MBE Advance Access published October 14, 2008

For Mol. Biol. Evol.
First submission on 28 July 2008
Revised version on 15 September 2008

Investigating protein-coding sequence evolution with probabilistic codon substitution models

Maria Anisimova[1,2,*] and Carolin Kosiol[3]

[1] Institute of Computational Science, Swiss Federal Institute of Technology (ETHZ), Zurich, Switzerland, [2] Swiss Institute of Bioinformatics, Lausanne, Switzerland, and [3] Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY, USA

*Corresponding author:

Dr Maria Anisimova

Institute of Computational Science

ETH Zurich

6 Universitatstrasse

8092 Zurich, Switzerland

Phone: +41 44 632 6076

Fax: +41 44 632 1374

Email: maria.anisimova@inf.ethz.ch

Key words: Markov model, maximum likelihood, Bayesian approach, codon evolution, positive selection

Abstract

This review is motivated by the true explosion in the number of recent studies both developing and ameliorating probabilistic models of codon evolution. Traditionally parametric, the first codon models focused on estimating the effects of selective pressure on the protein via an explicit parameter in the maximum likelihood framework. Likelihood ratio tests of nested codon models armed the biologists with powerful tools, which provided unambiguous evidence for positive selection in real data. This, in turn, triggered a new wave of methodological developments. The new generation of models view the codon evolution process in a more sophisticated way, relaxing several mathematical assumptions. These models make a greater use of physico-chemical amino acid properties, genetic code machinery and the large amounts of data from the public domain. The overview of the most recent advances on modeling codon evolution is presented here and a wide range of their applications to real data is discussed. On the downside, availability of a large variety of models, each accounting for various biological factors, increases the margin for misinterpretation; the biological meaning of certain parameters may vary among models, and model selection procedures also deserve greater attention. Solid understanding of the modeling assumptions and their applicability is essential for successful statistical data analysis.

Introduction

Protein-coding genes are the DNA sequences necessary for the production of functional proteins. Such sequences consist of nucleotide triplets called codons. During the protein production phase codons are translated into amino acids (AA) according to the organism's genetic code. While protein-coding genes form only 1.5% of the human genome, they are the major part of viral genomes, which are so compact that many genes use overlapping reading frames. The analysis of coding sequences can be performed on three different levels: using DNA, AA or codon sequences. While a dominating proportion of methods are DNA-based, these are often used on coding sequences. Unfortunately, DNA-based methods are not adapted for codon data. Treating all codon positions equally is likely to lead to misleading conclusions (Shapiro, Rambaut, and Drummond 2006; Bofkin and Goldman 2007). Different positions typically evolve with highly heterogeneous patterns, and so should be analyzed as different site-partitions.

On another hand, synonymous substitutions may be saturated for distant species (Maynard Smith and Smith 1996). Thus, amino acid models also often take preference even when synonymous substitutions may reveal important details. However, analyses of fast-evolving mammalian mitochondria and deep-rooted yeast data demonstrate that synonymous substitutions are very informative even for high divergences and substantially improve phylogenetic inference (Seo and Kishino 2008). Such a conclusion is based on a comparison of an AA model with a codon model, through a transformation of a 20×20 AA-matrix to a 61×61 codon-matrix. Interestingly, even though the simplest codon models do not account for physicochemical properties, they were found to fit data significantly better than AA models (Seo and Kishino 2008). The relationship between AA and codon models has been further explored via aggregated Markov models (Kosiol 2006). Such models can explain some of the observed non-Markovian behavior of AA sequences (e.g., Benner, Cohen, and Gonnet 1994) and suggests that protein sequence evolution should be modeled at the codon level rather than at the level of AA substitutions.

The first codon models became especially popular in positive selection studies of protein-coding genes, e.g., where a phenotypic change may be attributed to functional changes of a protein, caused by advantageous substitutions. One advantage of studying protein-coding sequences is the ability to distinguish between the nonsynonymous (AA-replacing) codon changes and the synonymous (AA-conserving) changes. Based on this distinction, the selective pressure on the protein-coding level can be measured through the comparison of the nonsynonymous and synonymous substitution rates, d_N and d_S respectively (Kimura 1977; Jukes and King 1979). In the traditional framework, if recurrent AA changes are advantageous, the nonsynonymous substitution rate is higher than the synonymous rate, and their ratio $\omega = d_N/d_S > 1$. In contrast, purifying selection acts to preserve the AA sequence, so that the nonsynonymous substitution rate is lower than the synonymous rate, causing $\omega < 1$. Neutrally evolving sequences exhibit similar nonsynonymous and synonymous rates, with $\omega \approx 1$.

In this article we will describe the advanced probabilistic codon models applied within maximum likelihood (ML) and Bayesian frameworks, whose development was driven by their greater use in genome-scale analysis and the increased availability of computational resources. While the first codon models primarily focused on detecting positive selection by the comparison of nonsynonymous and synonymous substitution rates (as discussed in Anisimova and Liberles 2007), most recent models allow to explore finer aspects of coding sequence evolution, including physicochemical properties of changes, synonymous rates variation, selective pressures on codon usage, and rates of instantaneous double and triple nucleotide mutations.

Methodological framework

Markovian codon models

Similar to DNA and AA, models of codon substitution are typically described by a Markov process, where the probability of a change from one state to another depends only on the current state and not on any past states. The process is fully determined by the matrix $Q = \{q_{ij}\}$ specifying instantaneous rates of

change between 61 sense codons. Mutations to/from stop codons are not allowed, since such events usually are not tolerated by a functional protein. The diagonal elements of Q are defined by a mathematical requirement that the rows sum up to zero, and so $q_{ii} = -\sum_{i \neq j} q_{ij}$ (Cox and Miller 1977).

For multiple sequence alignments, the substitution process runs in continuous time over a tree representing phylogenetic relations between the sequences. Transition probabilities $p_{ij}(t)$ from codon i to codon j over time $t > 0$ are given by the transition probability matrix $P(t) = \{p_{ij}(t)\} = e^{Qt}$, which is found as a solution of the differential equation $dP(t)/dt = P(t)Q$ with $P(0)$ being an identity matrix (Cox and Miller 1977). The instantaneous rate matrix is usually scaled so that the average rate of substitution at equilibrium equals 1. This means that tree branches are measured by the expected number of substitutions per site.

As a matter of a mathematical and computational convenience rather than biological reality, several simplifying assumptions are usually made. The substitution process is typically *homogeneous over time*, so the instantaneous rates are time-independent. The homogeneous process has an equilibrium distribution, which is also limiting when time approaches infinity. Globally time-homogeneous models use the same Q-matrix on all the branches of the tree. Locally time-homogeneous models are more realistic as they allow changing evolutionary patterns at tree nodes by using different Q-matrices on different branches (e.g., Yang 1998). Note that over-parameterized versions of locally time-homogeneous models quickly loose their advantages. Standard substitution models commonly allow any state to change into any other. Such Markov process is called *irreducible* and has a unique *stationary* distribution corresponding to the equilibrium codon frequencies $\pi = \{\pi_i\}$. *Time-reversibility* implies that the direction of the change between two states i and j is indistinguishable, so that $\pi_i p_{ij}(t) = \pi_j p_{ji}(t)$. This assumption helps to reduce the number of model parameters and is convenient when calculating the matrix exponential (Q-matrix of a reversible process has only real eigenvectors and eigenvalues; Keilson 1979). Fully unrestrained Q matrix for N characters defines an irreversible model with $N \times (N-1) - 1$ free parameters, while for a reversible process this number is $N \times (N+1)/2 - 2$. For

nucleotide data, estimation under unrestricted model is tricky (Yang 1994a; Klosterman et al. 2006) as (i) the computation of eigenvalues/vectors becomes more complex; (ii) estimates for branches descendent to the root are tightly correlated; and (iii) single gene samples typically do not contain sufficient information to estimate all parameters. Such difficulties mount with the increase in number of character states. Thus, AA and codons models are typically time-reversible and applied over unrooted trees.

The first two codon models, further referred to as MG (Muse and Gaut 1994) and GY (Goldman and Yang 1994), capitalized on the distinction between nonsynonymous and synonymous changes. Table 1 lists entries of Q-matrices defining several codon models. The MG model focused on estimating two separate parameters for synonymous and nonsynonymous substitution rates (α and β respectively). The GY model included the transition/transversion rate ratio κ, and modeled the selective effect indirectly using a multiplicative factor based on Grantham (1974) distances (table 1). Such approach had marginal success (Nielsen and Yang 1998) due to limitations of Grantham matrix and other aspects of AA classification. The GY model was later simplified to estimate the selective pressure explicitly using the single parameter ω (Yang 1998). This is essentially equivalent to the treatment in the MG model, as parameters α and β are unidentifiable on their own and only their ratio may be estimated.

More realistic codon models were subsequently developed based on these first models. GY-type models potentially can estimate 61 equilibrium codon frequencies (60 free parameters), whereas MG-type models rely on estimating only 12 equilibrium frequencies of target nucleotides at each of the three codon positions (9 free parameters). While less realistic, the latter approach has fewer parameters to be estimated and saves computational time. Also, small samples are often insufficient to estimate codon frequencies reliably. In practice, the codon frequencies are estimated empirically from data at hand. A model where all codon frequencies are estimated (empirically or by ML) is referred to as F61. Other variants may assume equal codon frequencies (Fequal) or estimate them from the observed frequencies

of four nucleotides (F1×4) or from three sets of frequencies of four nucleotides at the three codon positions (F3×4 model, usually describes data sufficiently well). Recently, variants of the GY and MG models have been compared within a Bayesian framework (Rodrigue, Lartillot, and Philippe, ms. submitted).

Maximum likelihood and Bayesian inference

The likelihood is a function of model parameters, and is proportional to the probability of the observed data (D), given the values of all parameters (P), and the substitution model (M): $L = p(\mathrm{D} \mid \mathrm{P}, \mathrm{M})$. ML estimation has convenient mathematical properties, making the method very attractive (Stuart, Ord, and Arnold 1999). For simplicity, the substitution process is often assumed to be *identical and independent for all sites* in a sequence, so that the total log-likelihood of data is a sum of site-log-likelihoods calculated via the pruning algorithm (Felsenstein 1981). ML parameter estimates (MLEs) are obtained by maximizing the likelihood function over the parameter space.

The Bayesian analysis introduces $p(\mathrm{P} \mid \mathrm{M})$, a prior probability distribution on the parameters of a phylogenetic model, representing biologist's beliefs about parameter distributions before collecting observations. Inferences about particular quantities are conducted by averaging over the posterior distribution. The posterior probability distribution of model parameters conditional on the data and the model is evaluated using Bayes' theorem:

$$p(\mathrm{P} \mid \mathrm{D}, \mathrm{M}) = \frac{p(\mathrm{D} \mid \mathrm{P}, \mathrm{M}) \times p(\mathrm{P} \mid \mathrm{M})}{p(\mathrm{D} \mid \mathrm{M})},$$

where $p(\mathrm{D} \mid \mathrm{M})$ is the normalizing constant, also known as the marginal likelihood of data, which is obtained by the integration over the parameter space.For discrete parameters (e.g., tree topology, site class), the integral is replaced by a summation over possible parameter values. In phylogenetic context, the Bayesian paradigm was introduced in the seminal works of Rannala and Yang (1996), Larget and Simon (1999), and Ronquist and Huelsenbeck (2003). Commonly used priors in the Bayesian phylogenetic analysis (Ronquist and Huelsenbeck 2003) assume that all phylogenetic trees are equally

probable *a priori* and that branch lengths are exponentially and independently distributed random variables. Codon frequencies π are drawn from a flat Dirichlet distribution, the transition-transversion ratio κ and the ω-ratio are ratio of two identically distributed exponential variables. Posterior probability distribution of model parameters is typically approximated by Markov chain Monte Carlo (MCMC; Metropolis et al. 1953; Hastings 1970).

Hypothesis testing and model selection

The accuracy of hypothesis testing and the biological relevance of parameter estimates depend on the validity of the model. While no model truly represents reality, capturing crucial and most visible evolutionary patterns is essential for informative inferences about underlying processes. Likelihood ratio tests (LRTs) compare nested hypotheses represented by their parameterizations (models). The LRT statistic is double the difference of log-likelihoods, and significance is evaluated using the asymptotic null distribution, such as the χ^2-distribution with degrees of freedom equal to the difference in the number of free parameters between the compared models (subject to regularity; Stuart, Ord, and Arnold 1999). When the theoretical null distribution is unknown or with an insufficient sample size, the empirical distribution may be used to approximate the significance threshold. Monte Carlo simulations are also applicable for non-nested models (Goldman 1993).

With more than two competing models, typical selection procedures in the ML framework include hierarchical or dynamical LRTs (hLRTs and dLRTs; Posada and Crandall 1998; 2001), AIC (Akaike 1973) and AIC_c (AIC corrected for sample size; Sugiura 1978); for review see Posada and Buckley (2004). For a series of nested models hLRTs and dLRT may be used, but correction for multiple testing is problematic. AIC and AIC_c are applicable to multiple non-nested models and do not require multiple testing correction. The models are simply rated by their ML-based scores, penalizing extra parameters. However, a simulation study of codon model selection showed that dLRTs starting with the most complex model and gradually removing parameters (backwards elimination) may be more accurate than approaches using AIC and AIC_c (Bao et al. 2007). The Bayesian analogue of AIC is

the BIC score (Schwarz 1978), it is also based on the maximized likelihood and is easily computed. Minin et al. (2003) suggested performance-based model selection based on BIC using the relative branch length error as a performance measure. Other Bayesian model selection procedures use posterior probabilities and Bayes factors (Jeffreys 1935; Wasserman 2000), requiring computationally expensive calculation of the marginal likelihood via MCMC. Nonetheless, the use of Bayes factors in model selection was successfully demonstrated for AA and codon models (Lartillot and Philippe 2006; Rodrigue, Philippe, and Lartillot 2006; Choi et al. 2007; Rodrigue, Lartillot, and Philippe, ms. submitted). Further work on codon model selection cannot be underestimated, especially given the great variety of useful codon models, as reviewed below.

When one model cannot be chosen with high confidence, it may be too risky to base conclusions on a single best-fitting model. Instead, several candidate models can be used to estimate a model-average of a parameter,.. For example, Kosakovsky Pond and Frost (2005c) used model-averaging approach to evaluate support for positive selection on different branches of a tree. Note that interpretation of parameters has to be compatible across models.

Accounting for variability of selective pressures

First codon models assumed constant nonsynonymous and synonymous rates among sites and throughout the phylogenetic history. Although most proteins evolve under purifying selection most of the time, positive selection may affect some lineages, and during episodes of adaptive evolution only a small fraction of sites in the protein have the capacity to increase the fitness of the protein via AA replacements (e.g., Gillespie 1991; Li 1997; Messier and Stewart 1997; Pupko and Galtier 2002). Thus approaches assuming constant selective pressure over time and over sites lack power in detecting genes affected by positive selection (Yang and Bielawski 2000). Consequently, various scenarios of variation in selective pressure were incorporated within GY and MG models. These models became very popular for detecting positive selection. Evidence of positive selection on a gene can be obtained by a LRT comparing two nested models, one of which (the null hypothesis) does not allow positive selection,

whereas another one does (the alternative hypothesis). Positive selection is detected if a model allowing sites or lineages under positive selection (with $\omega > 1$) fits data significantly better than the model restricting $\omega \leq 1$ at all sites and lineages. However, the asymptotic null distribution may vary from the standard χ^2 due to boundary problems or if some parameters become inestimable (e.g., Anisimova, Bielawski, and Yang 2001; Anisimova and Yang 2007).

Selective variability over time: branch models

One simple way to account for the variation of the selective pressure over time is to use a separate ω-ratio for each branch of a phylogeny ("free-ratio" model; Yang 1998). With T species, such model has an extra $2T-4$ free parameters in comparison with the "one-ratio" model (constant ω-ratio). A variety of branch models can be defined by constraining different sets of branches of a tree to have an individual ω. LRTs are used to decide (i) whether selective pressure is significantly different on a pre-specified set of branches, and (ii) whether these branches are under positive selection (see example in fig. 1A). Note that testing of multiple hypotheses on the same data requires a correction, so the overall false positive rate is kept at the required level (e.g., 5%). Correction for multiple testing reduces the power of the method, especially when many hypotheses are tested simultaneously (see discussion later). A biologically reasonable *a priori* hypothesis is not often available. Besides, even the most reasonable hypotheses have been proven wrong and the true scenario may be completely unexpected. Kosakovsky Pond and Frost (2005c) suggested detecting lineage-specific variation in selective pressure using a genetic algorithm (GA) – a computational analogue of evolution by natural selection. The GA approach was successfully applied to phylogenetic reconstruction (Lemmon and Milinkovitch 2002; Jobb, von Haeseler, and Strimmer 2004; Zwickl 2006). In the context of detecting lineage-specific positive selection, GA does not require an *a priori* hypothesis. Instead the algorithm samples regions of the whole hypotheses space according to their "fitness" measured by AIC_C . One hypothesis represents a scenario of lineage-specific variation in selective pressure on a T-taxon tree; it allows K branch-classes ($1 \leq K \leq 2T-3$), each with an individual ω parameter. The total number of different selection scenarios

(including one- and free-ratio models) is $\sum_{K=0}^{2T-3} S(2T-3,K)$, a sum of second kind Stirling numbers, each evaluating the number of ways to allocate $2T$-3 branches of an unrooted tree into K branch-classes. For example, with 5 sequences there exist 52 possible selection scenarios; but this increases to 115975 for 10 sequences. Thus, with many taxa it is necessary to limit the maximum number of branch-classes. The efficiency of the GA-based approaches also depends on their definitions of selection and mutation processes used to evolve the populations of hypotheses, and the stopping rule.

Selective variability among codons: site-models

Similar to among-site rate variation (Yang 1993; 1994b; Gu, Fu, and Li 1995; Mayrose, Friedman, and Pupko 2005), among-site variation of selective pressure may be described by various probability distributions of the ω-ratio, or of the nonsynoynous and synonymous rates. The simplest site-models use a discrete distribution with a pre-specified number of site-classes K (typically 3). Each site-class i =(0.. K-1) has an independent parameter ω_i estimated by ML together with proportions of sites p_i in each class. The discretized versions of continuous distributions (such as gamma and beta) or distributions mixture were also successfully applied (Yang et al. 2000; Kosakovsky Pond and Muse 2005). The distribution of selective pressure differs greatly from gene to gene and cannot be easily generalized to take a particular shape (Yang et al. 2000). To this end, the beta distribution (constrained between 0 and 1) is particularly useful as it accommodates a variety of possible distribution shapes. Yang et al. (2000) proposed a number of GY-type site-models, with some specifically designed to represent the null and alternative hypotheses in LRTs for positive selection. The performance of several LRTs was thoroughly tested in simulations (Anisimova, Bielawski, and Yang 2001; Swanson, Nielsen, and Yang 2003; Wong et al. 2004), identifying the most successful pairs of models. Consequently, small model-modifications were implemented to achieve the best accuracy vs. power properties of LRTs (Yang 2007). For example, the modified M1 model allows two site-classes, one with $\omega_0 < 1$ and another with $\omega_1 = 1$, representing a simplification of the neutral model of evolution, and

therefore can be used as the null hypothesis. The alternative model M2 extends M1 by adding a further (third) site-class with $\omega_2 \geq 1$ to accommodate sites evolving under positive selection. Another stringent LRT can be performed on the basis of the modified model M8 with two site-classes: one with sites where the ω-ratio obeys the beta distribution (so $0 \leq \omega \leq 1$, describing the neutral scenario), and the second, discrete class, with $\omega \geq 1$. Fixing the ω-ratio of this second class to 1, provides a sufficiently flexible null hypothesis, whereby all evolution can be explained by sites with ω from the beta distribution or from a discrete site-class with $\omega = 1$ (Swanson, Nielsen, and Yang 2003). Significance of the LRT comparing M1 vs. M2 is tested using the χ^2_2–distribution, whereas to compare M8 ($\omega = 1$) vs. M8 the mixture $\frac{1}{2}\chi^2_0 + \frac{1}{2}\chi^2_1$ should be used. *A posteriori* distribution of sites into site-classes may be estimated by Bayesian approaches (empirical, Yang et al. 2000; hierarchical, Huelsenbeck and Dyer 2004; BEB, Yang, Wong, and Nielsen 2005); for discussion see Scheffler and Seoighe (2005); Aris-Brosou (2006). In particular, when the LRT for positive selection is significant, sites under positive selection may be predicted if their posterior probability of coming from a class with $\omega \geq 1$ is sufficiently high (usually >0.95, but see Anisimova, Bielawski, and Yang 2002; Yang, Wong, and Nielsen 2005). Alternatively, Massingham and Goldman (2005) proposed a site-wise likelihood ratio estimation to detect sites under purifying or positive selection.

It is worth remembering that GY-type models assume a constant synonymous rate among sites, implying that rate variation among codons is solely due to the variation of the nonsynonymous rate. Recent studies question whether such an assumption is generally realistic (Chamary, Parmley, and Hurst 2006; Nackley et al. 2006; Kimchi-Sarfaty et al. 2007; Komar 2007). Anisimova, Nielsen and Yang (2003) suggested that failure to account for synonymous rate variation may be one of the reasons why LRTs for positive selection are vulnerable on data with high recombination rates (the other reason is relying on a single topology). Kosakovsky Pond and Muse (2005) incorporated synonymous as well as nonsynonymous rate variation in the MG model (table 1). Both d_N and d_S rates are described by general discrete distributions with K_n and K_s classes respectively, so that each site may come from any

of $K_n \times K_s$ combinations of nonsynonymous and synonymous rate classes (typically $K_n=K_s=3$). Since only products of rates and times can be estimated (Felsenstein 1981), the synonymous rate distribution is restricted to have a fixed mean. The ω -ratio can then be estimated for each combination as a ratio of the correspondent rates. Presence of a site-class with $\omega > 1$ can be taken as a support for positive selection on a gene, and the Bayesian approach is used to predict the allocation of sites into $K_n \times K_s$ possible d_N and d_S site classes. However, even if positive selection at some sites is indicated by MLEs, it should not be automatically accepted. This may be an artifact of ML estimation, especially since the estimation of ω relies on the ratio of two estimated parameters d_N and d_S. Unambiguous evidence of positive selection is obtained by showing that model with positive selection fits data significantly better compared to the nested null model that does not allow sites under positive selection. The proposed MG-type site models offer no such null hypothesis, and therefore no rigorous way of testing for positive selection. Instead, the extent of synonymous rate variation may be tested with a LRT comparing the null model restricting d_S to be constant vs. a more flexible model that allows d_S to vary; significance is determined using $\chi^2_{2K_s-2}$. Such LRTs may provide important insights in studies of protein anomalies related to synonymous changes in coding sequences (Nackley et al. 2006; Kimchi-Sarfaty et al. 2007; Sauna et al. 2007). Scheffler, Martin, and Seoighe (2006) extended MG-type models with d_N and d_S site-variation to allow a topology change at the detected recombination breakpoints. Indeed, fast-evolving pathogens (such as viruses) typically undergo frequent recombination that is likely to change either the whole shape of the underlying tree, or only the apparent branch lengths. While the efficiency of the approach depends on the success of inferring recombination breakpoints, the study demonstrated that taking into account alternative topologies achieves a substantial decrease of false positive inferences of selection while maintaining reasonable power. In a related development, Wilson and McVean (2006) used an approximation to a population genetics coalescent with selection and recombination. Inference was performed on both parameters simultaneously using the Bayesian approach with reversible-jump MCMC.

Site-models that do not use *a priori* partitioning of codons (as those described above) are known as random-effect (RE) models (Kosakovsky Pond and Frost 2005a). In contrast, fixed-effect (FE) models categorize sites based on a prior knowledge, e.g., according to tertiary structure for single genes, or by gene category for multi-gene data (Yang and Swanson 2002; Bao et al. 2007). FE models can also be defined by inferred recombination breakpoints, useful for inferences of positive selection from recombining sequences (Kosakovsky Pond et al. 2006b; Kosakovsky Pond et al. 2006c). Apart from modeling among-site variation of selective pressure, FE models can include among-site variation of other evolutionary parameters, such as background mutation rate, transition/transversion bias and codon frequencies. Given an appropriate partitioning, FE models are useful to study heterogeneity among partitions, although a priori information is often unavailable. FE models with each site being a partition lead to the "infinitely many parameter trap", and so should be avoided (Felsenstein 2004).

Whether codons are partitioned *a priori* or not, all the discussed above site-models require specification of the number of selection site-classes. While an arbitrary choice of 3 classes seems sufficient in most cases, using the Dirichlet process to infer the number of site-classes may be appealing (as implemented in the full Bayesian framework by Huelsenbeck et al. 2006).

Temporal and spatial variation of selective pressure

Several solutions were proposed to simultaneously account for differences in selective constraints among codons and the episodic nature of molecular evolution at individual sites. The GY-type branch-site models were primarily designed for detecting positive selection (Yang and Nielsen 2002; Yang, Wong, and Nielsen 2005; Zhang, Nielsen, and Yang 2005). For example, model MA assumes four classes of sites. Two classes contain sites evolving constantly over time: one under purifying selection with $\omega_0 < 1$, another with $\omega_1 = 1$. The other two site-classes allow selective pressure at a site to change over time on a pre-specified set of branches, known as the foreground; these variable classes are derived from the constant classes so that sites typically evolving with $\omega_0 < 1$ or $\omega_1 = 1$ are allowed to be under positive selection with $\omega_2 \geq 1$ on the foreground. Testing for positive selection on the hominid

branch (fig. 1A) involves a LRT comparing a constrained version of MA (with ω_2=1) vs. an unconstrained MA; significance is tested using $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$. Compared to branch models, the branch-site formulation improves the chance of detecting short episodes of adaptive pressure affecting only a small fraction of sites.

Specific models were proposed to study differences in selective constraints between pre-specified clades, e.g., those resulted from speciation, gene duplication or due to parasite adaptation to a different host (Forsberg and Christiansen 2003; Bielawski and Yang 2004). For example, model MD with three site classes (Bielawski and Yang 2004) has two sites classes evolving constantly over time, each with an individual ω-ratio, ω_0 and ω_1, while sites from the third class may evolve under different selective pressures in the two clades, with two clade-specific parameters ω_2 and ω_3. The null hypothesis assuming no difference between the clades, is constructed by setting ω_2=ω_3. This resulting model is equivalent to the site model M3 with three discrete site classes, each with individual unconstrained ω-ratios (Yang et al. 2000). Thus, the LRT comparing M3 vs MD evaluates the difference in selective constrains between the two clades (or any two arbitrary pre-specified non-overlapping branch sets of a tree), significance of which is tested with χ_1^2. For example, in figure 1A let us designate the primate clade as the foreground, then the LRT of M3 vs MD can be used to test if some sites in the primate clade evolved under significantly different selective pressures to those in the rodent clade.

Once again, the major drawback of described branch-site models is their reliance on a biologically viable *a priori* hypothesis. In context of detecting sites and lineages affected by positive selection, one possible solution is to perform multiple branch-site LRTs, each setting a different branch at the foreground (Anisimova and Yang 2007). In the hypothetical example of 7 species (fig. 1), a total of 11 tests (for an unrooted tree) are necessary in the absence of prior information. Multiple test correction has to be applied to control excessive false inferences. This strategy tends to be conservative but can be sufficiently powerful in detecting episodic instances of adaptation. As with all model-based

techniques, precautions are necessary for data with unusual heterogeneity patterns, which may cause deviations from the asymptotic null distribution and thus result in an elevated false positive rate (Anisimova and Yang 2007).

In the case of episodic selection where any combination of branches of a phylogeny can be affected, Bayesian approaches in lieu of the standard LRTs and multiple testing have been suggested. The multiple LRT approach is most concerned with controlling the false-positive rate of selection inference, and is less suited to infer the best-fitting selection history. In the hypothetical example (fig. 1) a total of 2^7-1 = 127 selection histories (excluding the history without selection on any branch) need to be considered. The Bayesian analysis allows a probability distribution over possible selection histories to be computed, and therefore permits estimates of prevalence of positive selection on individual branches and clades. Such an approach evaluates uncertainty in selection histories using their posterior probabilities and allows robust inference of interesting parameters such as the switching probabilities for gains and losses of positive selection (Kosiol et al. 2008).

MG-type models of d_N and d_S site-variation also may be extended to allow changes of selective regimes on different branches. This is achieved (similar to Yang 1998) by adding further parameters, one per branch, describing the deviation of selective pressure on a branch from the average level on the whole tree under the site model (Kosakovsky Pond and Muse 2005). Such a model is parameter-rich and can be used for exploratory purposes on data with long sequences, but does not provide a robust way of testing whether ω>1 on a branch is due to positive selection on a lineage or due to inaccuracy of the ML estimation. The branch-model selection with GA may also be adapted to incorporate d_N and d_S among-site variation, although this imposes a much heavier computational burden (Kosakovsky Pond and Frost 2005c).

In branch- and branch-site models, change in selection regime is always associated with nodes of a tree, but the selective pressure remains constant over the length of each branch (as in fig. 1A-B). Guindon et al. (2004) proposed a Markov-modulated model where switches of selection regimes may

occur at any time on the phylogeny (fig. 1C). In a covarion-like manner (Fitch 1971), this codon model combines two Markov processes: one governs the codon substitution (models M2 and M3, Yang et al. 2000), and the other specifies rates of switches between selective regimes. Such model does not require *a priori* knowledge of lineages evolving under positive selection. Changes between different selective regimes (purifying, neutral and positive) are not equiprobable, and the relative rates of changes from neutral to positive and purifying to positive may be estimated, which may be especially useful to study viral dynamics.

Modeling site-dependence

Despite the common assumption of site-independence, real data exhibits complex dependencies of evolutionary patterns among sites. For example, proteins often include conserved and variable linear domains, so that rates at neighbor-sites tend to be correlated; CpG and CpNpG effects and overlapping reading frames cause complex dependencies. Interactions among sites can also be non-local, necessary for protein stability and for its specific function.

In an evolutionary context, modeling general site-interdependencies is non-trivial, as it involves rate matrices of very large dimensions. In a brave attempt, Robinson et al. (2003) explicitly modeled structural constraints within a standard phylogenetic framework. The Markov process specified at the nucleotide level is, in fact, equivalent to the process generated by a $61^N \times 61^N$ matrix, with single entries describing rates of change from one *N*-codon sequence to another. The only allowed non-zero entries correspond to sequence changes due to no more than one nucleotide (table 1). Under this model, the effective rate of each type of possible nonsynonymous events at a given site is dependent on the states at other sites, and can change when these sites change states over time. The model relies on protein threading, and so requires a known 3D protein structure, which is assumed conserved for all analyzed homologues. Measures of solvent accessibility and pairwise sequence-structure compatibility correlate with free energy of the folded protein, and are therefore used to adjust rates of sequence change (table 1). Parameters are then estimated in the Bayesian framework by MCMC sampling over possible

pairwise histories. Based on an appropriate set of sequence fitness measures, the model can include site-dependencies other than those imposed by protein structure.

Context-dependent extensions of the GY and MG models accommodate the CpG effect (Pedersen, Wiuf, and Christiansen 1998; Jensen and Pedersen 2000; Siepel and Haussler 2004b), as well as methylation at CpA and CpT dinucleotides (Huttley 2004), and overlapping reading frames (Pedersen and Jensen 2001). Some models introduced dependency only within the same codon (Pedersen, Wiuf, and Christiansen 1998; Huttley 2004) so that likelihood is calculated using the site-independence. This approach fails to account for CpG dinucleotides formed at the codon boundaries. Other models are described by instantaneous rates at a base that depend upon the states at neighboring nucleotides (Jensen and Pedersen 2000; Pedersen and Jensen 2001). Assuming such conditional higher order Markov process makes ML parameter estimation intractable and, even with MCMC, it is only applicable to pairs of sequences. Christensen, Hobolth, and Jensen (2005) proposed a generalization, approximated with the pseudo-likelihood-based estimation and using expectation-maximization (EM) algorithm. But such approach is still applicable to very limited phylogenies. Siepel and Haussler (2004b) extended context-dependent substitution to a general phylogeny, at the expense of limiting the full process-based process defined by Jensen and Pedersen (2000). A second order Markov process running at the tips of a tree is only approximated, since interdependencies in the ancestral sequences are ignored. The likelihood is calculated with a modified pruning algorithm and optimized with EM.

Applications of CpG codon models to HIV and mammalian data confirmed that methylation plays significant role in the evolution of protein-coding sequences (Pedersen, Wiuf, and Christiansen 1998; Jensen and Pedersen 2000; Huttley 2004; Siepel and Haussler 2004b). Hobolth (2006) also included the CpNpG effect and applied their codon model to single coding sequences from tomato. Their analysis showed that CpG and CpNpG effects are not correlated suggesting their diverse biological roles.

Other models with local site-dependence include autocorrelated rates (for DNA, Yang 1995; and for proteins, Stern and Pupko 2006). Mayrose et al. (2007) described autocorrelation of synonymous and, separately, of nonsynonymous rates using two Hidden Markov Models (HMMs), with hidden states at each codon represented by synonymous and nonsynonymous rate-classes. The backward dynamic programming algorithm permits likelihood calculation (Durbin et al. 1998). LRTs may be used to test for synonymous and nonsynonymous rate variation and autocorrelation. The model is particularly relevant for viral sequences due to possible selection on regulatory and overlapping codon regions.

Empirical codon models

Unlike AA models, codon substitution models are traditionally parametric. Despite their apparent success, such models do not incorporate physico-chemical biases (but see Robinson et al. 2003; Sainudiin et al. 2005; Wong, Sainudiin, and Nielsen 2006; table 1), or simultaneous multiple nucleotide changes (but see Whelan and Goldman 2004). Empirical substitution matrices generalize evolutionary patterns, "averaging" over large quantities of data. Schneider, Cannarozzi, and Gonnet (2005) used pairwise alignments to estimate a PAM-style empirical codon matrix (CodonPAM), describing transition probabilities for a range of distances. CodonPAM may be extrapolated into a full model (e.g., Kosiol and Goldman 2005), although for deep divergences the resulting transitions will lack accuracy due to limitations of distance estimation. A full ML estimation of a general time-reversible codon model (ECM) involves 1891 free parameters (table 1), but became feasible thanks to the fast EM algorithm (Holmes and Rubin 2002; Klosterman et al. 2006). A visual comparison of matrices defining ECM and GY (figures 1A-B of Kosiol, Holmes, and Goldman 2007) is sufficient to see the apparent differences: significant proportions of changes (~24%) involve simultaneous multiple nucleotide substitutions. The double and triple nucleotide changes significantly improve the likelihoods of codon models. However, previous theoretical and experimental studies show noticeably lower proportions of double and triple changes (Averof et al. 2000; Bazykin et al. 2004; Whelan and Goldman 2004). There

is no clear evidence to suggest if such multiple changes are really instantaneous, or whether they seem instantaneous as a result of subsequent single nucleotide substitutions occurring on a much faster timescale than other single nucleotide changes. One explanation for the multiple nucleotide changes is that compensatory changes are fixed rapidly (population level rather than species level) even if the intermediate mutation is deleterious. Another clearly visible phenomenon stems from the nature of genetic code and underlying biases: the codons tend to cluster into almost invariant sets (AIS) having a high rate of changes among codons of each set, but very small rates between sets. The AIS method was suggested and explored based on AA substitution models. Applied to the empirical codon model it shows the importance of the genetic code and the physical-chemical properties of the amino acids for codon substitution patterns (Kosiol, Goldman, and Buttimore 2004; Kosiol 2006).

On the other hand, parametric models have been very successful in applications studying biological forces shaping protein evolution of individual genes and combining the advantages of parametric and empirical approaches offers a promising direction. Kosiol, Holmes, and Goldman (2007) explored a number of combined codon models that incorporated empirical AA exchangeabilities from ECM while using parameters to study selective pressure, transition/transversion biases and codon frequencies. Similarly, AA exchangeabilities from (suitable) empirical AA matrices may be used to alter probabilities of nonsynonymous changes, together with traditional parameters ω, κ, and codon frequencies π_j (Doron-Faigenboim and Pupko 2007). Such an approach accommodates site-specific variation of selective pressure and can be further extended to include lineage-specific variation. Combined empirical and parametric models will therefore become more frequent in selection studies. However, selecting an appropriate model is of utmost importance and needs further study. In particular, parameter interpretations may change with different model definitions, since empirical exchangeabilities already include average selective factors and other biases (Kosiol, Holmes, and Goldman 2007). Thus, selection among alternative parameterizations requires detailed attention.

More codon models and their applications

Studying selective pressure on a protein

Codon models have been especially successful at detecting positive selection and identifying codon sites responsible for adaptive diversification. It is a good practice to verify the robustness of conclusions by repeated inferences under different models or by model averaging. Codon models are now commonly used to identify candidate genes under positive selection in large-scale genomic studies (Clark et al. 2003; Nielsen et al. 2005; Arbiza, Dopazo, and Dopazo 2006; Anisimova et al. 2007; Anisimova and Liberles 2007; Studer et al. 2008). While the ω-ratio allows detection of recurrent diversifying positive selection, a separate parameter describing directional selection is easily accommodated within a standard codon model, whereby mutation rates towards (or away from) a pre-specified amino acid may be estimated (table 1; Seoighe et al. 2007). Such a model is time-directional and non-reversible, as it uses viral sequence pairs obtained from patients before and after the treatment. As more data is becoming available on the specific functional roles of amino acids, in particular from structural or mutagenesis studies, it is increasingly possible to find direct links between selection and function. Conversely, codon model based methods to identify individual residues under positive selection in proteins (e.g., BEB, Yang, Wong, and Nielsen 2005) are increasingly used to generate biological hypothesis for verification through laboratory experiments. For example, a small segment of the immune defense protein TRIM5α was identified to be under positive selection, and functional analysis using mutagenesis confirmed the importance of the segment in species-species viral inhibition (Sawyer et al. 2005). With the sequencing of more and more genomes, the methods to detect site-specific selection in genome scale analysis will be more and more informative, and there is a potentially productive feedback loop between computational phylogenetic methods and functional characterization of sites. Presence of interactions between sites involved in developing drug resistance may be tested based on the conditional selection model (Chen and Lee 2006). FE models can be adapted to detect differential population-specific adaptation of HIV to human populations (Kosakovsky Pond et al.

2006a). In general, identifying various kinds of selection are classic focal points for evolutionary biologists, and codon models are also proving invaluable in studies of pathogenic drug resistance, disease progression and epidemics dynamics, important in vaccine design and treatment strategies (Lemey et al. 2005; Lemey, Van Dooren, and Vandamme 2005; Chen and Lee 2006; Kosakovsky Pond et al. 2006a; Seoighe et al. 2007; Carvajal-Rodriguez et al. 2008; Kosakovsky Pond et al. 2008). Using codon models to study the evolution of gene families is also well documented (Bielawski and Yang 2003; Aguileta, Bielawski, and Yang 2004; Balakirev, Anisimova, and Ayala 2006; Studer et al. 2008). While selection studies are still predominant, plenty of other applications are emerging.

Codon usage bias and the missing link

Codon usage is non-random among both genes and species. Pressure to optimize translational efficiency, robustness and kinetics may lead to synonymous codon bias. A classic problem is to untangle the effects of translational selection and mutational biases. Selection against the non-optimal codons leads to a negative correlation between codon bias and lower synonymous substitution rates (e.g., Akashi and Eyre-Walker 1998). Codon usage bias is often studied with various codon adaptation indexes (e.g., CAI, Sharp and Li 1987; ENC, Wright 1990), while the synonymous rates may be estimated with ML under a codon model (Goldman and Yang 1994; Yang and Nielsen 2000). However, different ways of modeling unequal codon frequencies impose different assumptions about the mutation process, leading to different conclusions (Aris-Brosou and Bielawski 2006). Codon models with site- or context-dependencies seem very appealing for analyses of the codon bias, but come at a heavy computational cost. Codon usage, and asymmetric selective effects in particular, may also be studied using Markov models with fewer states, corresponding to groups of codons translated by distinct tRNAs (Higgs, Hao, and Golding 2007).

Particularly useful for studying codon bias, are codon models that make use of the ultimate link between intraspecific and population genetics parameters. Since the evolution in populations effectively shapes the intraspecific patterns, several studies attempted to recreate this important missing link

(Halpern and Bruno 1998; McVean and Vieira 1999; 2001; Nielsen and Yang 2003; Thorne et al. 2007; Yang and Nielsen 2008). The classical assumption is that the rate of codon change is a product of the mutation rate and the mutation fixation probability (Kimura 1962). Such "mutation-selection" models vary by constraints imposed on variability of these key components over time and among sites. For example, Nielsen et al. (2007) used one selection coefficient for optimal codon usage for each branch of a phylogeny, and estimated these jointly with the ω-ratio by ML. Since codon usage bias evolves over time (Duret 2002), such approach is useful to study ancestral codon usage bias (e.g., the model confirmed reduction in selection for optimal codon usage in *D. melanogaster*; Nielsen et al. 2007), but requires *a priori* knowledge of preferred and unpreffered codons. Improving on previous work, Yang and Nielsen (2008) separately considered the mutation and selection on codon usage, modeling the latter by individual codon fitness parameters (FMutSel model; table 1). Together with mutational-bias parameters, this allows to estimate optimal codon frequencies for a gene across multiple species. Testing whether the codon bias is due to the mutational-bias alone is straightforward with the LRT (FMutSel0 vs. FMutSel; with standard genetic code use χ^2_{41}).

Understanding how the interspecific parameters relate to population parameters gives further insights to how changing demographic factors influence observed intraspecific patterns. For example, the intraspecific selective pressure measured by the ω-ratio is affected by changes in population size (Nielsen and Yang 2003; Thorne et al. 2007), which should be taken into account when comparing species data.

Codon-based ancestral reconstruction and the Bayesian substitution mapping

Codon models are equally suitable for ML or Bayesian reconstruction of ancestral coding sequences (Yang, Kumar, and Nei 1995; Nielsen 2002; Weadick and Chang 2007). Given that ancestral reconstruction is particularly sensitive to model choice (Chang 2003), codon models may have advantages over DNA and AA models, since they provide a better data fit. Several studies used inferred ancestral sequences by parsimony or ML to count synonymous and nonsynonymous substitutions to

infer positive selection (e.g., Crandall and Hillis 1997; Messier and Stewart 1997; Suzuki 2004). Although tempting, inferred ancestral sequences should not be treated as observed, as uncertainties in inferences may introduce biases to subsequent estimations (Nielsen 2002). Rather, the inferred distribution of ancestral states provides a good starting point for experimental testing (Chang 2003; Ugalde, Chang, and Matz 2004). Ancestral proteins can then be recreated and studied in the laboratory (Weadick and Chang 2007; Hult et al. 2008). Poon et al. (2007a; 2007b) inferred interacting sites in the HIV *env* gene by estimating a distribution of ancestral codons and then using it to infer a Bayesian network representing a joint distribution of substitutions observed in the sequence. Parametric bootstrap was used to account for uncertainty in codon reconstruction.

The Bayesian mapping implementations of substitution models (also known as data augmentation) sample substitution mappings over a tree from their posterior distribution and estimate model parameters via MCMC sampling. This enables the formal treatment of uncertainty of the ancestral reconstruction. Such an approach is useful in inferences of phylogeny-associated data statistics, formulated as functions of such mappings. For coding sequences, estimating ages and distributions of synonymous and nonsynonymous substitutions (Nielsen 2001; Nielsen 2002) may reveal the origin and spread of heritable traits, and is useful for testing population genetics and molecular evolution models. Nielsen and Huelsenbeck (2002) applied the substitution mapping to infer sites under positive selection, reproducing results obtained by site-models (Yang et al. 2000). Extending this approach, Zhai, Slatkin, and Nielsen (2007) included variation of selective pressure over sites and over time, and illustrated method's performance on the surface antigen of influenza H3N2. The structure-dependent codon model described above (Robinson et al. 2003) uses a similar approach. Moreover, data augmentation may be used to infer co-evolving codon positions (as for AA data, Dimmic et al. 2005), and to implement other sophisticated codon models. Fortunately, recent improvements of the sampling procedure render the approach very efficient (Rodrigue, Philippe, and Lartillot 2008).

Phylogenetic reconstruction

Ignoring the genetic code structure or synonymous changes in coding sequences causes information loss at wide range of divergences (e.g., Ren, Tanaka, and Yang 2005; Seo and Kishino 2008). However, tree-inference from coding data is typically conducted under DNA and AA models. This is hardly surprising given the lack of efficient codon-based tree-search implementations. The progress is hampered by heavy computational costs associated with 61×61 matrices. Indeed, for large datasets there is currently no feasible way to infer a phylogeny under a codon model. For small datasets, this is possible with CODEML from the PAML package (Yang 2007), although the implemented heuristic algorithm is not the most efficient. One possible approach is to reconstruct several good starting trees under DNA and AA models and then improve these trees with efficient ML heuristics under codon models. Another promising direction to implement codon models within the Bayesian framework and sample the topology space with an efficient MCMC (Rodrigue, Philippe, and Lartillot 2008). Meanwhile, using DNA models for tree inference cannot be avoided, but three codon positions should be treated as different data partitions. The use of codon models may prove an asset when it comes to comparison of several candidate trees inferred under DNA or AA models (Ren, Tanaka, and Yang 2005).

Molecular dating

Diverse evolutionary forces may affect *absolute* synonymous and nonsynonymous substitution rates differently; while deviations in synonymous rates may be due to changes in generation time or mutations rate, nonsynonymous rates may also be affected by changes in effective population size and natural selection patterns (Seo, Kishino, and Thorne 2004). Thus, studying the absolute synonymous and nonsynonymous rates should be more informative than solely considering their estimated ratio. Inevitable confounding of time and rates is resolved by calibration based on fossil ages or the sampling times of rapidly evolving organisms, such as viruses. Recent DNA dating techniques relax molecular clock by allowing autocorrelated or independent rates (Kishino, Thorne, and Bruno 2001; Drummond et

al. 2006; Rannala and Yang 2007). A Bayesian approach via MCMC is then used to estimate divergence dates and mutation rates. Dating techniques were successfully adapted to codon models, so the absolute rates of synonymous and nonsynonymous changes can be estimated together with divergence dates (Seo, Kishino, and Thorne 2004; Lemey et al. 2007). As a result of absolute rates comparison in longtitudal HIV samples (*env*), Lemey et al. (2007) found that slower progression to AIDS is strongly associated with slower synonymous rates, suggesting slower viral replication and longer generation times.

Codon-based dating should fend off possible selective effects, simultaneously offering more informative inference. The synonymous changes are expected to be very informative for recent divergences, whereas nonsynonymous changes, once selection is accommodated, reveal details of distant relationships (Ren, Tanaka, and Yang 2005). Note that to increase the accuracy of divergence estimates, multiple gene data is necessary. This motivates further extension of codon-based dating techniques to multiple genes.

Computer implementations

A variety of GY-type codon models, including recent selection-mutation model, are implemented in the CODEML program of PAML (Yang 2007). Codon-based ML tree-inference with step-wise addition is equally available in CODEML, but is inefficient and slow with > 10-15 taxa. A large variety of MG-type models are available in the HYPHY package (Kosakovsky Pond, Frost, and Muse 2005) or on the web-server Datamonkey (Kosakovsky Pond and Frost 2005b). MrBayes is the Bayesian tree-inference software, which implements several simple site-models (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003). FitModel is the ML implementation of the switching codon model (Guindon et al. 2004), and SIMMAP implements a stochastic mapping of mutational histories on a phylogeny (Nielsen 2002; Bollback 2006). Darwin programming environment provides a series of functions dealing with codon matrices (Gonnet et al. 2000). Codon tools using the pair-wise empirical codon matrix (Schneider, Cannarozzi, and Gonnet 2005) are also accessible via a web server. Selecton web-server

(Stern et al. 2007) offers several site-models as well as the combined model described in Doron-Faigenboim and Pupko (2007). Equally, source-code is provided for download. BEAST package (Drummond and Rambaut 2007) allows Bayesian inferences with uncorrelated mutation rates on a tree, based on the simplest site model (Nielsen and Yang 1998).

Simulation of coding sequences

EVOLVER from the PAML package simulates codon data on a specified tree with branch lengths. Selection regimes can be specified for different sites and branches on a tree. Coalescent simulation of coding sequences with recombination is possible using CodonRecSim (Anisimova, Nielsen, and Yang 2003) and Recodon (Arenas and Posada 2007). In addition, Recodon enables simulations under more complex demographic scenarios and under a variety of codon models. SISSI (Gesell and von Haeseler 2005) allows simulation under a pre-specified dependency structure of the codons. CodonMutate function of Darwin uses the empirical pairwise matrix (Schneider et al. 2005) to simulate pairs of sequences. The program EvolveAGene (Hall 2005; 2008) evolves a real coding sequence along the tree using experimentally determined mutation spectrum (from *E. coli*). The simulated process is heterogeneous with mutation, selection and indels.

Future developments

The utility of codon models for molecular sequence analysis is beyond doubt, as demonstrated by rapid expansion of codon-based applications. Numerous studies examined selective pressures in proteins using codon models with site, branch and branch-site specific variation. Such models became commonplace in genome-scale analyses and have resulted in a greater understanding of the heterogeneity of the evolutionary process. Phylogenomics coupled with improvements in computer hardware, has allowed long-held and limiting assumptions about molecular evolution to be relaxed and a new generation of codon models to be developed. While many codon-based techniques were first implemented for DNA and AA data, such methodological transfer was very beneficial and should continue. For example, mixture and general heterogeneous (Koshi and Goldstein 1995; Koshi and

Goldstein 1997; Lartillot and Philippe 2004; Blanquart and Lartillot 2008; Whelan 2008) as well as nonreversible (Galtier, Tourasse, and Gouy 1999) codon models may become increasingly popular. Incorporating indels within probabilistic codon-based framework is another interesting direction that may become possible given recent developments for DNA and RNA (Rivas 2005; Bradley and Holmes 2007). Content-dependent codon models are still in their infancy and deserve further attention.

Already existing codon models have been challenged by their greater use in analyzing genomic data and have been brought to higher levels of sophistication. One promising direction is a further development and fine-tuning of combined empirical and parametric codon models (Kosiol, Holmes, and Goldman 2007; Doron-Faigenboim and Pupko 2007). The empirical component of these models reflects clear distinctions between different nonsynonymous changes, which are treated equally in traditional codon models. Furthermore, validity of the traditional selective measure d_N/d_S is often thought conditional on neutrality of evolution at synonymous sites. Yang and Nielsen (2008) use their mutation-selection model to argue that d_N/d_S-based inference does not require the neutrality at synonymous sites. The potential selection acting on synonymous sites is better thought of as selection on the DNA level affecting both synonymous and nonsynonymous sites equally. This apart, models allowing separate variable synonymous and nonsynonymous rates should provide further insights about evolutionary patterns at synonymous sites. Finally, strengthening the link between the macroevolution and the population genetics is vital for a better understanding of the interplay among different demographic factors and selection over time.

Caution should be taken against unnecessary over-parameterization. This is true with both ML and Bayesian implementations. While the Bayesian approach is more tolerant of parameter-rich models, large sample sizes necessary to estimate such models require longer MCMC runs with slower point likelihood calculations and trickier convergence. Essentially, the computational burden of codon-based analyses presents the next challenge to develop faster, efficient methods, and a greater use of heuristics and approximations.

More and more, codon models are used outside the traditional field of phylogenetic modeling and reconstruction in alignment programs, gene finding and functional annotation of genomes in general. Currently, codon alignment may be constructed by back-translating the aligned AA sequences given the corresponding unaligned DNA. Alternatively, empirical codon matrices could be used to construct codon alignments directly (Loytynoja and Goldman 2005; Schneider, Cannarozzi, and Gonnet 2005). Methods for simultaneous alignment and phylogenetic inference based on codon models were also proposed (e.g., Suchard and Redelings 2006). Codon patterns and codon substitution models are used in functional annotation of genomes and gene-finders using phylogenetic hidden Markov models (e.g., Siepel and Haussler 2004a).

The future of modeling codon sequence evolution and the field of genomics are intertwined; 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 codon evolution.

Acknowledgements

We thank Nicolas Rodrigue and the anonymous reviewers for valuable comments on the manuscript. M. A. is supported by the Swiss Federal Institute of Technology (ETH, Zurich), and C.K. is partially funded by NSF grants DBI-0644111 and NSF0516310.

Figure 1

Figure legend. – A hypothetical example used to illustrate the utility of codon models with variation of selective pressure. In the example gene tree, node S represents a speciation event, giving rise to the primate and the rodent clades; and node D is a gene duplication event preceding the speciation of rat and mouse.

(A) Hypothesis testing with branch and branch-site clade models. A set of branches to test for functional divergence or positive selection is selected *a priori* (before analyzing the data), based on a biologically justified hypothesis. *To test for functional divergence* after speciation event S, use the LRT to compare the one-ratio model ($\omega_P = \omega_R$) vs. two-ratio model allowing different selective pressures in primates and rodents ($\omega_P \neq \omega_R$). Rejecting the null shows that the two copies of a gene evolve under different selective pressures. Such test is more powerful if clade models are used (Bielawski and Yang 2004). Imagine that due to some environmental changes, the gene function adapted to new conditions on the branch leading to the hominids. *To test for positive selection*, designate the branch in question with a separate parameter $\omega_{hominid}$. With branch models, set $\omega_P = \omega_R$ or $\omega_P \neq \omega_R$, depending on the outcome of the previous test. Then use a LRT to compare the null "$\omega_P = \omega_{hominid}$" against the alternative "$\omega_P \neq \omega_{hominid}$". Even when such test is significant and the MLE of $\omega_{hominid} > 1$, only a further testing confirms whether this is not due to an estimation error. This can be done by yet another LRT of the null "$\omega_{hominid}=1$" against an alternative where $\omega_{hominid}$ is freely estimated. A more direct and powerful way is to use LRTs based on branch-site models with the hominid branch at the foreground (Yang, Wong, and Nielsen 2005).

(B) A schematic illustration of a changing distribution of sites in branch-site and clade models. Here we test for positive selection after the duplication event D. Note that a site might experience a change of selective regime only at a node of a tree. The substitution process at each site is homogeneous along each branch, but not throughout the phylogeny.

(C) An illustration of a possible selection history at one site in Markov modulated models (Guindon et al. 2004). Each site can switch between existing selective regimes at any time on the phylogeny. These changes are not associated with tree nodes, unlike in branch, branch-site and clade models. Thus, the substitution process is no longer homogeneous along each branch. However, the model is still time-reversible.

References

Aguileta, G., J. P. Bielawski, and Z. Yang. 2004. Gene conversion and functional divergence in the beta-globin gene family. J Mol Evol **59**:177-189.

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. Pp. 267-281 *in* B. N. Petrov, and F. Csaki, eds. Second International Symposium on Information Theory. Akademiai Kiado, Budapest.

Akashi, H., and A. Eyre-Walker. 1998. Translational selection and molecular evolution. Curr Opin Genet Dev **8**:688-693.

Anisimova, M., J. Bielawski, K. Dunn, and Z. Yang. 2007. Phylogenomic analysis of natural selection pressure in Streptococcus genomes. BMC Evol Biol **7**:154.

Anisimova, M., J. P. Bielawski, and Z. Yang. 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Mol Biol Evol **18**:1585-1592.

Anisimova, M., J. P. Bielawski, and Z. Yang. 2002. Accuracy and power of bayes prediction of amino acid sites under positive selection. Mol Biol Evol **19**:950-958.

Anisimova, M., and D. Liberles. 2007. The quest for positive election in the era of comparative genomics. Heredity.

Anisimova, M., R. Nielsen, and Z. Yang. 2003. Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. Genetics **164**:1229-1236.

Anisimova, M., and Z. Yang. 2007. Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites. Mol Biol Evol **24**:1219-1228.

Arbiza, L., J. Dopazo, and H. Dopazo. 2006. Positive selection, relaxation, and acceleration in the evolution of the human and chimp genome. PLoS Comput Biol **2**:e38.

Arenas, M., and D. Posada. 2007. Recodon: coalescent simulation of coding DNA sequences with recombination, migration and demography. BMC Bioinformatics **8**:458.

Aris-Brosou, S. 2006. Identifying sites under positive selection with uncertain parameter estimates. Genome **49**:767-776.

Aris-Brosou, S., and J. P. Bielawski. 2006. Large-scale analyses of synonymous substitution rates can be sensitive to assumptions about the process of mutation. Gene **378**:58-64.

Averof, M., A. Rokas, K. H. Wolfe, and P. M. Sharp. 2000. Evidence for a high frequency of simultaneous double-nucleotide substitutions. Science **287**:1283-1286.

Balakirev, E. S., M. Anisimova, and F. J. Ayala. 2006. Positive and negative selection in the beta-esterase gene cluster of the Drosophila melanogaster subgroup. J Mol Evol **62**:496-510.

Bao, L., H. Gu, K. A. Dunn, and J. P. Bielawski. 2007. Methods for selecting fixed-effect models for heterogeneous codon evolution, with comments on their application to gene and genome data. BMC Evol Biol **7 Suppl 1**:S5.

Bazykin, G. A., F. A. Kondrashov, A. Y. Ogurtsov, S. Sunyaev, and A. S. Kondrashov. 2004. Positive selection at sites of multiple amino acid replacements since rat-mouse divergence. Nature **429**:558-562.

Benner, S. A., M. A. Cohen, and G. H. Gonnet. 1994. Amino acid substitution during functionally constrained divergent evolution of protein sequences. Protein Eng **7**:1323-1332.

Bielawski, J. P., and Z. Yang. 2004. A maximum likelihood method for detecting functional divergence at individual codon sites, with application to gene family evolution. J Mol Evol **59**:121-132.

Bielawski, J. P., and Z. Yang. 2003. Maximum likelihood methods for detecting adaptive evolution after gene duplication. J Struct Funct Genomics **3**:201-212.

Blanquart, S., and N. Lartillot. 2008. A site- and time-heterogeneous model of amino acid replacement. Mol Biol Evol **25**:842-858.

Bofkin, L., and N. Goldman. 2007. Variation in evolutionary processes at different codon positions. Mol Biol Evol **24**:513-521.

Bollback, J. P. 2006. SIMMAP: stochastic character mapping of discrete traits on phylogenies. BMC Bioinformatics **7**:88.
Bradley, R. K., and I. Holmes. 2007. Transducers: an emerging probabilistic framework for modeling indels on trees. Bioinformatics **23**:3258-3262.
Carvajal-Rodriguez, A., D. Posada, M. Perez-Losada, E. Keller, E. J. Abrams, R. P. Viscidi, and K. A. Crandall. 2008. Disease progression and evolution of the HIV-1 env gene in 24 infected infants. Infect Genet Evol **8**:110-120.
Chamary, J. V., J. L. Parmley, and L. D. Hurst. 2006. Hearing silence: non-neutral evolution at synonymous sites in mammals. Nat Rev Genet **7**:98-108.
Chang, B. S. 2003. Ancestral gene reconstruction and synthesis of ancient rhodopsins in the laboratory. Integr. Comp. Biol. **43**:500-507.
Chen, L., and C. Lee. 2006. Distinguishing HIV-1 drug resistance, accessory, and viral fitness mutations using conditional selection pressure analysis of treated versus untreated patient samples. Biol Direct **1**:14.
Choi, S. C., A. Hobolth, D. M. Robinson, H. Kishino, and J. L. Thorne. 2007. Quantifying the impact of protein tertiary structure on molecular evolution. Mol Biol Evol **24**:1769-1782.
Christensen, O. F., A. Hobolth, and J. L. Jensen. 2005. Pseudo-likelihood analysis of codon substitution models with neighbor-dependent rates. J Comput Biol **12**:1166-1182.
Clark, A. G., S. Glanowski, R. Nielsen, P. D. Thomas, A. Kejariwal, M. A. Todd, D. M. Tanenbaum, D. Civello, F. Lu, B. Murphy, S. Ferriera, G. Wang, X. Zheng, T. J. White, J. J. Sninsky, M. D. Adams, and M. Cargill. 2003. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. Science **302**:1960-1963.
Cox, D. R., and H. D. Miller. 1977. The theory of stochastic
processes. Chapman & Hall, London.
Crandall, K. A., and D. M. Hillis. 1997. Rhodopsin evolution in the dark. Nature **387**:667-668.
Dimmic, M. W., M. J. Hubisz, C. D. Bustamante, and R. Nielsen. 2005. Detecting coevolving amino acid sites using Bayesian mutational mapping. Bioinformatics **21 Suppl 1**:i126-135.
Doron-Faigenboim, A., and T. Pupko. 2007. A combined empirical and mechanistic codon model. Mol Biol Evol **24**:388-397.
Drummond, A. J., S. Y. Ho, M. J. Phillips, and A. Rambaut. 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol **4**:e88.
Drummond, A. J., and A. Rambaut. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol **7**:214.
Durbin, R., S. R. Eddy, A. Krogh, and G. Mitchison. 1998. Biological sequence analysis. Cambridge University press, Cambridge.
Duret, L. 2002. Evolution of synonymous codon usage in metazoans. Curr Opin Genet Dev **12**:640-649.
Felsenstein, J. 2004. Inferring phylogenies. Sinauer Associates, Sunderland Massachusetts.
Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol **17**:368-376.
Fitch, W. M. 1971 Rate of change of concomitantly variable codons. J. Mol. Evol. **1**:84–96.
Forsberg, R., and F. B. Christiansen. 2003. A codon-based model of host-specific selection in parasites, with an application to the influenza A virus. Mol Biol Evol **20**:1252-1259.
Galtier, N., N. Tourasse, and M. Gouy. 1999. A nonhyperthermophilic common ancestor to extant life forms. Science **283**:220-221.
Gillespie, J. H. 1991. The causes of molecular evolution. Oxford University Press, Oxford.
Goldman, N. 1993. Statistical tests of models of DNA substitution. J Mol Evol **36**:182-198.
Goldman, N., and Z. Yang. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol **11**:725-736.
Gonnet, G. H., M. T. Hallett, C. Korostensky, and L. Bernardin. 2000. Darwin v. 2.0: an interpreted computer language for the biosciences. Bioinformatics **16**:101-103.

Grantham, R. 1974. Amino acid difference formula to help explain protein evolution. Science **185**:862-864.
Gu, X., Y. X. Fu, and W. H. Li. 1995. Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. Mol Biol Evol **12**:546-557.
Guindon, S., A. G. Rodrigo, K. A. Dyer, and J. P. Huelsenbeck. 2004. Modeling the site-specific variation of selection patterns along lineages. Proc Natl Acad Sci U S A **101**:12957-12962.
Hall, B. G. 2008. Simulating DNA coding sequence evolution with EvolveAGene 3. Mol Biol Evol **25**:688-695.
Hall, B. G. 2005. Comparison of the accuracies of several phylogenetic methods using protein and DNA sequences. Mol Biol Evol **22**:792-802.
Halpern, A. L., and W. J. Bruno. 1998. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. Mol Biol Evol **15**:910-917.
Hastings, W. 1970. Monte Carlo sampling methods using Markov chains and their application. Biometrika **57**:97-109.
Higgs, P. G., W. Hao, and G. B. Golding. 2007. Identification of conflicting selective effects on highly expressed genes. Evol. Bioinformatics **2**:1-13.
Hobolth, A., R. Nielsen, Y. Wang, F. Wu, and S. D. Tanksley. 2006. CpG + CpNpG analysis of protein-coding sequences from tomato. Mol Biol Evol **23**:1318-1323.
Holmes, I., and G. M. Rubin. 2002. An expectation maximization algorithm for training hidden substitution models. J Mol Biol **317**:753-764.
Huelsenbeck, J. P., and K. A. Dyer. 2004. Bayesian estimation of positively selected sites. J Mol Evol **58**:661-672.
Huelsenbeck, J. P., S. Jain, S. W. Frost, and S. L. Pond. 2006. A Dirichlet process model for detecting positive selection in protein-coding DNA sequences. Proc Natl Acad Sci U S A **103**:6263-6268.
Huelsenbeck, J. P., and F. Ronquist. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics **17**:754-755.
Hult, E. F., C. J. Weadick, B. S. Chang, and S. S. Tobe. 2008. Reconstruction of ancestral FGLamide-type insect allatostatins: A novel approach to the study of allatostatin function and evolution. J Insect Physiol **54**:959-968.
Huttley, G. A. 2004. Modeling the impact of DNA methylation on the evolution of BRCA1 in mammals. Mol Biol Evol **21**:1760-1768.
Jeffreys, H. 1935. Some tests of significance, treated by the theory of probability. Proc. Camb. Phil. Soc. **31**:203-222.
Jensen, J. L., and A. K. Pedersen. 2000. Probabilistic models of DNA sequence evolution with context dependent rates of substitution. Adv. Appl. Probab. **32**:499-517.
Jobb, G., A. von Haeseler, and K. Strimmer. 2004. TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. BMC Evol Biol **4**:18.
Jukes, T. H., and J. L. King. 1979. Evolutionary nucleotide replacements in DNA. Nature **281**:605-606.
Keilson, J. 1979. Markov chain models: rarity and exponentiality. Springer-Verlag, New York.
Kimchi-Sarfaty, C., J. M. Oh, I. W. Kim, Z. E. Sauna, A. M. Calcagno, S. V. Ambudkar, and M. M. Gottesman. 2007. A "silent" polymorphism in the MDR1 gene changes substrate specificity. Science **315**:525-528.
Kimura, M. 1977. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. Nature **267**:275-276.
Kimura, M. 1962. On the probability of fixation of mutant genes in a population. Genetics **47**:713-719.
Kishino, H., J. L. Thorne, and W. J. Bruno. 2001. Performance of a divergence time estimation method under a probabilistic model of rate evolution. Mol Biol Evol **18**:352-361.
Klosterman, P. S., A. V. Uzilov, Y. R. Bendana, R. K. Bradley, S. Chao, C. Kosiol, N. Goldman, and I. Holmes. 2006. XRate: a fast prototyping, training and annotation tool for phylo-grammars. BMC Bioinformatics **7**:428.

Komar, A. A. 2007. Genetics. SNPs, silent but not invisible. Science **315**:466-467.
Kosakovsky Pond, S. L., and S. D. Frost. 2005a. Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol Biol Evol **22**:1208-1222.
Kosakovsky Pond, S. L., and S. D. Frost. 2005b. Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. Bioinformatics **21**:2531-2533.
Kosakovsky Pond, S. L., and S. D. Frost. 2005c. A genetic algorithm approach to detecting lineage-specific variation in selection pressure. Mol Biol Evol **22**:478-485.
Kosakovsky Pond, S. L., S. D. Frost, Z. Grossman, M. B. Gravenor, D. D. Richman, and A. J. Brown. 2006a. Adaptation to different human populations by HIV-1 revealed by codon-based analyses. PLoS Comput Biol **2**:e62.
Kosakovsky Pond, S. L., S. D. Frost, and S. V. Muse. 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics **21**:676-679.
Kosakovsky Pond, S. L., and S. V. Muse. 2005. Site-to-site variation of synonymous substitution rates. Mol Biol Evol **22**:2375-2385.
Kosakovsky Pond, S. L., A. F. Poon, S. Zarate, D. M. Smith, S. J. Little, S. K. Pillai, R. J. Ellis, J. K. Wong, A. J. Leigh Brown, D. D. Richman, and S. D. Frost. 2008. Estimating selection pressures on HIV-1 using phylogenetic likelihood models. Stat Med.
Kosakovsky Pond, S. L., D. Posada, M. B. Gravenor, C. H. Woelk, and S. D. Frost. 2006b. GARD: a genetic algorithm for recombination detection. Bioinformatics **22**:3096-3098.
Kosakovsky Pond, S. L., D. Posada, M. B. Gravenor, C. H. Woelk, and S. D. Frost. 2006c. Automated phylogenetic detection of recombination using a genetic algorithm. Mol Biol Evol **23**:1891-1901.
Koshi, J. M., and R. A. Goldstein. 1997. Mutation matrices and physical-chemical properties: correlations and implications. Proteins **27**:336-344.
Koshi, J. M., and R. A. Goldstein. 1995. Context-dependent optimal substitution matrices. Protein Eng **8**:641-645.
Kosiol, C. 2006. Markov models of protein sequence evolution. University of Cambridge, Cambridge.
Kosiol, C., and N. Goldman. 2005. Different versions of the Dayhoff rate matrix. Mol Biol Evol **22**:193-199.
Kosiol, C., N. Goldman, and N. H. Buttimore. 2004. A new criterion and method for amino acid classification. J Theor Biol **228**:97-106.
Kosiol, C., I. Holmes, and N. Goldman. 2007. An empirical codon model for protein sequence evolution. Mol Biol Evol **24**:1464-1479.
Kosiol, C., T. Vinar, R. R. da Fonseca, M. J. Hubisz, C. D. Bustamante, R. Nielsen, and A. Siepel. 2008. Patterns of positive selection in six Mammalian genomes. PLoS Genet **4**:e1000144.
Larget, B., and D. Simon. 1999. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. Mol Biol Evol **16**:750-759.
Lartillot, N., and H. Philippe. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol **21**:1095-1109.
Lartillot, N., and H. Philippe. 2006. Computing Bayes factors using thermodynamic integration. Syst Biol **55**:195-207.
Lemey, P., I. Derdelinckx, A. Rambaut, K. Van Laethem, S. Dumont, S. Vermeulen, E. Van Wijngaerden, and A. M. Vandamme. 2005. Molecular footprint of drug-selective pressure in a human immunodeficiency virus transmission chain. J Virol **79**:11981-11989.
Lemey, P., S. L. Kosakovsky Pond, A. J. Drummond, O. G. Pybus, B. Shapiro, H. Barroso, N. Taveira, and A. Rambaut. 2007. Synonymous substitution rates predict HIV disease progression as a result of underlying replication dynamics. PLoS Comput Biol **3**:e29.
Lemey, P., S. Van Dooren, and A. M. Vandamme. 2005. Evolutionary dynamics of human retroviruses investigated through full-genome scanning. Mol Biol Evol **22**:942-951.

Lemmon, A. R., and M. C. Milinkovitch. 2002. The metapopulation genetic algorithm: An efficient solution for the problem of large phylogeny estimation. Proc Natl Acad Sci U S A **99**:10516-10521.
Li, W. H. 1997. Molecular evolution. Sinauer Associates, Sunderland Massachesetts USA.
Loytynoja, A., and N. Goldman. 2005. An algorithm for progressive multiple alignment of sequences with insertions. Proc Natl Acad Sci U S A **102**:10557-10562.
Massingham, T., and N. Goldman. 2005. Detecting amino acid sites under positive selection and purifying selection. Genetics **169**:1753-1762.
Maynard Smith, J., and N. H. Smith. 1996. Synonymous nucleotide divergence: what is "saturation"? Genetics **142**:1033-1036.
Mayrose, I., A. Doron-Faigenboim, E. Bacharach, and T. Pupko. 2007. Towards realistic codon models: among site variability and dependency of synonymous and non-synonymous rates. Bioinformatics **23**:i319-327.
Mayrose, I., N. Friedman, and T. Pupko. 2005. A gamma mixture model better accounts for among site rate heterogeneity. Bioinformatics **21 Suppl 2**:ii151-158.
McVean, G. A., and J. Vieira. 2001. Inferring parameters of mutation, selection and demography from patterns of synonymous site evolution in Drosophila. Genetics **157**:245-257.
McVean, G. A., and J. Vieira. 1999. The evolution of codon preferences in Drosophila: a maximum-likelihood approach to parameter estimation and hypothesis testing. J Mol Evol **49**:63-75.
Messier, W., and C. B. Stewart. 1997. Episodic adaptive evolution of primate lysozymes. Nature **385**:151-154.
Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. 1953. Equations of state calculations by fast computing machines. J. Chem. Physics **21**:1087-1092.
Minin, V., Z. Abdo, P. Joyce, and J. Sullivan. 2003. Performance-based selection of likelihood models for phylogeny estimation. Syst Biol **52**:674-683.
Muse, S. V., and B. S. Gaut. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol Biol Evol **11**:715-724.
Nackley, A. G., S. A. Shabalina, I. E. Tchivileva, K. Satterfield, O. Korchynskyi, S. S. Makarov, W. Maixner, and L. Diatchenko. 2006. Human catechol-O-methyltransferase haplotypes modulate protein expression by altering mRNA secondary structure. Science **314**:1930-1933.
Nielsen, R. 2002. Mapping mutations on phylogenies. Syst Biol **51**:729-739.
Nielsen, R. 2001. Mutations as missing data: inferences on the ages and distributions of nonsynonymous and synonymous mutations. Genetics **159**:401-411.
Nielsen, R., V. L. Bauer DuMont, M. J. Hubisz, and C. F. Aquadro. 2007. Maximum likelihood estimation of ancestral codon usage bias parameters in Drosophila. Mol Biol Evol **24**:228-235.
Nielsen, R., C. Bustamante, A. G. Clark, S. Glanowski, T. B. Sackton, M. J. Hubisz, A. Fledel-Alon, D. M. Tanenbaum, D. Civello, T. J. White, J. S. J, M. D. Adams, and M. Cargill. 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol **3**:e170.
Nielsen, R., and J. P. Huelsenbeck. 2002. Detecting positively selected amino acid sites using posterior predictive P-values. Pac Symp Biocomput:576-588.
Nielsen, R., and Z. Yang. 2003. Estimating the distribution of selection coefficients from phylogenetic data with applications to mitochondrial and viral DNA. Mol Biol Evol **20**:1231-1239.
Nielsen, R., and Z. Yang. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics **148**:929-936.
Pedersen, A. K., C. Wiuf, and F. B. Christiansen. 1998. A codon-based model designed to describe lentiviral evolution. Mol Biol Evol **15**:1069-1081.
Pedersen, A. M., and J. L. Jensen. 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**:763-776.

Poon, A. F., F. I. Lewis, S. L. Pond, and S. D. Frost. 2007a. An evolutionary-network model reveals stratified interactions in the V3 loop of the HIV-1 envelope. PLoS Comput Biol **3**:e231.
Poon, A. F., F. I. Lewis, S. L. Pond, and S. D. Frost. 2007b. Evolutionary interactions between N-linked glycosylation sites in the HIV-1 envelope. PLoS Comput Biol **3**:e11.
Posada, D., and T. R. Buckley. 2004. Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. Syst Biol **53**:793-808.
Posada, D., and K. A. Crandall. 2001. Selecting the best-fit model of nucleotide substitution. Syst Biol **50**:580-601.
Posada, D., and K. A. Crandall. 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics **14**:817-818.
Pupko, T., and N. Galtier. 2002. A covarion-based method for detecting molecular adaptation: application to the evolution of primate mitochondrial genomes. Proc Biol Sci **269**:1313-1316.
Rannala, B., and Z. Yang. 1996. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. J Mol Evol **43**:304-311.
Rannala, B., and Z. Yang. 2007. Inferring speciation times under an episodic molecular clock. Syst Biol **56**:453-466.
Ren, F., H. Tanaka, and Z. Yang. 2005. An empirical examination of the utility of codon-substitution models in phylogeny reconstruction. Syst Biol **54**:808-818.
Rivas, E. 2005. Evolutionary models for insertions and deletions in a probabilistic modeling framework. BMC Bioinformatics **6**:63.
Robinson, D. M., D. T. Jones, H. Kishino, N. Goldman, and J. L. Thorne. 2003. Protein evolution with dependence among codons due to tertiary structure. Mol Biol Evol **20**:1692-1704.
Rodrigue, N., H. Philippe, and N. Lartillot. 2006. Assessing site-interdependent phylogenetic models of sequence evolution. Mol Biol Evol **23**:1762-1775.
Rodrigue, N., H. Philippe, and N. Lartillot. 2008. Uniformization for sampling realizations of Markov processes: applications to Bayesian implementations of codon substitution models. Bioinformatics **24**:56-62.
Ronquist, F., and J. P. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics **19**:1572-1574.
Sainudiin, R., W. S. Wong, K. Yogeeswaran, J. B. Nasrallah, Z. Yang, and R. Nielsen. 2005. Detecting site-specific physicochemical selective pressures: applications to the Class I HLA of the human major histocompatibility complex and the SRK of the plant sporophytic self-incompatibility system. J Mol Evol **60**:315-326.
Sauna, Z. E., C. Kimchi-Sarfaty, S. V. Ambudkar, and M. M. Gottesman. 2007. The sounds of silence: synonymous mutations affect function. Pharmacogenomics **8**:527-532.
Sawyer, S. L., L. I. Wu, M. Emerman, and H. S. Malik. 2005. Positive selection of primate TRIM5alpha identifies a critical species-specific retroviral restriction domain. Proc Natl Acad Sci U S A **102**:2832-2837.
Scheffler, K., D. P. Martin, and C. Seoighe. 2006. Robust inference of positive selection from recombining coding sequences. Bioinformatics **22**:2493-2499.
Scheffler, K., and C. Seoighe. 2005. A Bayesian model comparison approach to inferring positive selection. Mol Biol Evol **22**:2531-2540.
Schneider, A., G. M. Cannarozzi, and G. H. Gonnet. 2005. Empirical codon substitution matrix. BMC Bioinformatics **6**:134.
Schwarz, G. 1978. Estimating the dimension of a model. Ann. Statist. **6**:461-464.
Seo, T. K., and H. Kishino. 2008. Synonymous substitutions substantially improve evolutionary inference from highly diverged proteins. Syst Biol **57**:367-377.

Seo, T. K., H. Kishino, and J. L. Thorne. 2004. Estimating absolute rates of synonymous and nonsynonymous nucleotide substitution in order to characterize natural selection and date species divergences. Mol Biol Evol **21**:1201-1213.
Seoighe, C., F. Ketwaroo, V. Pillay, K. Scheffler, N. Wood, R. Duffet, M. Zvelebil, N. Martinson, J. McIntyre, L. Morris, and W. Hide. 2007. A model of directional selection applied to the evolution of drug resistance in HIV-1. Mol Biol Evol **24**:1025-1031.
Shapiro, B., A. Rambaut, and A. J. Drummond. 2006. Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. Mol Biol Evol **23**:7-9.
Sharp, P. M., and W. H. Li. 1987. The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res **15**:1281-1295.
Siepel, A., and D. Haussler. 2004a. Combining phylogenetic and hidden Markov models in biosequence analysis. J Comput Biol **11**:413-428.
Siepel, A., and D. Haussler. 2004b. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. Mol Biol Evol **21**:468-488.
Stern, A., A. Doron-Faigenboim, E. Erez, E. Martz, E. Bacharach, and T. Pupko. 2007. Selecton 2007: advanced models for detecting positive and purifying selection using a Bayesian inference approach. Nucleic Acids Res **35**:W506-511.
Stern, A., and T. Pupko. 2006. An evolutionary space-time model with varying among-site dependencies. Mol Biol Evol **23**:392-400.
Stuart, A., K. Ord, and S. Arnold. 1999. Kendall's advanced theory of statistics. Arnold, London.
Studer, R., L. Duret, S. Penel, and M. Robinson-Rechavi. 2008. Pervasive positive selection on duplicated and non duplicated vertebrate protein coding genes. Genome Res.
Suchard, M. A., and B. D. Redelings. 2006. BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny. Bioinformatics **22**:2047-2048.
Sugiura, N. 1978. Further analysis of the data by Akaike's information criterion and the finite corrections. Commun. Stat. Theory Methods **A7**:13-26.
Suzuki, Y. 2004. New methods for detecting positive selection at single amino acid sites. J Mol Evol **59**:11-19.
Swanson, W. J., R. Nielsen, and Q. Yang. 2003. Pervasive adaptive evolution in mammalian fertilization proteins. Mol Biol Evol **20**:18-20.
Thorne, J. L., S. C. Choi, J. Yu, P. G. Higgs, and H. Kishino. 2007. Population genetics without intraspecific data. Mol Biol Evol **24**:1667-1677.
Ugalde, J. A., B. S. Chang, and M. V. Matz. 2004. Evolution of coral pigments recreated. Science **305**:1433.
Wasserman, L. 2000. Bayesian model selection and model averaging. J. Math. Psychol. **44**:92–107.
Weadick, C. J., and B. S. Chang. 2007. Long-wavelength sensitive visual pigments of the guppy (Poecilia reticulata): six opsins expressed in a single individual. BMC Evol Biol **7 Suppl 1**:S11.
Whelan, S. 2008. Spatial and temporal heterogeneity in nucleotide sequence evolution. Mol Biol Evol.
Whelan, S., and N. Goldman. 2004. Estimating the frequency of events that cause multiple-nucleotide changes. Genetics **167**:2027-2043.
Wilson, D. J., and G. McVean. 2006. Estimating diversifying selection and functional constraint in the presence of recombination. Genetics **172**:1411-1425.
Wong, W. S., R. Sainudiin, and R. Nielsen. 2006. Identification of physicochemical selective pressure on protein encoding nucleotide sequences. BMC Bioinformatics **7**:148.
Wong, W. S., Z. Yang, N. Goldman, and R. Nielsen. 2004. Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. Genetics **168**:1041-1051.
Wright, F. 1990. The 'effective number of codons' used in a gene. Gene **87**:23-29.
Yang, Z. 1994a. Estimating the pattern of nucleotide substitution. J Mol Evol **39**:105-111.

Yang, Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol **15**:568-573.
Yang, Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol Biol Evol **10**:1396-1401.
Yang, Z. 1994b. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol **39**:306-314.
Yang, Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol **24**:1586-1591.
Yang, Z. 1995. A space-time process model for the evolution of DNA sequences. Genetics **139**:993-1005.
Yang, Z., and J. P. Bielawski. 2000. Statistical methods for detecting molecular adaptation. Trends Ecol Evol **15**:496-503.
Yang, Z., S. Kumar, and M. Nei. 1995. A new method of inference of ancestral nucleotide and amino acid sequences. Genetics **141**:1641-1650.
Yang, Z., and R. Nielsen. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol **17**:32-43.
Yang, Z., and R. Nielsen. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol **19**:908-917.
Yang, Z., and R. Nielsen. 2008. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. Mol Biol Evol **25**:568-579.
Yang, Z., R. Nielsen, N. Goldman, and A. M. Pedersen. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics **155**:431-449.
Yang, Z., and W. J. Swanson. 2002. Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. Mol Biol Evol **19**:49-57.
Yang, Z., W. S. Wong, and R. Nielsen. 2005. Bayes empirical bayes inference of amino acid sites under positive selection. Mol Biol Evol **22**:1107-1118.
Zhai, W., M. Slatkin, and R. Nielsen. 2007. Exploring variation in the d(N)/d(S) ratio among sites and lineages using mutational mappings: applications to the influenza virus. J Mol Evol **65**:340-348.
Zhang, J., R. Nielsen, and Z. Yang. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol **22**:2472-2479.
Zwickl, D. J. 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion The University of Texas, Austin.

Table 1

Off-diagonal entries of different Q-matrices defining Markov codon substitution models

Model (code/descriptive name)	q_{ij} $(i \neq j)$	If i and j differ by	Parameters (some may not be free)	References
First Markovian codon models:				
GY (original)	$\kappa\pi_j e^{(-d_{AA_iAA_j}/V)}$ $\pi_j e^{(-d_{AA_iAA_j}/V)}$ 0	1 transition 1 transversion > 1 nucleotide	V, κ, $\{\pi_j\}$, $j=1,...,61$	Goldman and Yang (1994)
GY (simplified)	$\omega\kappa\pi_j$ $\omega\pi_j$ $\kappa\pi_j$ π_j 0	1 nonsynonymous transition 1 nonsynonymous transversion 1 synonymous transition 1 synonymous transversion > 1 nucleotide	ω, κ, $\{\pi_j\}$, $j=1,...,61$	Goldman and Yang (1994)
MG	$\beta\pi_{j_n}$ $\alpha\pi_{j_n}$ 0	1 nonsynonymous substitution 1 synonymous substitution > 1 nucleotide	α, β, $\{\pi_{j_n}\}, j=1,...,61$, $j_n=\{A,T,G,C\}$	Muse and Gaut (1994)
Accounting for variability in selective pressures:				
GY	$\omega^h\kappa\pi_j$ $\omega^h\pi_j$ $\kappa\pi_j$ π_j 0	1 nonsynonymous transition 1 nonsynonymous transversion 1 synonymous transition 1 synonymous transversion > 1 nucleotide	ω^h, h may vary over site classes or branches, κ, $\{\pi_j\}$, $j=1,\ldots,61$	Yang (1998); Nielsen and Yang (1998); Yang et al. (2000)
MG×GTR	$\beta^k\theta_{i_nj_n}\pi_{j_n}$ $\alpha^h\theta_{i_nj_n}\pi_{j_n}$	1 nonsynonymous substitution 1 synonymous substitution	α^h, β^k, h and k may vary over site classes or branches, $\theta_{i_nj_n}$,$\{\pi_{j_n}\}$, $i,j=1,...,61$,	Kosakovsky Pond and Muse (2005); Kosakovsky Pond and Frost (2005)

	0	>1 nucleotide	j_n={A,T,G,C}	
Dating model	$\omega^h \kappa \pi_j u^h$ $\omega^h \pi_j u^h$ $\kappa \pi_j u^h$ $\pi_j u^h$ 0	1 nonsynonymous transition 1 nonsynonymous transversion 1 synonymous transition 1 synonymous transversion >1 nucleotide	ω^h, u^h h may vary over site classes or branches, κ, $\{\pi_j\}$, j=1,...,61	Seo et al. (2004)

Dependencies on protein secondary structure and amino acid properties:

Structure-dependent	$\kappa\omega\pi_{j_h} e^{\left(\left(E_p(i)-E_p(j)\right)f_s+\left(E_p(i)-E_p(j)\right)f_p\right)}$ $\omega\pi_{j_h} e^{\left(\left(E_p(i)-E_p(j)\right)f_s+\left(E_p(i)-E_p(j)\right)f_p\right)}$ $\kappa\pi_{j_h}$ π_{j_h} 0	1 nonsynonymous transition 1 nonsynonymous transversion 1 synonymous transition 1 synonymous transversion >1 nucleotide	ω, κ, f_s, f_p $\{\pi_{j_n}\}$, j_h={A,T,G,C}, here i and j represent whole DNA sequences	Robinson et al. (2003)
Physicochemical (amino acids are *a priori* partitioned by one physicochemical property)	$\gamma^h\omega\kappa\pi_j$ $\omega\kappa\pi_j$ $\gamma^h\omega\pi_j$ $\omega\pi_j$ $\kappa\pi_j$ π_j 0	1 nonsyn. property-altering transition 1 nonsyn. property-conserving transition 1 nonsyn. property-altering transversion 1 nonsyn. property-conserving transversion 1 synonymous transition 1 synonymous transversion >1 nucleotide	γ^h h may vary over site classes, ω, κ, $\{\pi_j\}$, j=1,...,61	Wong et al. (2006); Setting ω=1 simplifies to the model of Sainudiin et al. (2005)
Directional selection (for pre-specified amino acid Y)	$\omega_Y\kappa\pi_j$ $\omega_Y\pi_j$ $\omega\,\kappa\pi_j$ $\omega\,\pi_j$ $\kappa\pi_j$ π_j 0	1 nonsynonymous transition to Y 1 nonsynonymous transversion to Y 1 nonsyn. transition to other than Y 1 nonsyn. transversion to other than Y 1 synonymous transition 1 synonymous transversion >1 nucleotide	ω_Y, ω κ, $\{\pi_j\}$, j=1,...,61	Seoighe et al. (2007)

Empirical models and combined empirical and mechanistic models

ECM	$s_{ij}\pi_j$	any codon change	$\{s_{ij}\}$, $\{\pi_j\}$, i, j = 1,...,61	Kosiol et al. (2007)

ECM+ω +κ	$\omega\kappa(i,j)s_{ij}\pi_j$ $\kappa(i,j)s_{ij}\pi_j$	nonsynonymous substitution synonymous substitution	ω, $\{s_{ij}\}$, $\{\pi_j\}$, $i, j = 1,\ldots,61$ number and definitions of parameters describing $\kappa(i,j)$ may vary	Kosiol et al. (2007); see the original article for other ECM variants
MEC	$\omega\kappa(i,j)s_{(AAi\rightarrow AAj)}\pi_j$ $\kappa(i,j)s_{(AAi\rightarrow AAj)}\pi_j$	nonsynonymous substitution synonymous substitution	ω, $\{\pi_j\}$, $i, j = 1,\ldots,61$, $AA_i = 1,...,20$ number and definitions of parameters describing $\kappa(i,j)$ may vary	Doron-Fagenbiom and Pupko (2007)
Accounting for codon bias:				
Preferred-codon model	$q'_{ij}P_{S+}$ $q'_{ij}P_{S-}$ q'_{ij}	change from an unpreferred codon to a preferred change from a preferred codon to an unpreferred else	parameters defining $\{q'_{ij}\}$, $i, j = 1,...,61$ P_{S+}, P_{S-}	Nielsen et al. (2007)
FMutSel	$\omega^h\theta_{i_n j_n}h(S_{ij})\pi_{j_n}$ $\theta_{i_n j_n}h(S_{ij})\pi_{j_n}$ 0	1 nonsynonymous substitution 1 synonymous substitution >1 nucleotide	ω^h, $\{\theta_{i_n j_n}\}$, $\{\pi_{j_n}\}$, $\{F_j\}$, where $S_{ij}=F_j-F_i$, $i, j =1,..,61$, $n=\{1, 2, 3\}$, $j_n=\{A,T,G,C\}$	Yang and Nielsen (2008)

Note.—Definitions of model parameters:

α is the synonymous substitution rate;

β is the nonsynonymous substitution rate;

ω is the measure of selective pressure, but the exact meaning may vary across models;

ω_Γ is the measure of selective pressure on nonsynonymous substitutions towards the target amino acid Y;

ω^h is the site-class or branch specific measure of selective pressure, with h referring to a particular class-site or branch;

V is parameter representing the tendency of a gene to undergo nonsynonymous changes (negatively correlated with $1/\omega$);

$d_{AA_iAA_j}$ is the Grantham distance between AAs encoded by codons i and j (= 0 if i and j encode the same AA);

γ^h is the property-altering nonsynonymous substitutions rate relative to the background rate of nonsynonymous substitutions ω , with h referring to a class-site;

u^h is the background substitution rate, with h indicating a specific branch;

κ is the transition/transversion rate ratio;

$\kappa(i,j)$ represents the transition-transversion bias between codons i and j, and can be formulated in several ways; a possible definition is that $\kappa(i,j)= \kappa^x$, where κ is the transition (or transversion) bias and x is the number of nucleotide transitions (or transversions) between i and j;

π_{j_n} is the equilibrium frequency of a target nucleotide j_n in codon j at codon position n (described as mutation bias in model FMutSel);

π_j is the equilibrium frequency of codon j (often estimated from data using the observed nucleotides frequencies at three codon positions, i.e., F3×4 model); in semi-parametric models these can be estimated individually for each gene;

$\theta_{i_n j_n}$ is the exchangeability rate between nucleotides i_n and n_j (also parameters of the DNA-based GTR model);

s_{ij} is the empirical exchangeability rate between codons i and j;

$s_{(AAi \to AAj)}$ is the empirical exchangeability rate between amino acids AA_i and AA_j, coding for codons i and j respectively;

P_{S+} is the ratio of the fixation probability from unpreferred to preferred codon and the fixation probability of a neutral mutation: $P_{S+} = S/(1–e^{–S})$, where $S = 2Ns$, N is the effective population size, and s is the selective coefficient ($0<s<<1$);

P_{S-} is the ratio of the fixation probability from preferred to unpreferred codon and the fixation probability of a neutral mutation: $P_{S-} = –S/(1–e^{S})$, where $S = 2Ns$, N is the effective population size, and s is the selective coefficient ($0<s<<1$);

$h(S_{ij})$ is the ratio of the fixation probability of a substitution from i to j and the fixation probability of a neutral mutation: $h(S_{ij})= S_{ij}/(1–e^{–Sij})$, where $S_{ij} = 2Ns_{ij}$, N is the effective population size, and s_{ij} is the selective coefficient of i to j change ($|s_{ij}|<<1$);

F_j is the population-scaled fitness of codon j, and so the population-scaled selection coefficient of an i to j change is $S_{ij} = F_j – F_i$;

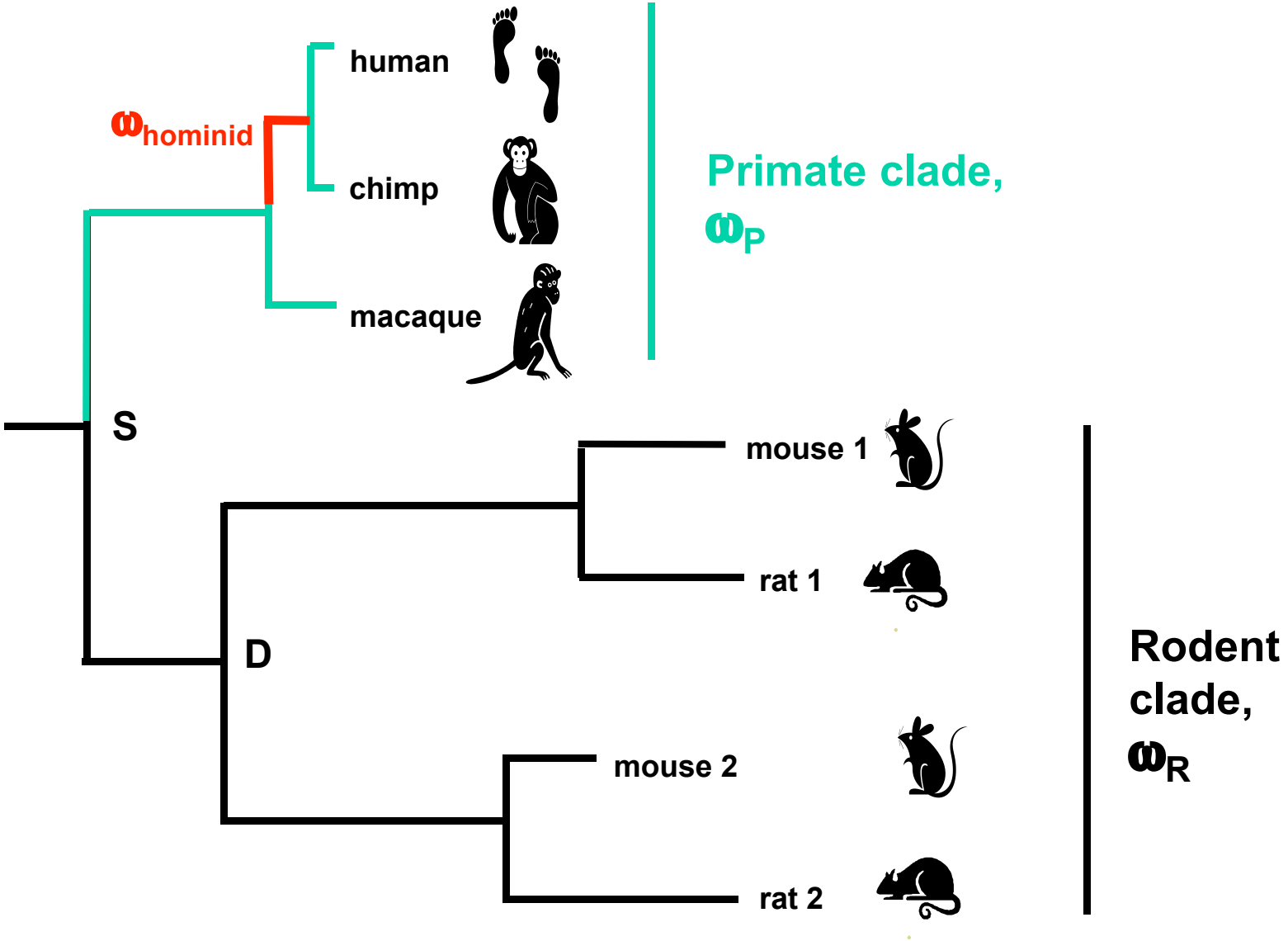
q'_{ij} is a transition rate from codon i to codon j, which may be defined as in any existing (e.g., MG or GY) or modified codon model;

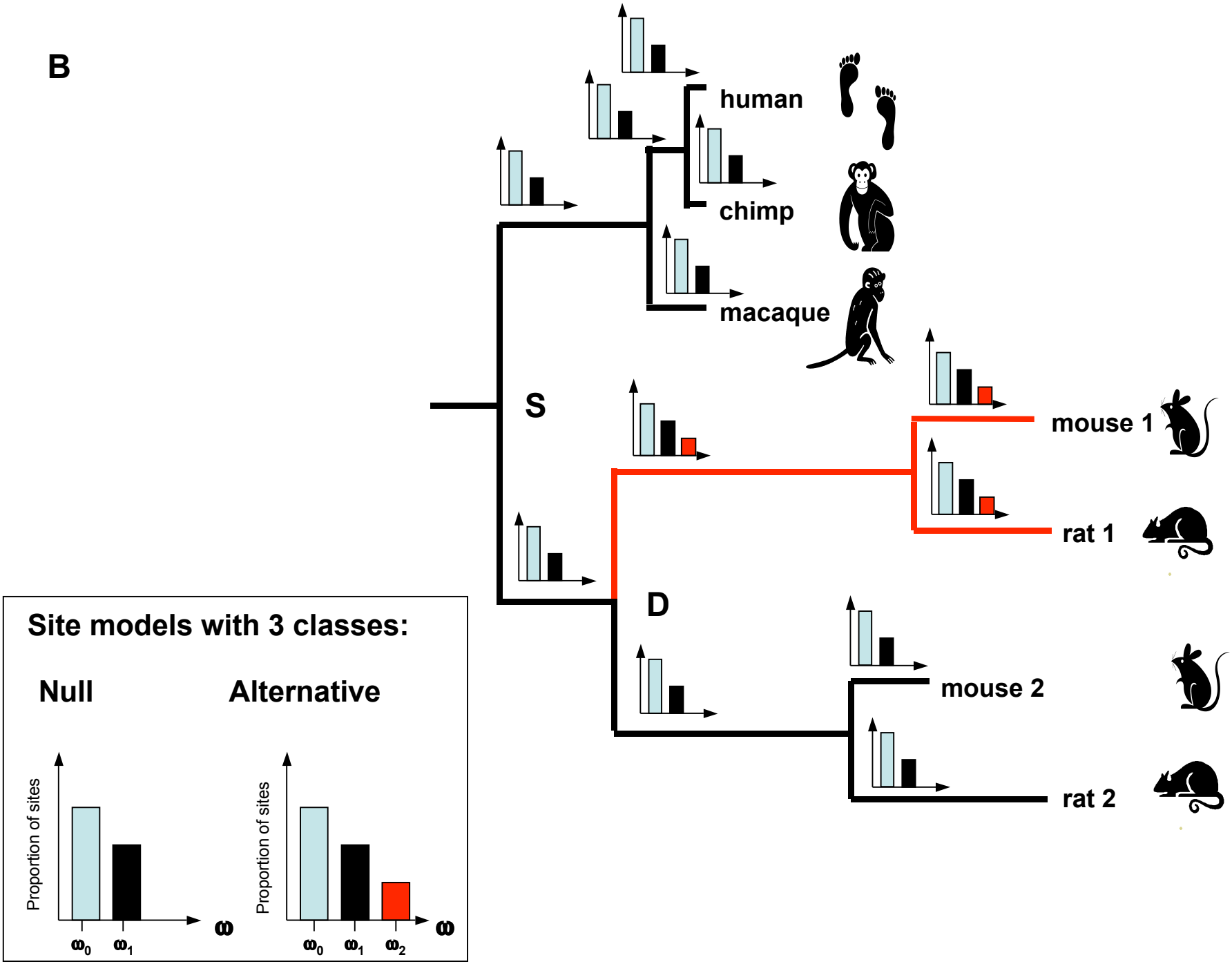
$E_s(i)$ is the solvent accessibility measure of sequence-structure compatibility for sequence i;

$E_p(i)$ is the paiwise measure of sequence-structure compatibility for sequence i;

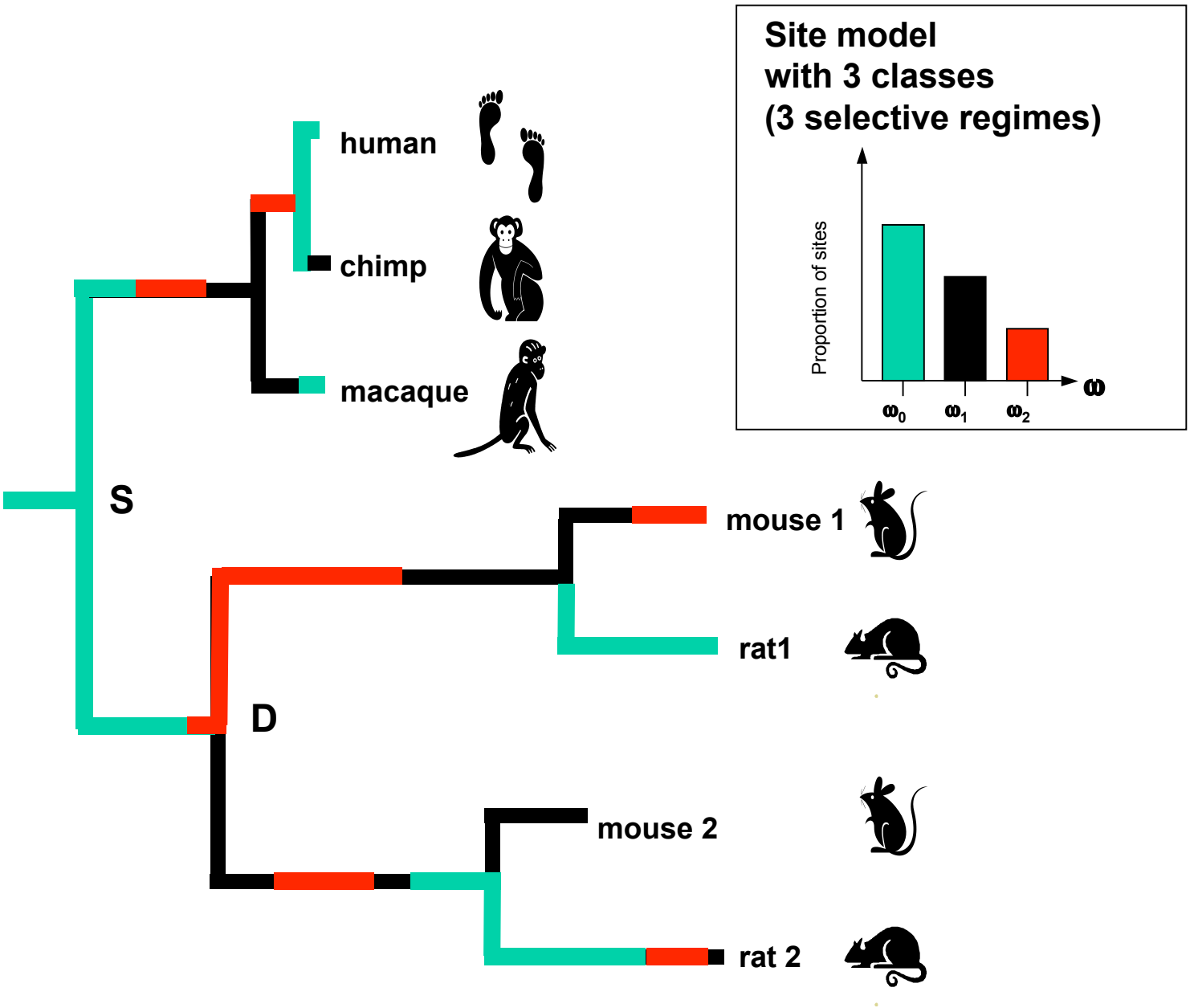
f_s, f_p are parameters reflecting contributions of nonsynonymous rates coming from sequence-structure fit

A


human
ω hominid
chimp
macaque
Primate clade,
ω P
S
mouse 1
rat 1
D
mouse 2
rat 2
Rodent
clade,
ω R


B
human
chimp
macaque
S
D
mouse 1
rat 1
mouse 2
rat 2
Site models with 3 classes:
Null
Alternative
Proportion of sites
ω
ω0
ω1
Proportion of sites
ω
ω0
ω1
ω2

C


Site model
with 3 classes
(3 selective regimes)
Proportion of sites
ω
ω0
ω1
ω2
human
chimp
macaque
S
D
mouse 1
rat1
mouse 2
rat 2