbors, and so on. This cascading effect makes computation very difficult because the evolution of sequences has to be described as a whole. Recently, there have been several proposed approaches to tackling this problem. Many of these studies have concentrated on the hyper-mutation of CG, usually

through complex approximations that allow efficient computation [87–89]. The SDT model of Whelan and Goldman [20] used codon labeling to investigate whether mutations frequently cause multiple nucleotides to change. Robinson et al. [19] produced a model that included the dependency between codons resulting from the secondary and tertiary structure of proteins. This embraces the codon as well as the amino acid level and is placed between both data levels in Fig. 1. The majority of current models describing context dependency are experimental and, at present, there are no studies comparing the performance of competing methodology. They represent an exciting and open avenue of research, where novel biological discoveries are waiting to be made. With the exception of RNA stem evolution, it may be some time before they are widely used by evolutionary and genomic scientists.

6. Adaptive evolution in proteins

The selective forces acting upon a protein are highly informative about its biological function and evolutionary history [90]. For example, the interactions of proteins through their regulatory and metabolic networks are also reflected in the selection acting upon them. Recently, it was demonstrated that the more interactions a protein has with other molecules, the slower it evolves, and that proteins operating in complexes (e.g., involved in translation or DNA repair) are, on average, more restricted than those with simple housekeeping functions [91]. In viruses, the sites on envelope proteins interact with host molecules and are targets for the immune system, leading a host-viral ''arms race,'' and the amino acids at the interacting sites evolve under continuous positive selection. These sites are usually in the envelope of viruses and not involved directly in antigen functions, such as docking or binding to host cells, and their identification through evolutionary analysis can be useful for the design of vaccines.

6.1. Homogenous selection

Models describing evolution at the codon level allow estimation of the average selective forces acting on a sequence alignment. The ratio of rates between non-synonymous and synonymous substitutions, referred to as ω in the parlance of evolutionary modeling, is used as a direct measure of these forces. It can be used to detect when coding DNA is evolving neutrally, under negative (purifying) selection, and under positive (adaptive) selection. The design of codon models means that ω accounts for the structure of the genetic code, greatly aiding direct, biological interpretation of its value. When there are few selective pressures acting, sequences are said to be evolving neutrally and the relative rates of synonymous mutation and non-synonymous are roughly equal (ω is approximately 1). When a protein has an important function its sequence is highly conserved through evolution and ω takes a value substantially less than 1. Conversely, when proteins adapt quickly to their environment, non-synonymous mutations are strongly selected for and ω will take a value greater than 1. The pioneering studies examining the effect of natural selection in proteins estimated an average ω (or equivalent value) across all the sites in an alignment and usually drew similar conclusions: adaptive evolution is rare and most proteins are under strong purifying selection [92]. The lack of positive selection suggested by these studies is most probably an underestimate of the true amount. The contents of a genome have been evolving for millions of years and are highly adapted to the functions they perform. Consequently, purifying selection will have been acting on the majority of sites in a protein to maintain this function and the average value of ω would be expected to be low. Positive selection would normally be expected to affect only a few key residues in a protein, to successfully find its footprint during molecular evolution requires a more sophisticated approach.

6.2. Variation in selection

Mixture models are an effective way to describe the variable selective forces acting on protein sequences. Important selective constraints occur through a protein's structure and its interaction with its environment, not through immediate proximity in the linear coding sequence. The use of LRTs to compare pairs of models using different statistical distributions to describe the variation in ω has proven an effective way of identifying adaptive evolution [13]. For these tests, the null model describes the evolution of a protein as a distribution containing only neutral and purifying selection ($\omega \leqslant 1$), and the alternative model describes a similar distribution that also allows positive selection (ω can take all values). A popular choice of models for forming these hypotheses are M7 and M8 (see Yang et al. [13]). M7 (the null model) describes variation in ω between 0 and 1 with a β-distribution, which can take a variety of shapes that describe a wide range of potential selective scenarios and requires only two simple parameters to be estimated from the data. M8 (the alternate model) contains the β-distribution of M7, but also includes a single variable category of ω to describe positive selection. When statistical tests show that M8 explains the evolution of the protein significantly better than M7 and the additional category of ω is greater than 1, positive selection is inferred. Many new incidences of adaptive evolution have been found using this approach and extensions of these methods allow the detection of the specific sites in a protein that are undergoing positive selection [13,93]. When these adapting sites are mapped on to the three-dimensional structure of proteins, they have identified regions known to have chemical significance through other experimental study. For example, this approach has identified sites undergoing positive selection in HIV that laboratory studies have previously proved to be targeted by the immune system. Readers should also be aware that an updated version of this test for positive selection is now

available, a comparison between M8a and M8b, which has favorable statistical properties compared to original M7/M8 test (see Swanson et al. [94] and the PAML documentation [81] for more details).

The mixture models described above assume the selective pressures acting on a protein remain constant through evolutionary time. This is a suitable assumption when sites are under consistent selective pressure to change, but inappropriate when molecular adaptation occurs during relatively short periods of evolutionary time. To detect episodic adaptive evolution, models have been updated to allow ω to vary during evolution. These function either by allowing ω to vary in all of the branches in a phylogenetic tree (across-branch models) [95] or, more recently, by allowing ω to vary both across sites and across a prespecified group of branches in a phylogeny (branch-site models) [96]. These models are still relatively limited in their ability to describe evolution, but there has been rapid development and they have proved useful in real biological applications, including animal physiology. During the evolution of primates, the lysozyme enzyme has been associated with foregut fermentation in colobine monkeys. The application of across-branch modeling has been successfully employed to demonstrate that positive selection occurred in this lineage and provided evidence that hominids also display this type of molecular adaptation [95].

Modern likelihood-based methods have proved successful in detecting molecular adaptation by providing sensitive analyses combined with rigorous statistical testing procedures. There is, however, ongoing debate regarding the performance of these approaches when there is serious uncertainty in the tree topology, errors in alignment, or limited amount of information in data. Wong et al. [97] used simulation approaches to demonstrate that across-site methods are stable in all but the most extreme conditions. Further developments in the methodology used in these amongst-sites models should also reduce error [93,94]. Branch-site models have been shown to struggle to discriminate between positive selection and neutral evolution under certain simulation conditions [98], although this has not stopped them being successfully employed on real data [99]. Researchers developing new methods work hard to ensure their accuracy, but reasonable care should always be taken when preparing data and performing analyses.

7. Summary

Molecular evolutionary studies offer an effective method for using genomic information to investigate many biomedical phenomena. This review highlights some of the benefits of using likelihood-based methods to investigate molecular evolution: the clear assumptions made when devising evolutionary models; the statistical properties that allow accurate parameter estimation and hypothesis testing; and its efficient use of information in sequence data. Factors that have led to likelihood becoming the most widely used inference procedure in molecular phylogenetics in recent years. Progress in phylogenetic methods, coupled with improvements in computer hardware, has allowed long-held and limiting assumptions about molecular evolution to be relaxed and a new generation of evolutionary models to be developed. This is typified by innovations for describing the heterogeneity of the evolutionary process within sequences, where highly advanced methodology has become commonplace in examining the selective pressures in proteins. The development of sophisticated modeling technology requires careful and diligent study to ensure that its products are biologically interpretable and over-parameterization is avoided. The fruits of this novel and exciting research are beginning to cast new light on how biological molecules function and interact with their environment. The future of statistical modeling in molecular evolution is bright; the completion of further genome projects will provide ample data, allowing new and exciting studies, which in turn will feed forward to the development of more realistic descriptions of evolution.

References

[1] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature 2001;409:860–921.
[2] Watson JD, Crick FHC. A structure for deoxyribose nucleic acid. Nature 1953;171:737–8.
[3] Durbin R, Eddy S, Krogh A, Mitchison G. Biological sequence analysis. Cambridge University Press; 1998.
[4] Whelan S, Lió P, Goldman N. Molecular phylogenetics: state-of-the-art methods for looking into the past. Trends Genet 2001;17:261–72.
[5] Mouse Genome Sequencing Consortium. Initial sequencing of the mouse genome. Nature 2002;420:520–62.
[6] Felsenstein J. Inferring phylogenies. Massachusetts: Sinauer Associates; 2003.
[7] Rat Genome Sequencing Consortium. Genome sequence of the brown Norway rat yields insights into mammalian evolution. Nature 2004;428:493–521.
[8] International Chicken Genome Sequencing Consortium. Sequence and comparative analysis of the chicken provide unique perspectives on vertebrate evolution. Nature 2004;432:695–716.
[9] The ENCODE Project Consortium. The ENCODE (Encyclopedia of DNA Elements) project. Science 2004;306:636–40.
[10] Qian B, Goldstein RA. Detecting distant homologs using phylogenetic tree-based HMMs. Proteins 2003;52:446–53.
[11] Goldman N, Thorne JL, Jones DT. Using evolutionary trees in protein secondary structure prediction and other comparative sequence analysis. J Mol Biol 1996;263:196–208.
[12] Dimmic MW, Mindell DP, Goldstein RA. Modeling evolution at the protein level using an adjustable amino acid fitness model. Pac Symp Biocomput 2000;18–29.
[13] Yang Z, Nielsen R, Goldman N, Pedersen AMK. Codon-substitution models for the heterogeneous selection pressure at amino acid sites. Genetics 2000;155:431–49.
[14] Suzuki Y, Gojobori T. A method for detecting positive selection at single amino acid sites. Mol Biol Evol 1999;16:1315–28.
[15] Zhu G, Golding GB, Dean AM. The selective cause of an ancient adaption. Science 2005;307:1279–82.
[16] Yang Z, Goldman N, Friday A. Comparison of models for nucleotide substitution used in maximum likelihood phylogenetic estimation. Mol Biol Evol 1994;1:725–36.
[17] Buckley TR. Model misspecification and probabilistic tests of topology: evidence from empirical data sets. Syst Biol 2002;51:509–23.

[18] Rokas A, Williams B, King N, Carroll SB. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 2003;425:798–804.
[19] Robinson DM, Jones DT, Kishino H, Goldman N, Thorne JL. Protein evolution with dependence among codons due to tertiary structure. Mol Biol Evol 2003;20:1692–704.
[20] Whelan S, Goldman N. Estimating the frequency of events that cause multiple-nucleotide changes. Genetics 2004;167:2027–43.
[21] Fisher RA. Theory of statistical estimation. Proc Camb Phil Soc 1925;22:700–25.
[22] Edwards AWF. Likelihood. Cambridge: Cambridge University Press; 1972.
[23] Kimura M. A simple method for estimating evolutionary rates of base substitution through comparative studies of nucleotide sequences. J Mol Evol 1980;6:111–20.
[24] Lió P, Goldman N. Models of molecular evolution and phylogeny. Gen Res 1998;8:1233–44.
[25] Posada D, Crandall KA. The effect of recombination on the accuracy of phylogeny estimation. J Mol Evol 2002;54:396–402.
[26] Doolittle WF. Phylogenetic classification and the universal tree. Science 1999;284:2124–9.
[27] Chang JT. Full reconstruction of Markov models on evolutionary trees: identifiability and consistency. Math Biosci 1996;137:51–73.
[28] Press HP, Teukolsky SA, Vetterling WT, Flannery BP. Numerical recipes in C. Cambridge University Press; 1992.
[29] Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 1987;4:406–25.
[30] Sorhannus U. The effect of positive selection on a sexual reproduction gene in *Thalassiosira weissflogii* (Bacillariophyta): results obtained from maximum likelihood and parsimony-based methods. Mol Biol Evol 2003;20:1326–8.
[31] Suzuki Y, Nei M. False-positive selection identified by ML-based methods: examples from the *Sig1* gene of the diatom *Thalassiosira weissflogii* and the *tax* gene of a human T-cell lymphotropic virus. Mol Biol Evol 2004;21:914–21.
[32] Whelan S, Goldman N. A general empirical model of protein evolution derived from multiple protein families using a maximum likelihood approach. Mol Biol Evol 2001;18:691–9.
[33] Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 1981;17:368–76.
[34] Hasegawa M, Kishino H, Yano T. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 1985;22:160–74.
[35] Cao Y, Adachi J, Janke A, Pääbo S, Hasegawa M. Phylogenetic relationships among eutherian orders estimated from inferred sequences of mitochondrial proteins: instability of a tree based on a single gene. J Mol Evol 1994;39:519–27.
[36] Goldman N, Whelan S. A novel use of equilibrium frequencies in models of sequence evolution. Mol Biol Evol 2002;19:1821–31.
[37] Dayhoff MO, Schwartz RM, Orcutt BC. A model of evolutionary change in proteins. In: Dayhoff MO, editor. Atlas of protein sequence and structure; 1978, p. 345–352.
[38] Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 1992;89:10915–9.
[39] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–10.
[40] Thompson JD, Higgins DG, Gibson TJ. Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994;22:4673–80.
[41] Whelan S, Goldman N. Statistical tests of gamma-distributed rate heterogeneity in models of sequence evolution in phylogenetics. Mol Biol Evol 1999;16:1292–9.
[42] Goldman N, Whelan S. Distributions of statistics used for the comparison of models of sequence evolution in phylogenetics. Mol Biol Evol 2000;17:975–8.
[43] Huelsenbeck J, Rannala B. Phylogenetic methods come of age: testing hypotheses in an evolutionary context. Science 1997;276:227–32.
[44] Efron B, Tibshirani RJ. An introduction to the bootstrap. Chapman and Hall; 1993.
[45] Goldman N. Statistical tests of models of DNA substitution. J Mol Evol 1993;37:650–61.
[46] Akaike H. A new look at the statistical model identifications. IEEE Trans Automat Contr 1974;AC-19:716–23.
[47] Sugiura N. Further analysis of the data by Akaike's information criterion and the finite corrections. Commun Stat A—Theory Methods 1978;7:13–26.
[48] Schwarz G. Estimating the dimension of a model. Ann Statist 1978;6:461–4.
[49] Posada D, Crandall KA. MODELTEST: testing the model of DNA substitution. Bioinformatics 1998;14:817–8.
[50] Abascal F, Zardoya R, Posada D. ProtTest: selection of best-fit models of protein evolution. Bioinformatics 2005;21:2104–5.
[51] Swofford DL, Olsen GJ, Waddell PJ, Hillis M. Phylogenetic inference. In: Hillis DM, Moritz C, Mable BK, editors. Molecular systematics. Massachusetts: Sinauer Associates; 1996. p. 407–514.
[52] Goldman N, Anderson JP, Rodrigo AG. Likelihood-based test of topologies in phylogenetics. Syst Biol 2000;49:652–70.
[53] Felsenstein J. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 1985;39:783–91.
[54] Shimodaira H, Hasegawa M. Multiple comparisons of log-likelihoods with applications to phylogenetics inference. Mol Biol Evol 1999;16:1114–6.
[55] Shimodaira H, Hasegawa M. CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinformatics 2001;17:1246–7.
[56] Shimodaira H. An approximately unbiased test of phylogenetic tree selection. Syst Biol 2002;51:492–508.
[57] Hillis D, Bull J. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Syst Biol 1993;42:182–92.
[58] Felsenstein J, Kishino H. Is there something wrong with the bootstrap on phylogenies? A reply to Hillis and Bull. Syst Biol 1993;42:193–200.
[59] Efron B, Halloran E, Holmes S. Bootstrap confidence levels for phylogenetic trees. Proc Natl Acad Sci USA 1996;93:13429–34.
[60] Hasegawa M, Kishino H. Confidence limits on the maximum-likelihood estimate of the homonoid tree from mitochondrial-DNA sequences. Evolution 1989;43:672–7.
[61] Hardison R, Roskin KM, Yang S, Diekhans M, Kent WJ, Weber R, et al. Covariation in frequencies of substitution, deletion, transposition and recombination during eutherian evolution. Genome Res 2003;13:13–26.
[62] Gardiner-Garden M, Frommer M. CpG islands in vertebrate genomes. J Mol Biol 1987;196:261–82.
[63] Steel MA, Székely LA, Hendy MD. Reconstructing trees when sequence sites evolve at variable rates. J Comp Biol 1994;1:153–63.
[64] Thorne JL, Kishino H, Painter IS. Estimating the rate of evolution of the rate of molecular evolution. Mol Bio Evol 1998;15:1647–57.
[65] Lopez P, Casane D, Philippe H. Heterotachy, an important process of protein evolution. Mol Biol Evol 2002;19:1–7.
[66] Spencer M, Susko E, Roger AJ. Likelihood, parsimony, and heterogeneous evolution. Mol Biol Evol 2005;22:1161–4.
[67] Goldman N, Yang Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 1994;11:725–36.
[68] 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;11:715–24.
[69] Kosiol C, Goldman N. Different versions of the Dayhoff rate matrix. Mol Bio Evol 2004;22:193–9.
[70] Margulies EH, Blanchette M, NISC Comparative Sequencing Program, Haussler D, Green ED. Identification and characterization of multi-species conserved sequences. Genome Res. 2003;13:2507–18.

[71] Uzzel T, Corbin KW. Fitting discrete probability distributions to evolutionary events. Science 1971;172:1089–96.
[72] Yang Z. Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol Biol Evol 1993;19:1396–401.
[73] Yang Z. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol 1994;39:306–14.
[74] Yang Z. Among-site rate variation and its impact on phylogenetic analyses. Trends Ecol Evol 1996;11:367–72.
[75] Krogh A, Brown M, Mian IS, Sjölander K, Haussler D. Hidden Markov models in computational biology: applications to protein modelling. J Mol Biol 1994;235:1501–31.
[76] Steel MA. Should phylogenetic models be trying to 'fit an elephant'? TIG 2005;21:307.
[77] Lartillot N, Philippe H. A Bayesian mixture model for across-site heterogeneities in the amino acid replacement process. Mol Biol Evol 2004;21:1095–109.
[78] Pagel M, Meade A. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequences or character-state data. Syst Biol 2004;53:571–81.
[79] Pupko T, Huchon D, Cao Y, Okada N, Hasegawa M. Combining multiple datasets in a likelihood analysis: which models are best. Mol Biol Evol 2002;19:2294–307.
[80] Yang Z. Maximum likelihood models for combined analyses of multiple sequence data. J Mol Evol 1996;42:587–96.
[81] Yang Z. PAML: a program package for phylogenetic analysis by maximum likelihood. CABIOS 1997;15:555–6.
[82] Felsenstein J, Churchill HA. A hidden Markov model approach to variation among sites in rate of evolution. Mol Biol Evol 1996;13:93–104.
[83] Thorne JL. Models of protein sequence evolution and their applications. Curr Opin Genet Dev 2000;10:602–5.
[84] Meyer IM, Durbin R. Comparative ab initio prediction of gene structures using pair HMMs. Bioinformatics 2002;18:1309–18.
[85] Flicek P, Keibler E, Hu P, Korf I, Brent MR. Leveraging the mouse genome for gene prediction in human: from whole-genome shotgun reads to a global synteny map. Genome Res 2003;13:46–54.
[86] Siepel A, Haussler D. Phylogenetic hidden Markov models. In: Nielsen R, editor. Statistical methods in molecular evolution. New York: Springer; 2005. p. 325–51.
[87] Pedersen AMK, Jensen JL. A dependent rates model and MCMC based methodology for the maximum likelihood analysis of sequences with overlapping reading frames. Mol Biol Evol 2001;18:763–76.
[88] Siepel A, Haussler D. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. Mol Biol Evol 2004;21:468–88.
[89] Lunter G, Hein J. A nucleotide substitution model with nearest-neighbour interactions. Bioinformatics 2004;20:i216–23.
[90] Yang ZB. Statistical methods for detecting molecular adaptation. Trends Ecol Evol 2000;51:423–32.
[91] Aris-Brosou S. Determinants of adaptive evolution at the molecular level: the extended complexity hypothesis. Mol Biol Evol 2005;22:200–9.
[92] Endo T, Ikeo K, Gojobori T. Large-scale search for genes on which positive selection may operate. Mol Biol Evol 1996;13:685–90.
[93] Massingham T, Goldman N. Detecting amino acid sites under positive selection and purifying selection. Genetics 2005;169:1753–62.
[94] Swanson WJ, Nielsen R, Yang Q. Pervasive adaptive evolution in mammalian fertilization proteins. Mol Biol Evol 2003;20:18–20.
[95] Yang Z. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol 1998;15:568–73.
[96] Yang Z, Nielsen R. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol 2002;19:908–17.
[97] Wong WSW, Yang Z, Goldman N, Nielsen R. Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. Genetics 2004;168:1041–51.
[98] Zhang J. Frequent false detection of positive selection by the likelihood method with branch-site models. Mol Biol Evol 2004;21:1332–9.
[99] Bielawski JP, Yang Z. A maximum likelihood method for detecting functional divergence at individual codon sites, with application to gene family evolution. J Mol Evol 2004;59:121–32.