CodonTest: Modeling Amino Acid Substitution Preferences in Coding Sequences

Wayne Delport[1], Konrad Scheffler[2], Gordon Botha[2], Mike B. Gravenor[3], Spencer V. Muse[4], Sergei L. Kosakovsky Pond[5]*

1 Department of Pathology, University of California, San Diego, La Jolla, California, United States of America, **2** Computer Science Division, Department of Mathematical Sciences, Stellenbosch University, Stellenbosch, South Africa, **3** School of Medicine, University of Swansea, Swansea, United Kingdom, **4** Department of Statistics, North Carolina State University, Raleigh, North Carolina, United States of America, **5** Department of Medicine, University of California, San Diego, La Jolla, California, United States of America

Abstract

Codon models of evolution have facilitated the interpretation of selective forces operating on genomes. These models, however, assume a single rate of non-synonymous substitution irrespective of the nature of amino acids being exchanged. Recent developments have shown that models which allow for amino acid pairs to have independent rates of substitution offer improved fit over single rate models. However, these approaches have been limited by the necessity for large alignments in their estimation. An alternative approach is to assume that substitution rates between amino acid pairs can be subdivided into K rate classes, dependent on the information content of the alignment. However, given the combinatorially large number of such models, an efficient model search strategy is needed. Here we develop a Genetic Algorithm (GA) method for the estimation of such models. A GA is used to assign amino acid substitution pairs to a series of K rate classes, where K is estimated from the alignment. Other parameters of the phylogenetic Markov model, including substitution rates, character frequencies and branch lengths are estimated using standard maximum likelihood optimization procedures. We apply the GA to empirical alignments and show improved model fit over existing models of codon evolution. Our results suggest that current models are poor approximations of protein evolution and thus gene and organism specific multi-rate models that incorporate amino acid substitution biases are preferred. We further anticipate that the clustering of amino acid substitution rates into classes will be biologically informative, such that genes with similar functions exhibit similar clustering, and hence this clustering will be useful for the evolutionary fingerprinting of genes.

Citation: Delport W, Scheffler K, Botha G, Gravenor MB, Muse SV, et al. (2010) CodonTest: Modeling Amino Acid Substitution Preferences in Coding Sequences. PLoS Comput Biol 6(8): e1000885. doi:10.1371/journal.pcbi.1000885

Editor: Wen-Hsiung Li, University of Chicago, United States of America

Received April 3, 2010; **Accepted** July 14, 2010; **Published** August 19, 2010

Funding: This research was supported by the Joint DMS/NIGMS Mathematical Biology Initiative through Grant NSF-0714991, the National Institutes of Health (AI47745), and by a University of California, San Diego Center for AIDS Research/NIAID Developmental Award to WD and SLKP (AI36214). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: spond@ucsd.edu

Introduction

Modern molecular evolution has benefited greatly from the development of a sound probabilistic framework for modeling the evolution of homologous gene sequences [1]. In particular, codon substitution models [2,3] have facilitated the estimation of the ratio of non-synonymous to synonymous substitution rates (referred to as dN/dS, K_a/K_s, ω), which can be interpreted as an indicator of the strength and type of natural selection (see [4] or [5] for recent reviews). Codon models are fundamentally mechanistic because they use the structure of the genetic code to partition codon substitutions into classes. Initially, and in most subsequent applications of codon models, all one-nucleotide substitutions were stratified into synonymous (rate α, using the notation of [2]) and non-synonymous (rate β) classes. Despite several early attempts, e.g. [3], none of the widely-adopted codon models incorporated physicochemical properties of the two residues being exchanged. In contrast, most protein substitution models are derived by estimating the relative rates of amino-acid substitutions in large protein databases [6–8], and consistently report dramatic differences in the relative replacement rates of different residues.

The persisting dissonance between how codon and protein models approach amino acid substitution rates has fostered multiple recent efforts to develop what we will call *multi-rate* codon models (or more accurately, multi- nonsynonymous rate models), in contrast to the existing single-rate model. These models divide amino acid pairs (or codon pairs) into multiple rate categories, such that every category has its own rate which governs substitutions between the pairs in that category. In the most extreme case, every amino acid or codon pair belongs to a different category and thus has its own rate – potentially leading to a very large number of parameters that need to be estimated. Several strategies have been proposed for limiting the number of parameters in multi-rate models.

Doron-Faigenboim et al. [9] proposed to overlay existing empirically derived amino acid substitution matrices (e.g. [7] or [8]) onto single-rate codon models by weighted partitioning of the empirical rate of substitution between two protein residues. Kosiol, Holmes & Goldman [10] directly estimated all 1,830 codon-to-codon substitution rates in an empirical codon model – a codon equivalent of the nucleotide GTR model [11], assuming the universal genetic code. However, this effort required a truly

Author Summary

Evolution in protein-coding DNA sequences can be modeled at three levels: nucleotides, amino acids or codons that encode the amino acids. Codon models incorporate nucleotide and amino acid information, and allow the estimation of the rate at which amino acids are replaced (dN) versus the rate at which they are preserved (dS). The dN/dS ratio has been used in thousands of studies to detect molecular footprints of natural selection. A serious limitation of most codon models is the unrealistic assumption that all non-synonymous substitutions occur at the same rate. Indeed, amino acid models have consistently demonstrated that different residues are exchanged more or less frequently, depending on incompletely understood factors. We derive and validate a computational approach for inferring codon models which combine the power to investigate natural selection with data-driven amino acid substitution biases from alignments. The addition of amino acid properties can lead to more powerful and accurate methods for studying natural selection and the evolutionary history of protein-coding sequences. The pattern of amino acid substitutions specific to a given alignment can be used to compare and contrast the evolutionary properties of different genes, providing an evolutionary analog to protein family comparisons.

massive training dataset encompassing alignments from 7,332 protein families of the Pandit database [12]. The resulting empirical codon model (ECM) encodes evolution patterns averaged over many proteins. However, no single empirically-derived substitution rate matrix appears to be generalizable across multiple genes and taxonomic groups, as evidenced by a plethora of specialized substitution models, e.g. for mammalian mitochondrial genomes [13], plant chloroplast genes [14], viral reverse transcriptases [15] or HIV-1 genes [16].

More mechanistic parameters can be introduced to improve biological realism of codon-models. The linear combination of amino acid properties (LCAP) model [17] expresses exchangeability of a pair of codons as an (exponentiated) linear combination of differences in five independently validated amino acid physicochemical properties. This parameterization incorporates weighting (or importance) coefficients inferred from the data to allow for differences in protein evolution between genes, shown to be significant and biologically meaningful in yeast proteins [18], and once again underscoring the utility of gene-specific evolutionary models.

All multi-rate codon models published to date have shown clear improvements in model fit over the single-rate model. However, multi-rate models in which substitutions were randomly assigned to classes easily outperform the single-rate model [19] and thus it is a poor performance benchmark. At the other extreme of model space is the full time-reversible codon model, with 1,830 parameters (or 526, if only single nucleotide substitutions are modeled), which will certainly suffer from massive over-fitting on single gene alignments. Over-parameterization can be reduced by "smoothing", i.e. by grouping the rates into exchangeability classes based on the physicochemical properties of amino acids [20]. However, without a rigorous model selection framework, it is difficult to ascertain how well any particular smoothing approach fits the data. To appreciate how large the space of potential models is, consider that there are are approximately 2×10^{22} possible multi-rate codon models with $K=2$ nonsynonymous rate classes, and approximately 2×10^{50} possible models for $K=5$. Given such a large search space it is impossible to evaluate even a small fraction of possible models exhaustively, and one cannot presume that any given model or a small set of models are sufficiently representative without exploring the alternatives.

Huelsenbeck et al. [21] examined a Bayesian approach to estimate empirical amino acid substitution models in which amino acid exchangeability classes are assigned using a Dirichlet process. However, a prior distribution needs to be specified for the number of classes ($K=2$, 5, or 10), and mechanistic features of codon evolution are excluded. Models which combine empirical codon models and mechanistic parameters, such as β/α and transition-transversion bias [10], have been shown to outperform the models which include only a single effect. This evidence highlights the necessity to model both mutational effects, which result in substitution preferences for particular amino acids, and selective effects, the result of fitness differences of alternate phenotypes. In this manuscript, we present an information-theoretic model selection procedure that extends the concept of ModelTest [22], formulated for nucleotide model selection, to codon models. Unlike ModelTest, which examines 56 *a priori* defined models, we use a Genetic Algorithm (GA) to search the combinatorially large set of codon models (i.e. select the number of rate classes), to assign amino acid substitution rates to these classes, infer rate parameters and, finally, report a set of credible models given the data. Our group has successfully applied GAs to a variety of problems in evolutionary biology, including inference of lineage-specific selective regimes [23], detecting recombination in homologous sequence alignments [24], and model selection for paired RNA sequences [25], where the GA was able to recover biologically relevant properties and outperformed all known mechanistic models.

Using simulated data, we demonstrate that GA model selection (under a sufficiently stringent model selection criterion) is not susceptible to over-fitting, and that codon alignments of typical size contains sufficient signal to reliably allocate non-synonymous substitutions into a small number of rate classes, typically $2-8$. On empirical data sets, GA-selected codon substitution models consistently outperformed published empirical and mechanistic models. In addition to selecting a single best fitting model, the GA also estimates a set of credible models for an alignment. A weighted combination of models in the credible set enable model averaged phylogenetic [26] and substitution rate matrix [25] inference and further reduces the risk of over-fitting. We anticipate that improvements in model realism will translate into improved sequence alignment, phylogeny estimation, and selection detection. Moreover, we hypothesize that the clustering of non-synonymous substitution rates into groups with the same rate parameter is shared by genes with similar biological and structural properties, and hence this clustering is informative for improving evolutionary fingerprinting of genes [27].

Methods

Model definition

Models considered in this paper assume that codon substitutions along a branch in a phylogenetic tree can be described by an appropriately parameterized continuous-time homogeneous and stationary Markov process; an assumption ubiquitous in codon-evolution literature. The substitution process is uniquely defined by the rate matrix, Q, whose elements q_{ij} denote the instantaneous substitution rate from codon i to codon j. Using A_i to label the amino-acid encoded by codon i, and assuming a universal genetic code with three stop codons (other codes can be handled with

obvious modifications), matrix Q comprises 61×61 such elements, where

$$q_{ij} = \begin{cases} r(A_i,A_j)\theta_{ij}\pi_{ij}, & i \neq j, \text{ and } i \rightarrow j \text{ involves} \\ & \text{one nucleotide substitution,} \\ 0, & i \neq j \text{ and } i \rightarrow j \text{ involves two or} \\ & \text{three nucleotide substitutions,} \\ -\sum_{k \neq i} q_{ik}, & i = j. \end{cases} \quad (1)$$

Here, π_{ij} denote equilibrium frequency parameters, θ_{ij} denote nucleotide mutational biases, and $r(A_i, A_j) = r(A_j, A_i)$ denote the substitution rates between amino acids encoded by codons i and j. How to infer $r(A_i, A_j)$ is the primary focus of this paper. We consider two different parameterizations of π_{ij}: the GY parameterization [3], where π_{ij} is the equilibrium frequency of the target *codon*, and the MG parameterization [2], where $\pi_{ij} = \phi_a^p$ is a *nucleotide* frequency parameter for the position that is being substituted ($p = 1, 2, 3$; $a = A, C, G, T$). For the GY parameterization, we estimate codon equilibrium frequencies by their proportions in the data (the F61 estimator, 60 parameters for the universal genetic code). For the MG parameterization, we estimate the nine frequency parameters by maximum likelihood [28]. The equilibrium frequency of codon xyz can then be computed as

$$\pi_{xyz} = \frac{\phi_x^1 \phi_y^2 \phi_z^3}{1 - \phi_X},$$

where $X = \{TAA, TAG, TGA\}$ and $\phi_X = \sum_{xyz \in X} \phi_x^1 \phi_y^2 \phi_z^3$.

Finally, we set $\theta_{ij} = \theta_{ji}$, $\theta_{AG} = 1$ and estimate 5 other rates (θ_{AC}, θ_{AT}, θ_{CG}, θ_{CT}, θ_{GT}) by maximum likelihood; this parameterization follows the MG94×REV model from [29].

Inferring non-synonymous substitution rates

By varying the parametric complexity of the non-synonymous substitution rate $r(A_i, A_j)$ encoding in equation (1), we can span the range of models from the single rate model (SR, current default standard, 1 non-synonymous rate parameter), to the general codon time-reversible model (REV) with each amino-acid pair substitution exchanged at its own rate. Only 75 out of 190 total amino-acid pairs can be exchanged via a single nucleotide substitution, for example $F(TT\,R)$ and $L(TTY)$ are one such pair, but $A(GC\,N)$ and $H(C\,AY)$ are not. Consequently, the REV model has 75 non-synonymous rate parameters. The purpose of our study is to explore the model space between these two extremes, taking into account the limitations of information content in single gene alignments. Note that most existing multi-rate models can be represented with an appropriate choice of $r(A_i, A_j)$ in equation (1). Empirical models (e.g. ECM) replace $r(A_i, A_j)$ with numerical values estimated from large training data sets, whereas mechanistic models (e.g. LCAP) assume that rates can be modeled via a function measuring differences/similarities in physicochemical properties of residues (Table 1).

We focus on structured (or rate clustering) models: those which assume that substitution rates can be partitioned/structured into K classes, where each class has a single estimated rate parameter. These structured models may be defined using amino acid similarity classes [30], but instead of adopting *a priori* classes of rates, we propose to *infer* their number and identity from the data. A structured model with N substitutions (e.g. $N = 75$ for the Universal genetic code) in K classes can be represented as a vector M of length N, where each element is an integer between 1 and K labeling the class. For example if the vector entries corresponding to $I \leftrightarrow L$, $L \leftrightarrow V$ and $S \leftrightarrow W$ substitutions have values 1, 1 and 3, then $r(I, L) = r(L, V) = C_1$ and $r(S, W) = C_3$. As an analogy, the HKY85 nucleotide model [31] is a structured model with vector, $M_{HKY85} = (0_{AC}, 1_{AG}, 0_{AT}, 0_{CG}, 1_{CT}, 0_{GT})$, where the substitutions between 6 nucleotide pairs (indicated by a subscript) are placed into transition (1) and transversion (0) classes. Given the structure of a codon model, e.g. $(0_{LI}, 1_{LH}, 0_{LV}, 1_{LS}, 2_{LF}, \ldots 3_{RW})$, it can be fitted to the data using standard maximum likelihood phylogenetic algorithms, e.g. as implemented in HyPhy [32]. The resulting set of rate estimates $\hat{C}_1, \ldots \hat{C}_K$ instantiate a structured model and induce a corresponding empirical model, e.g. (0.25, 0.35, 0.25, 0.35, 0.8, ...1.5).

Because the space of structured codon models is combinatorially large, we utilize a GA previously used to solve an analogous model selection problem for paired RNA data [25]. Parameter space is defined by two components: a discrete component which assigns pairwise non-synonymous substitutions between codons to K rate classes using the structured vector described above, and a

Table 1. Various approaches to estimating residue-dependent non-synonymous substitution rates.

Model	$r(A, B)$	p	Description
Single rate	C	1	
Random – X	$C_{rand(1, X)}$	X	Rates randomly assigned to X classes
ECM	c_{ij}	0	Codon level rates c_{ij} are inferred from a large training data set
ECM+ω	ωc_{ij}	1	Codon level rates c_{ij} are inferred from a large training data set
			Correction parameter ω inferred from the data
LCAP	$\exp\left[\sum_{i=1}^{5} C_i \Delta_i(A,B)\right]$	5	Based on a weighted combination of 5 physico-chemical distances Δ_i
GA - X	$C_{g(A, B)}$	X	X and $g(A, B) \rightarrow 0 \ldots X-1$ are inferred by the GA
REV	$C_{A\,B}$	75	Each unique residue pair within one nucleotide substitution has its own rate

p = number of model parameters estimated from the data. C denotes rates that are estimated by maximum likelihood by the data and c – those that are estimated in other ways.

doi:10.1371/journal.pcbi.1000885.t001

continuous component comprising a vector of branch lengths, nucleotide substitution rates, frequency parameters and non-synonymous rates $C_1, \ldots C_K$. The discrete component is optimized by the GA, while the continuous component is estimated using numerical non-linear optimization procedures, given the structure of the model. We initially approximate branch lengths using the SR model and update them whenever the GA iteration improves the fitness score by more than 50 *mBIC* points (see below) as compared to the most recent model for which branch lengths have been estimated. Further details of the genetic algorithm are described in detail in [25], and for the sake of brevity we do not present it here.

We are left with the problem of inferring the number of rate classes K. This is done by starting with $K=1$ and iteratively proposing to increment K. For each proposal, the model with $K+1$ rate classes is optimized using the optimized K-class model as initialization. If the proposal results in a model with a better fitness value (see below), it is accepted and a new proposal generated. The process terminates when the $K+1$-class proposal does not beat the K-class model.

We initially assigned a fitness value to each model using $BIC=-2\log L+p\log s$ where s is the sample size and p is the number of parameters in the model [33]. The "sample size" of a sequence alignment is difficult to quantify with a single number, since it depends on both the number of sequences in the alignment and the lengths of those sequences. We use the number of characters to approximate "sample size" to make the model selection criterion maximally conservative. While it is straightforward to count the number of estimated parameters in any given structured model, setting p to that number leads to model overfitting (results not shown), because the topological component (the assignment of rates to classes) adds further "degrees of freedom" to the model. To determine the appropriate penalty term, we conducted simulations; there is precedent for this in statistical literature on generalized information criteria (e.g. [34]). We removed the effect of phylogeny by simulating nine sets of two-sequence alignments (0.2 divergence): each set of simulations consisted of 100 replicates with between 10^4 and 10^6 codons (in 10^4 increments). The sets had 1 to 5 rate classes (Figure 1), representing rate classification problems that ranged from easy (large numerical differences between class rates, e.g. 0.25 and 1.0) to difficult (small numerical differences, e.g. 0.25 and 0.3). We constructed generating multi-rate models by assigning rates to K bins randomly with equal probability. For each simulation set we plotted the difference in log likelihood (scaled by the sample size = log of characters) between the correct model (K rates), and models with $K-1$ and $K+1$ rates, respectively. Simulations indicated that doubling the number of parameters in the BIC penalty term ensured sufficient power, and controlled false positives for all simulation sets (Figure 1). We used this modified BIC, $mBIC=-2\log L+2p\log s$ to assign fitness to every model examined by a GA run and select those with the lowest *mBIC*.

Simulated data analysis

We also simulated realistic "gene-size" alignments on 16 and 32 taxon trees. Nucleotide frequencies were uniform (0.25) for each position, and the nucleotide bias component was set to HKY85 with transition/transversion ratio, $\kappa=4$. We generated 100 data sets for each K:rate vector combination, under the single rate, and a fixed Random-K model (Table 2). These data allowed us to assess the performance of the model when the true underlying model was known.

For each simulation scenario, we report the proportion of replicates P_m for which the GA inferred the correct number of rate classes K, the proportion of underfitted replicates P_u (too few rate classes were inferred) and the proportion of overfitted replicates P_o (too many rate classes were inferred). For the replicates where the correct number of rate classes was inferred, we computed the Rand statistic (P_c, [35]) on the generating and inferred model structures to quantify the similarity between two clusterings rates. The Rand statistic quantifies the similarity between two clusterings (A & B) of the same set of N objects and can be defined as $(N_{00}+N_{11})/(N_{00}+N_{01}+N_{10}+N_{11})$, where N_{00} is the number of objects (pairs of substitution rates) that belong to different classes in both A and B, N_{01} (N_{10}) is the number of objects that belong to different (same) classes in A, but the same (different) class in B, and N_{11} is the number of objects that belong to the same class in both A and B. Clearly, $P_c=1$ for perfect agreement ($N_{11}=N$) and $P_c=0$ for perfect disagreement ($N_{00}=N$).

Empirical data analysis

We prepared a collection of reference empirical data sets (see Table 3), to be used for benchmarking GA, published and extreme-case models. The collection included three protein family alignments from Pandit [12] selected randomly from all alignments with >80 taxa, a randomly selected Yeast protein alignment [18], a group M HIV-1 *pol* alignment [36] and an Influenza A virus (IAV) *HA* alignment comprising H3N2, H5N1, H2N2 and H1N1 serotypes. The latter was assembled by random selection of 30 post-2005 sequences for each serotype from the NCBI Influenza database [37]. Finally, we examined the vertebrate rhodopsin protein, recently analyzed for molecular mechanisms of phenotypic adaptation by [38]. We inferred a structured multi-rate model for each of these data sets using the genetic algorithm and *mBIC* model fitness function defined above. A comparison of the GA-fitted model against existing models is unfair, since the former was selected among a set of candidate models using the test alignment. To confirm that GA models were generalizable, we evaluated the fit of the GA models and that of existing models for both the reference datasets, and independent test alignments for the same taxonomic groups (validation data sets). Two HIV-1 pol gene alignments were obtained for subtypes B [39] and C [40]. Subtype assignments were confirmed using the SCUEAL sub-typing tool [36], and inter- and intra-subtype recombinants were pruned from the analysis. For IAV *HA* we used independent alignments for serotypes H5N1 and H3N2, filtered from the NCBI Influenza database [37], and from [41], respectively.

We fitted five reference models to each dataset: (i) the single-rate model, (ii) a Random-3 and a Random-5 model, (iii) the empirical codon model (ECM, [10]), (iv) the Linear Combination of Amino Acid Properties (LCAP) model [17,18], and (v) the reversible (REV) model (see Table 1).

For every dataset, the corresponding GA-run was processed to obtain three different alignment-specific multi-rate models.

1. A structured GA model (GA_s): this is the best-fitting model (with value $mBIC_0$), which defines K rate clusters. The numerical values of corresponding K substitution rates are inferred using maximum likelihood. This model is a direct analog of the single "best" substitution model reported by the familiar ModelTest [22] nucleotide model selection procedure.
2. A numerical model-averaged GA model (GA_r), which is computed by weighting the numerical rate estimates from all models in the credible set using *mBIC*-based Akaike weights (as in [25]). Briefly, for the $i-$th model examined by the GA, we compute its evidence ratio versus the GA_s model as $r_i=\exp[(mBIC_0-mBIC_i)/2]$, which can be thought of as

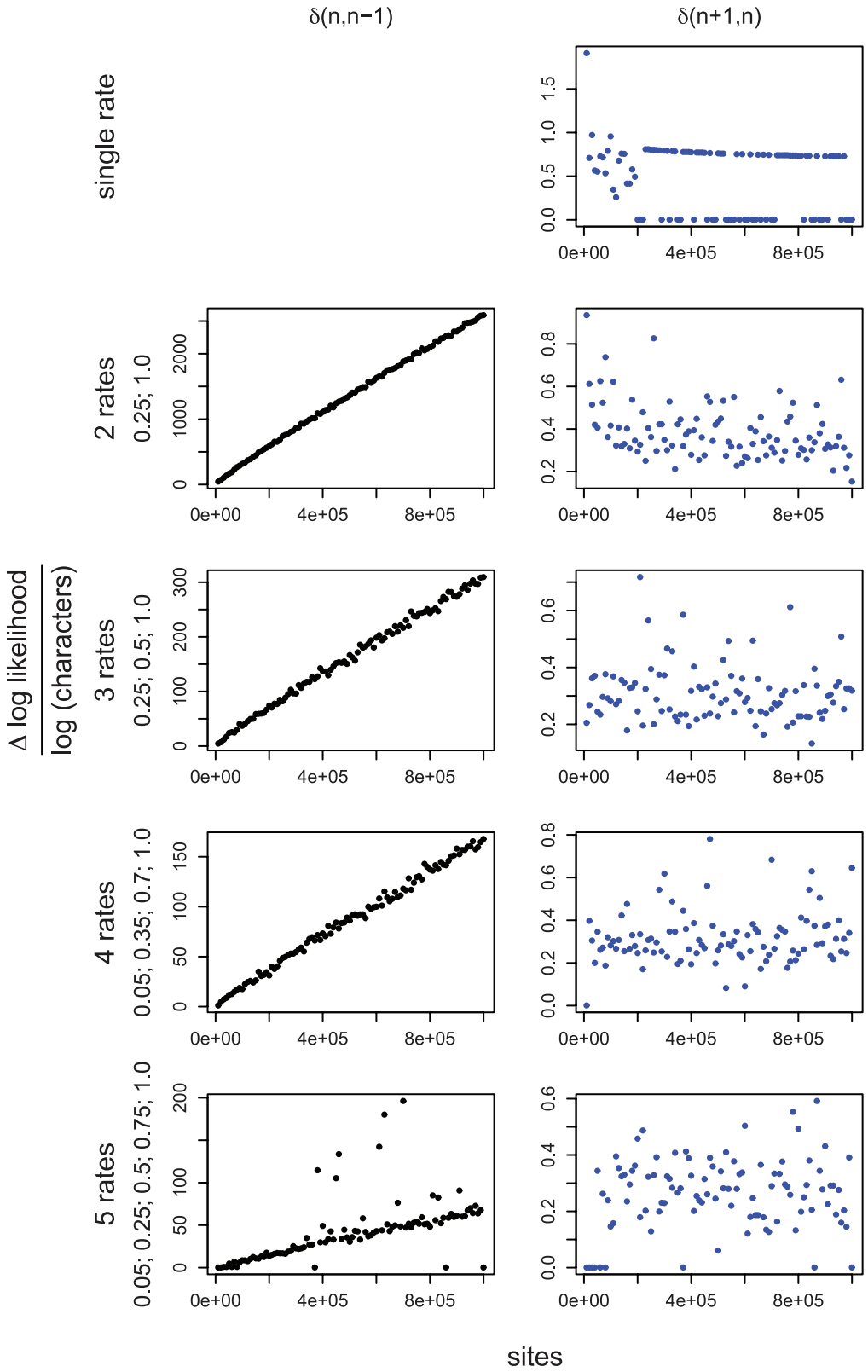


Figure 1. Simulation studies used to derive the appropriate penalty term for *mBIC*. Each panel plots the difference in log likelihood ($\log L$) normalized by the logarithm of the sample size (number of characters), between best fitting GA models with n and k rates ($\delta(n, k)$), against the number of sites in the alignment. For simulations with a single rate class we plotted $\delta(2, 1)$, top right. Figures for multiple rate simulations (2–5 rates) show $\delta(n, n-1)$ as black dots (left column); and $\delta(n+1, n)$ as blue dots (right column). Values to the right of row report simulated rates for each class. The left column is a reflection of power, whereas the right column – of the degree of over-fitting. For the case where a single rate was simulated, the degree of over-fitting is the rate of false positives. The desired behavior for $mBIC$ is achieved when the model with n rate classes is preferred to

models with $n-1$, and $n+1$ rate classes. For a modified BIC criterion $mBIC = -2\log L + cp\log s$ with $c=2$, the former happens if $\delta(n, n-1)>1$ (more definitively with increasing sample size), and the latter if $\delta(n+1, n)<1$ (regardless of sample size).
doi:10.1371/journal.pcbi.1000885.g001

the probability that model i is the best model to explain the data, in the sense of minimizing the Kullback-Leibler divergence from the "true" unobserved model [42]. In addition to the $G\,A_s$ model, we also construct a set of credible models, i.e. all those models whose r_i is sufficiently large (≥ 0.01). From this credible set we compute a model averaged estimate of any parameter p, by a weighted sum of the estimate under model i, p_i as $\sum_i w_i p_i$, where the Akaike weight of model i, w_i is defined as $w_i = r_i / \sum_j r_j$. This $G\,A_r$ model is an analog of an empirical substitution model (e.g. ECM), and has no rate parameters that are estimated from validation data sets. By combining information from multiple models, statistical noise may be reduced (e.g. [26]).

3. The numerical $G\,A_r$ model with the addition of a single non-synonymous substitution rate parameter ($G\,A_r+\omega$) which multiplies all non-synonymous substitution rates in the Q matrix. The direct analog is the $ECM+\omega$ model of [10], and its purpose is to add a dataset specific "adjustment" to the baseline numerical model, since the estimated parameters of the baseline numerical model are weighted over the credible set and fixed at these estimates when applied to other datasets.

We used both BIC [33] and Likelihood ratio tests, where appropriate, for model comparison. These goodness-of-fit comparisons allowed us to evaluate whether a model estimated on reference alignments yielded a significant improvement over the other models when fitted to independent alignments for the same taxonomic groups. All models were implemented with the F61 frequency parameterization, in addition to their original frequency parameterizations, because the methodology used to estimate the ECM model precluded the use of other frequency parameterizations for across-the-board comparison. Alignments and phylogenetic trees were provided for the Pandit data set. In all other cases, alignments were generated using codon alignment tools implemented in HyPhy [32]. Maximum likelihood phylogenetic trees were estimated using PhyML [43] under a GTR [44] model of nucleotide substitution and among-site rate variation modeled as a discretized gamma distribution with 4 rate-classes [45]. Empirical alignments and trees are available at http://www.hyphy.org/pubs/cms/.

Rate matrix comparisons

The entries of the substitution rate matrix Q can be used to estimate the expected number of substitutions per site per unit time, $E(t) = -t\sum_i \pi_i q_{ii}$, and to determine the value of the time parameter (assuming all other parameters are known) t_1 which yields $E(t_1)=1$. Furthermore, the expression for the number of expected one-nucleotide substitutions between codons i and j, in time t, at a site is given by $E_{ij}(t) = \pi_i q_{ij} + \pi_j q_{ji} = 2\pi_i q_{ij}$ (the simplification is the consequence of time-reversibility). Given two amino-acid residues x and y which can be exchanged by a single nucleotide substitution, we can further define $E_{xy}(t) = \sum_{A_i=x, A_j=y} E_{ij}(t)$, where A_i denotes the residue encoded by codon i. Consider a $75-$element substitution spectrum vector $S_Q(t) = (E_{A,G}(t), \ldots, E_{K,R}(t))$, which describes the relative abundance or paucity of a particular type of amino-acid pair substitution under the model defined by Q. Given two models, Q_1 and Q_2, we propose to compare their similarity by computing the distance between the corresponding substitution spectrum vectors evaluated at the corresponding "normalized" times:

Table 2. The performance of GA model selection with $mBIC$ in estimating the number and membership of K rate classes as well as rate values from simulated data.

C	taxa	D	simulated rates	σ	P_m	P_u	P_o	P_c
1	2	0.2	n/a	n/a	0.99	n/a	0.01	n/a
2	2	0.2	(0.25, 1.0)	(0.004, 0.010)	1.00	0	0	1.00
			(0.25, 0.3)	(0.012, 0.009)	0.98	0.02	0	0.860
3	2	0.2	(0.25, 0.5, 1.0)	(0.011, 0.015, 0.053)	1.00	0	0	0.996
			(0.25, 0.35, 0.5)	(0.004, 0.011, 0.008)	0.97	0.03	0	0.971
4	2	0.2	(0.05, 0.35, 0.7, 1.0)	(0.006, 0.021, 0.040, 0.041)	0.99	0.01	0	0.993
			(0.5, 0.65, 0.75, 1.0)	(0.004, 0.007, 0.006, 0.006)	0.82	0.18	0	0.936
5	2	0.2	(0.05, 0.25, 0.5, 0.75, 1.0)	(0.003, 0.012, 0.008, 0.014, 0.012)	0.91	0.09	0	0.981
			(0.5, 0.65, 0.75, 0.85, 1.0)	(0.003, 0.005, 0.006, 0.007, 0.010)	0.67	0.33	0	0.927
1	16	0.2	n/a	n/a	1.00	0	0	n/a
2	16	0.2	(0.25, 1.0)	(0.016, 0.044)	1.00	0	0	0.923
3	16	0.2	(0.25, 0.5, 1.0)	(0.022, 0.045, 0.052)	0.23	0.77	0	0.713
		0.2	(0.25, 0.75, 1.5)	(0.019, 0.050, 0.061)	1.00	0	0	0.837
		0.5	(0.25, 0.5, 1.0)	(0.014, 0.022, 0.037)	1.00	0	0	0.861
3	32	0.2	(0.25, 0.5, 1.0)	(0.018, 0.026, 0.038)	0.89	0.11	0	0.817

D measures the simulated pairwise sequence divergence (expected substitutions/nucleotide site); σ, standard deviation (averaged over replicates) of estimated rates from the generating values; P_m, the proportion of simulations for which the correct number of rate classes are inferred; P_u, the proportion of simulations which are under-fitted, P_o, the proportion of simulations which are over-fitted, and P_c, the mean Rand C-statistic [35] between rate clusters in the generating model and that in the inferred models.
doi:10.1371/journal.pcbi.1000885.t002

Table 3. Empirical data set characteristics.

source	Taxon	Gene	# taxa	# sites	D	K	C_k
Pandit/Pfam (PF03477)*	Multiple	ATP cone	72	312	66.6[†]	5	(0.007, 0.036, 0.144, 0.341, 3.108)
Pandit/Pfam (PF06455)*	Multiple	NADH5 C	82	552	1.68	4	(0.043, 0.208, 0.456, 0.910)
Pandit/Pfam (PF02780)*	Multiple	Transketolase C	83	393	3.00	6	(0.002, 0.033, 0.094,
							0.268, 0.678, 4.744)
[38]*	Vertebrate	Rhodopsin	38	990	0.44	4	(0.018, 0.116, 0.371, 0.724)
[18]	Yeast	Pyruvate kinase	16	1389	0.51	4	(0.024, 0.093, 0.226, 0.608)
(YAL038W)*							
NCBI*	HIV-1 group M	*pol*	142	2847	0.15	7	(0.047, 0.114, 0.211, 0.350,
							0.532, 0.998, 1.562)
[39]	HIV-1 subtype B	*pol*	371	1497	0.06	n/a	n/a
[40]	HIV-1 subtype C	*pol*	348	1170	0.09	n/a	n/a
NCBI*	Seasonal IAV	*HA*	349	987	0.09	3	(0.350, 1.211, 3.287)
NCBI	IAV A H5N1	*HA*	279	1545	0.04	n/a	n/a
[41]	IAV A H3N2	*HA*	68	987	0.02	n/a	n/a

D is mean pairwise nucleotide divergence (substitutions/site, estimated under the single rate codon model), K is the number of rates estimated in the GA, C_k are the maximum likelihood estimates for the rates.
*Reference alignments for which GA models were estimated. All GA results presented are for the model with best $mBIC$.
[†]ATP cone is comprised of highly divergent sequences, with only 22% average pairwise amino-acid identity; synonymous rates appear to be saturated.
doi:10.1371/journal.pcbi.1000885.t003

$$D(Q_1, Q_2) = \left\| S_{Q_1}\left(t_1^{Q_1}\right) - S_{Q_2}\left(t_1^{Q_2}\right) \right\| \quad (2)$$

Any norm on the standard $75-$dimension real valued vector space can be used, but for the purposes of this paper we consider the L_2 norm, and the corresponding induced Euclidean distance metric.

Implementation

All models and data sets utilized in this study are implemented as scripts in the HyPhy Batch Language (HBL), and are be available with the current source release of HyPhy [32]. In addition, we have made the GA codon model selector available as an analysis option at http://www.datamonkey.org [46]. The GA model selection code requires an MPI cluster environment with typical runtimes of approximately 36–48 hours for an intermediate-sized alignment (50 taxa) and 32 compute nodes.

Results

Power and accuracy analysis on simulated data

Results from both two- and multi-taxon simulations (Table 2, Figure 1) indicated that $mBIC$ controlled the rates of overfitting, defined as the proportion of replicates that overestimated the number of rate classes K, P_o. For null (single-rate model) simulations ($K=1$), false positive rates were 0.01 for two-taxon simulations and <0.01 for 16-taxon simulation. Neither two- nor multi-taxon simulations showed over-fitting across any simulation scenarios (Table 2). We deliberately designed the procedure to be conservative, since over-fitting is a major concern in statistical model selection. The power to select the correct number of rate classes K (P_m) behaved as expected: increasing, and eventually reaching 100%, given sufficiently divergent sequences and well resolved rate classes (Table 2). Indeed, the limited information content of alignments where simulated rate classes are similar (i.e 3 rates of $0.25, 0.35, 0.5$), and/or where pairwise sequence divergence is low (0.2), was evident as increased model under-fitting (Table 2), P_u. Model under-fitting was substantially reduced when information content was increased, either by boosting the disparity in rate classes, or by elevating sequence divergence and/or number of taxa (Table 2). Further evidence that the GA procedure has high power is provided by the positive association of the difference between $mBIC$ scores of the correct model with K rates, and one with K-1 rates, and separation between simulated rates, pairwise sequence divergence or number of taxa (Table S1). The ability to assign individual rates to the correct group (as measured by the Rand statistic) was similarly improved, while the variance in numerical rate parameter estimates decreased, for more divergent sequences and rate classes, suggesting that the GA search procedure recaptures most of the rate class structure, given sufficient information.

Empirical data analysis

We compared the fit of 6 codon substitution models (Table 1) on 11 empirical data sets (Table 3), spanning a range of proteins, taxonomic groups and divergence levels, using the BIC to measure goodness-of-fit. Using the GA procedure, we inferred distinct multi-rate models from 7 of these data sets (labelled with asterisks in Table 3). The remaining 4 alignments were used for validation such that we could determine the generalizability of two of the GA-fitted models (HIV and IAV) to other alignments from the same taxonomic groups. In 5 cases, the GA model outperforms every other model (often by a large margin), and in 2 cases it comes in second after the parameter rich REV model (Table 4). Note that the GA model outperforms REV in all 7 cases under the more conservative $mBIC$ criterion (which was used to inform the GA). Data set specific GA models consistently fit the data better than state-of-the-art empirical (ECM) and mechanistic (LCAP) models.

An intuitive understanding of the model selection process via the GA may be gained by thinking of it as a non-linear curve fitting problem, where the "true" curve is the unobserved distribution of biological substitution rates (Figure 2). We consider the 61×61 substitution rate matrix for a codon model, extract non-

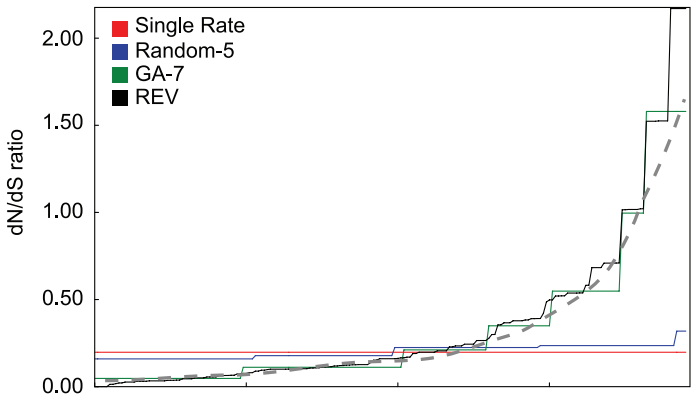
Table 4. Comparison of empirical model fits using BIC.

	S+F61	ECM+F61	ECM+F61+ω	LCAP+F61	GA$_s$+F61	REV+F61
ATP cone*	42176.4 (5)	41563.4 (3)	41329.6 (2)	49049 (6)	**41214.6 (1)**	41831.6 (4)
NADH5 C*	69057.9 (3)	69148.1 (5)	69099 (4)	72329.4 (6)	68086.3 (2)	**67211.8 (1)**
Transketolase C*	63509.4 (5)	61436.2 (2)	61443.7 (3)	67819.7 (6)	**61227.8 (1)**	61469.4 (4)
Rhodopsin *	27918.7 (5)	28583.3 (6)	27769.6 (3)	27614.7 (2)	**27322.7 (1)**	27781.3 (4)
Yeast Protein YAL038W*	21219.1 (5)	22246.1 (6)	20988.8 (2)	21098.2 (3)	**20822.7 (1)**	21142.7 (4)
HIV-1 *pol* Group M*	148650 (4)	158788 (6)	156792 (5)	146381 (3)	145338 (2)	**145209 (1)**
HIV-1 *pol* subtype B	113583 (4)	119721 (6)	119196 (5)	111249 (3)	**108251 (1)**	110113 (2)
HIV-1 *pol* subtype C	127143 (4)	134719 (6)	133794 (5)	125407 (3)	124434 (2)	**123346 (1)**
Influenza A *HA**	17803.9 (3)	19479.7 (6)	18883.3 (5)	17750.6 (2)	**17558.8 (1)**	18110.3 (4)
Influenza A *HA* H5N1	**28326.2 (1)**	28987.1 (6)	28911.7 (5)	28382.8 (3)	28347.2 (2)	28904.2 (4)
Influenza A *HA* H3N2	**7527.03 (1)**	7649.29 (4)	7658.29 (5)	7562.29 (3)	7546.24 (2)	8096.39 (6)

The best model (with smallest BIC) is shown in boldface and the rank of each model is provided in parentheses.
*Reference alignments from which GA models were estimated.
doi:10.1371/journal.pcbi.1000885.t004

synonymous rates for the 196 above-diagonal entries which correspond to one-step non-synonymous substitutions and rank them in an increasing order to obtain monotonically increasing rate curves as shown in (Figure 2). Note that because the ratios for all substitutions between the same pair of amino-acids (of which there are 75 pairs) are identical, this will create steps in such curves. In the case of one non-synonymous substitution rate (SR) the curve is a flat line at the estimated average non-synonymous substitution rate across all residue pairs. This is easily improved on by a random model which assigns non-synonymous substitutions randomly to one of 5 rate classes. At the other extreme lies the general time reversible models with 75 estimated rates. Since we have no *a priori* reason to believe that any two non-synonymous substitution rates will be exactly the same, REV is the most biologically realistic of the models which assume time-reversibility and only single nucleotide substitutions. However, fitting the parameter rich REV model to limited data is statistically unsound. The GA-approach, instead, searches for the best (in an information theoretic sense) step-wise smoothing of the biological distribution given the data available (Figure 2).

The "generalist" ECM model sacrifices gene-level resolution, in some cases so dramatically that it underperforms the single-rate model, even with the correction factor ω (Table 4). For instance, ECM appears to be ill suited for the analysis of viral genes. LCAP, on the other hand, performs poorly for highly divergent data sets; indeed the original validation of LCAP took place on relatively closely related yeast species [18], and the mechanistic properties assumed by the model may be insufficient in alignments spanning multiple genera and taxonomic groups. To test whether GA structured models are generalizable, we estimated two viral



Model	Log-L	Rate Parameters	Min	dN/dS Median	Max
Single Rate	-71337.6	1	0.197	0.197	0.197
Random-5	-71275	5	0.159	0.200	0.318
GA-7	-69413.1	7	0.047	0.327	1.580
REV	-69366.7	75	0.000	0.354	2.170

Figure 2. Evolutionary rate estimation as "curve fitting." An example from HIV-1 polymerase gene alignment for which the *GA* inferred 7 non-synonymous rate classes. The idealized biological rate distribution (unobservable) is depicted by the dashed line. The goodness of fit, the complexity of the models, and the range of maximum likelihood parameter estimates are listed in the table.
doi:10.1371/journal.pcbi.1000885.g002

models: one for HIV-1 polymerase and one for human IAV hemagglutinin. We then applied each of these models (holding the inferred class structure fixed) to two additional samples of sequences from the same gene, obtained independently from the training sample. In all 4 cases GA_s outperformed ECM, ECM+ω and LCAP by wide margins, lending credence to the claim that data-driven structured models recover substitutional biases that are shared by other samples shaped by similar evolutionary parameters. Curiously, for very low divergence (and low information content) intra-serotype IAV alignments, the single rate model was preferred to all other models by BIC, suggesting that there are biologically interesting alignments, which do not contain sufficient amino-acid variability to indicate the use of a multi-rate model.

As a test of protein-specificity of GA_s models, we randomly selected four Pandit data sets to assess how well GA_s models inferred from unrelated proteins fitted these data (Table S2). Not surprisingly, ECM was the best model in 3/4 cases, because it was derived as the best "average" protein model. LCAP topped the list in one case, but placed outside the top three in the other three cases. The GA structured models, being tailored to specific proteins, tended to differ from each other (Table S3) and did not perform well on proteins from different families. However, the GA structured models for ATP cone and Transketolase C did outperform the LCAP model in 3/4 cases, which suggests some similarity between the respective protein families in those cases. This indicates the GA models fitted to different proteins may be generalizable, with the degree limited by taxonomy, protein function or both. The generalizability of GA models could further be quantified by evolutionary fingerprinting of genes [27]; see also Figure 3(b).

Further analysis of GA multi-rate models

A GA search run typically examines between two- and a hundred-thousand potential models, e.g. 28770 models with 1 to 8 rate classes for the HIV-1 group M *pol* dataset. GA_s, which we compared to existing models in the previous section, is simply the single "best" model, i.e. the model that minimized the *mBIC* criterion among all those examined during the run. Further, we estimate the credible set of models as those models whose evidence ratio versus the best model is sufficiently large (see methods). Among 28770 models fitted to HIV-1 pol by the GA, 567 belonged to the credible set. Given sufficient data and knowing that the true model is in the set examined by the GA, e.g. in the long 2-sequence simulations discussed above, the size of the credible set frequently shrinks to 1 (the true model). These structured (GA_s) and model-averaged (GA_r) models can be analyzed further to draw inferences of the substitution process.

For instance, the structured GA_s model identifies which residue pairs are exchanged rarely, relative to the baseline synonymous rate. In Figures 4 and 5 we cluster the pairs of residues which have the same rate of non-synonymous substitution; residues are labelled by Stanfel class and physicochemical properties. Note that the same residue can be present as a node in multiple clusters because the GA partitions residue pairs (i.e. the rates between them), not the residues themselves. The model reveals a startling heterogeneity of substitution rates in HIV-1 *pol*: the single rate dN/dS estimate of 0.15 is resolved into 7 rate classes (Figure 4), with relative non-synonymous substitution rates ranging from 0.047 (20 residue pairs) to 1.561 (3 residue pairs); a similar range is revealed for other datasets (Table 3). It is remarkable that some of the non-synonymous substitutions occur at rates matching or exceeding the gene-average rate of synonymous substitutions. This can be interpreted, for instance, as lack of selective constraint on particular residue substitutions gene-wide, or evidence of directional selection when some residues are preferentially replaced with others. Regardless of how this result is interpreted, a remarkable complexity of substitution patterns is revealed by the analysis. We hypothesize that such patterns reflect complex dynamics of substitutional preferences that may be shared by multiple samples of the same genes. This hypothesis is supported (by the goodness-of-fit of GA_s vs other models) on HIV-1 and IAV samples in this study (Table 4), and we are currently undertaking the GA analysis of several thousand alignments to confirm this finding.

One of the benefits of using the GA_s model instead of REV or other models is that the former model automatically classifies all substitutions into similarity groups, supplying a data-driven analog of "conservative" or "radical" substitutions, previously defined *a priori* based on chemical properties of the residues, or a more sophisticated multi-property basis defined in the LCAP model. For example, the 75 substitution rates are partitioned into seven classes in the GA_s model inferred from HIV-1 *pol*, and into 4 rate classes

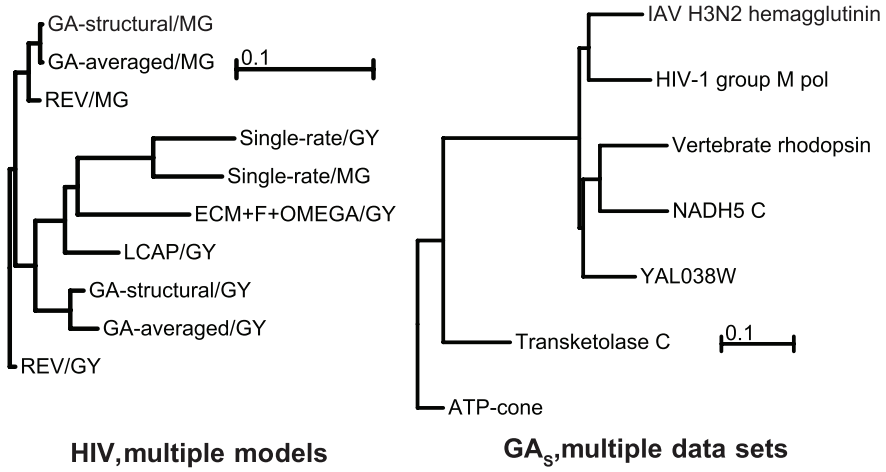


Figure 3. Neighbor-joining [57] trees built from matrices of pairwise substitution spectrum distances (Eq. 2) computed between different models fitted to the HIV-1 group M *pol* alignment, and between GA_s models inferred from different alignments.
doi:10.1371/journal.pcbi.1000885.g003

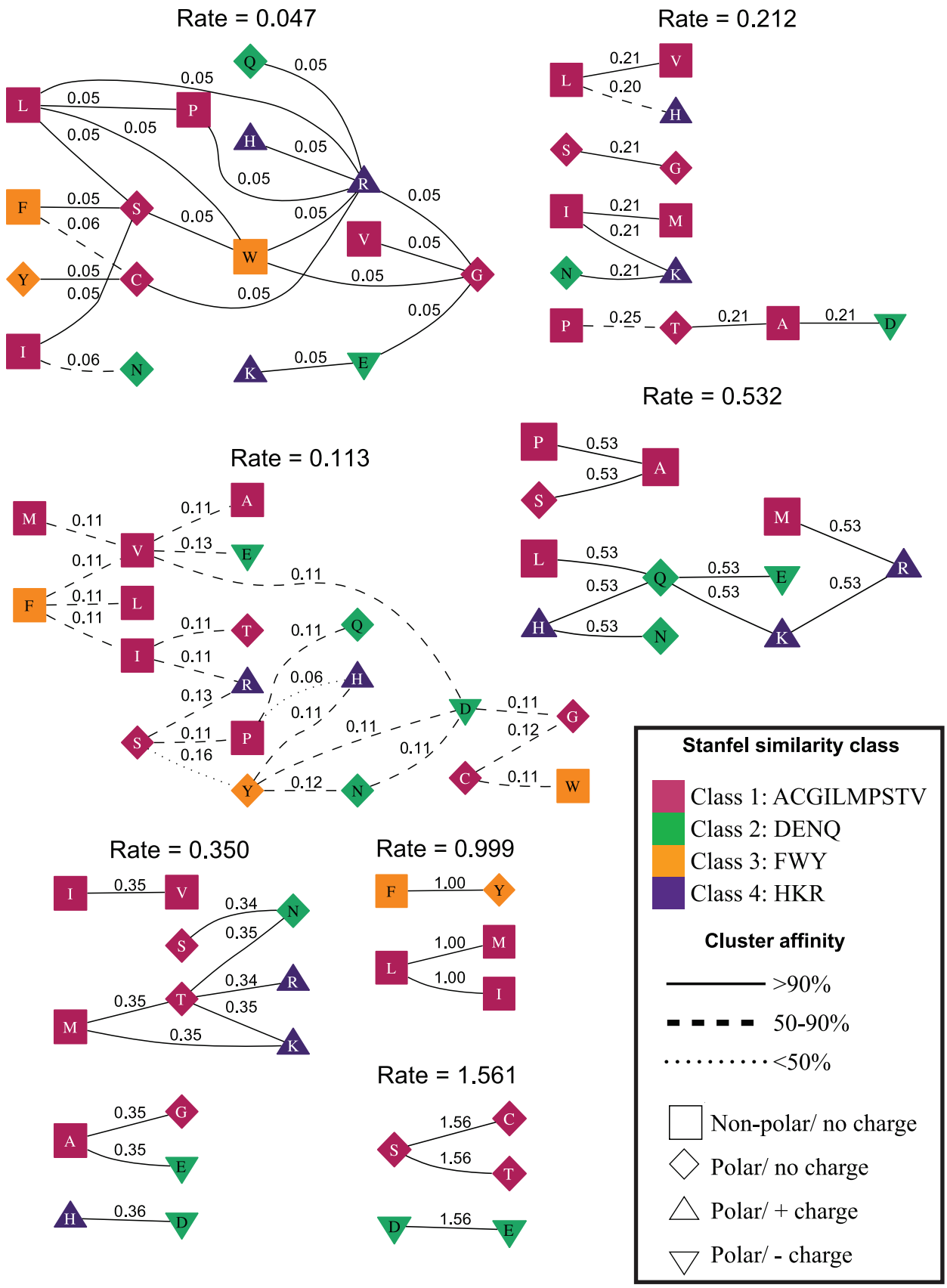

Rate = 0.047
Rate = 0.212
Rate = 0.532
Rate = 0.113
Rate = 0.350
Rate = 0.999
Rate = 1.561
Stanfel similarity class
Class 1: ACGILMPSTV
Class 2: DENQ
Class 3: FWY
Class 4: HKR
Cluster affinity
>90%
50-90%
<50%
Non-polar/ no charge
Polar/ no charge
Polar/ + charge
Polar/ - charge

Figure 4. Evolutionary rate clusters in structured GA models (GA_s) inferred from the HIV-1 group M *pol* alignment. Each cluster is labeled with the maximum likelihood estimate of its rate inferred under GA_s. The residues (nodes) are annotated by their biochemical properties and Stanfel class, and the rates (edges) are labeled with model-averaged (GA_r) rate estimates. The style of an edge is determined by its cluster affinity, where high cluster affinities indicate that a large proportion of models in the credible set were consistent with the structured GA_s model.
doi:10.1371/journal.pcbi.1000885.g004

for the GA_s model fitted to a smaller, but more divergent vertebrate rhodopsin alignment (Figure 5).

Multi-model inference is instrumental in assessing how robust the clustering assignment made by GA_s is. In Figures 4 and 5, we present this information by labeling individual substitution rates with their model averaged values. An examination of the numerical differences between rate estimates (for a particular amino-acid pair) obtained under GA_s and GA_r can reveal ambiguities in assigning a particular rate to a class. More formally, we can compute a model averaged support for the probability that rates $R_1 \leftrightarrow R_2$ and $R_3 \leftrightarrow R_4$ (for residues $R_1 \neq R_2$, $R_3 \neq R_4$) are in the same class, as described above, or that the corresponding edges e_{12} and e_{34} are in the same component of the rate graph (Figure 4). If C is a cluster defined by $G\,A_s$ (with the number of nodes in C, $|C| \geq 2$), we define the *cluster affinity* of an edge $e \in C$ as the mean of the model averaged estimates of the probabilities that edge e and other edges in C belong to the same cluster:

$$A(e, C) = (|C|-1)^{-1} \sum_{\substack{h \in C \\ h \neq e}} \Pr\{h \text{ and } e \text{ cluster together}\}$$

If $A(e, C)$ is below a certain threshold, for instance 0.5 for majority rule, then cluster membership of edge e is ambiguous. For example, the $S \leftrightarrow Y$ substitution pair with a model-averaged non-synonymous rate of 0.16 is one of two rate pairs with low ($<50\%$) cluster affinity for HIV-1 (Figure 4). Two of the inferred GA_s rate classes have non-synonymous rates of 0.113 and 0.212, respectively, and the placement of model-averaged rate for $S \leftrightarrow Y$ between the two values is indicative of the alternate assignment of this substitution pair to these two rate classes among models of the credible set. A larger training data set may be able to infer an additional intermediate rate class between 0.113 and 0.212. While GA_r yields more robust numeric estimates of substitution rates for a single data set, GA_s has better *BIC* fit on validation HIV and IAV alignments (results not shown).

The relationship between substitution rates and residue properties

The expectation that substitutions which preserve amino acid physicochemical properties occur at a lower rate than property-altering substitutions has previously been evaluated in the maximum likelihood codon model context [20,47]. However, in published work, property-altering and property-conserving amino acid classes are defined *a priori*, whereas in the GA approach amino acid substitution pairs are first partitioned into classes based on rate similarity, and thereafter property preserving versus property-altering rates can be compared. The increased substitution rate of property preserving substitutions, holds largely – but not universally – for GA_s and GA_r rates, as evidenced in Figures 6 and 7. For example, in the vertebrate rhodopsin sample, the median rate of charge-changing substitutions is significantly lower than the charge-preserving substitutions, but the two medians are not significantly different in the HIV-1 *pol* sample. The rates were negatively correlated ($p<0.05$, one-sided Pearson product moment test, no multiple test correction) with 4 out of 5 property-based distances (polarity, volume, isoelectric point and hydropathy) that form the basis of the LCAP model. However, while the broad pattern follows the expectation, the consistently better fit of *GA*-based models, and the presence of strong outliers, such as $H \leftrightarrow Q$ and $M \leftrightarrow R$ in the 0.532 cluster of HIV-1 rates (Figure 4), suggests that our data driven approach detects significant deviations from purely biochemical rate expectation. These deviations could be attributed to selective pressures which promote property changes, or could arise because not all biologically relevant important properties have been included into structured models.

One benefit of our approach over the "amino acid class" models [20,47] is that transitivity of rates (i.e. the requirement that if $X \leftrightarrow Y$, and $Y \leftrightarrow Z$ are in the same rate class, then so is $X \leftrightarrow Z$) is not enforced by the GA models. Because we focus on modeling single-nucleotide substitution rates only, the structure of the genetic code itself contradicts transitivity. For instance both E (encoded by GAR) $\leftrightarrow$ $G(GGN)$ and $G \leftrightarrow R(AGR)$ are one-step substitutions, but $E \rightarrow R$ is not. Further, since amino acid class models only estimate two non-synonymous rates (within and between classes), it is a necessary condition that non-synonymous rates which change amino acid property be shared irrespective of how much the property is being changed. For instance, substitution rates which change charge from negative to positive will be the same as those which change charge from negative/positive to uncharged. If amino acid substitutions that result in a positive charge are favored, then these transitive conditions are not representative of the substitution process. Furthermore, the amino acid class models assume all substitutions within classes occur at the same rate. This is a very strong assumption since some amino acids with the same physicochemical property class are separated by more than one nucleotide substitution, e.g. positively charged amino acids $H(CAY)$ and $K(AAR)$. Although we do not account for multiple nucleotide substitutions in the GA model directly (but see below), previous work has demonstrated that these occur at lower rates than single-nucleotide substitutions [9,10,48].

Model clustering

Using the substitution spectrum distance defined in Equation (2), it is easy to construct a hierarchical clustering of several models fitted to the same dataset, as well as between models fitted to different datasets. The former is useful to interpret how much difference in predicted substitution patterns over a unit of evolutionary time there is between different descriptions of the same data, whilst the latter naturally extends the concept of evolutionary fingerprinting of non-homologous genes [27]. For HIV-1 pol (Figure 3), GA_s and GA_r models both clustered closely with the rate substitution pattern predicted by the REV model, followed by LCAP, ECM+F+ω, and finally – distant single rate models. The similarity between REV and GA models was especially strong for the MG parameterization, under which the GA models were inferred. In a between-genes model comparison (Figure 3), the two viral alignments clustered together, as did the two most divergent alignments (ATP-cone and Transketolase C).

Effects of substitution models on statistical inference

Statistical inference procedures based on phylogenetic models have varying degrees of robustness with respect to the substitution rate matrix used in the analysis. For a multi-rate model, it is

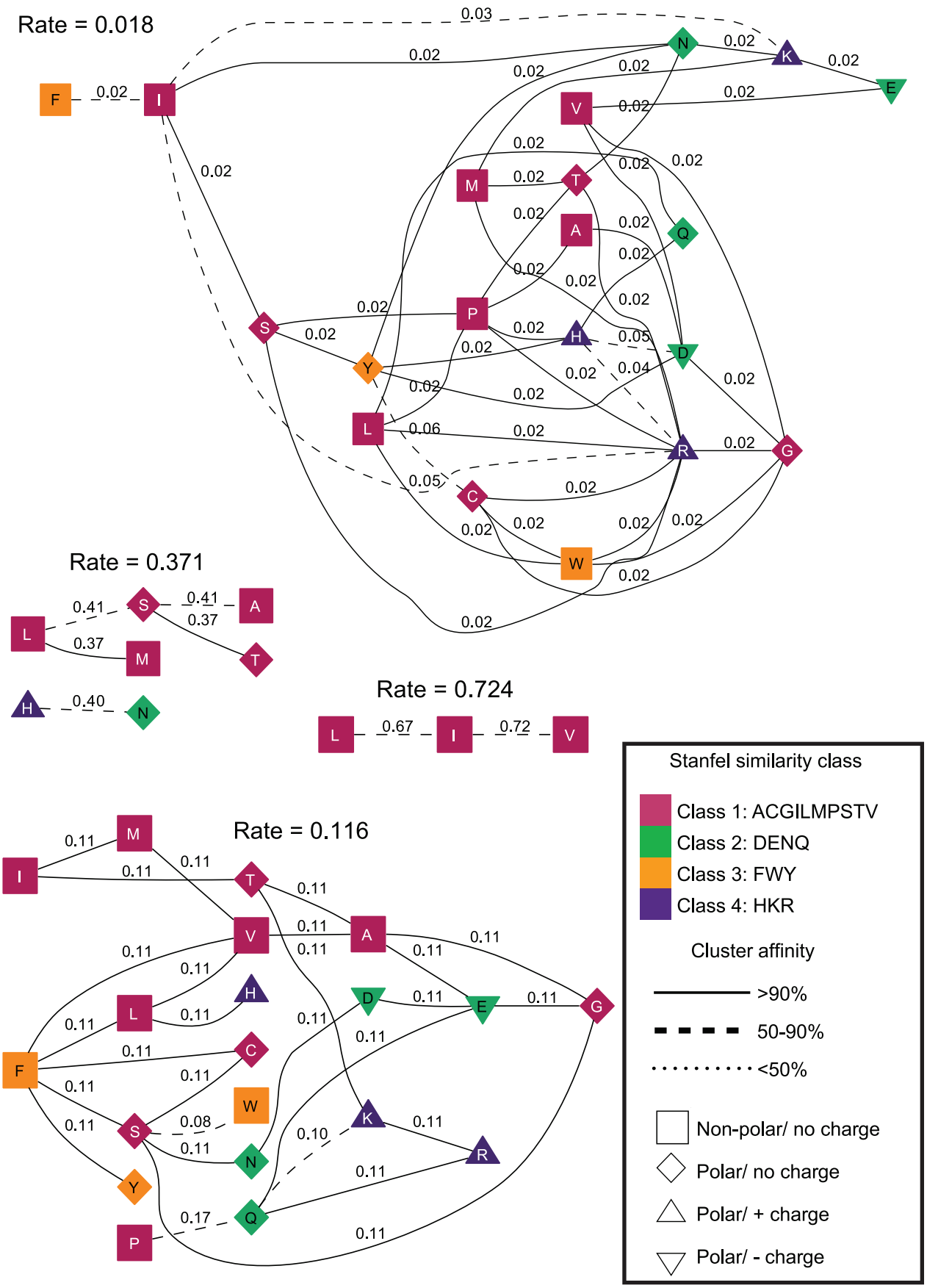

Rate = 0.018
Rate = 0.371
Rate = 0.724
Rate = 0.116
Stanfel similarity class
Class 1: ACGILMPSTV
Class 2: DENQ
Class 3: FWY
Class 4: HKR
Cluster affinity
>90%
50-90%
<50%
Non-polar/ no charge
Polar/ no charge
Polar/ + charge
Polar/ - charge

Figure 5. Evolutionary rate clusters in structured GA models (GA_s) inferred from the vertebrate rhodopsin protein alignment. Each cluster is labeled with the maximum likelihood estimate of its rate inferred under GA_s. The residues (nodes) are annotated by their biochemical properties and Stanfel class, and the rates (edges) are labeled with model-averaged (GA_r) rate estimates. The style of an edge is determined by its cluster affinity, where high cluster affinities indicate that a large proportion of models in the credible set were consistent with the structured GA_s model.
doi:10.1371/journal.pcbi.1000885.g005

intuitively clear that the types of inference that rely on "mean" rates should be minimally affected, whereas those that depend on the individual residue rates can be affected significantly. We examine several such measures inferred from two of the datasets in this study.

Branch length estimates are essentially unchanged when moving from the single-rate (SR) model to a GA_s model. On the example HIV-1 *pol* dataset, the total tree length changed from to 5.29 (SR) expected substitutions/nucleotide to 5.41 (GA_s), and the lengths of individual branches were nearly perfectly linearly correlated with

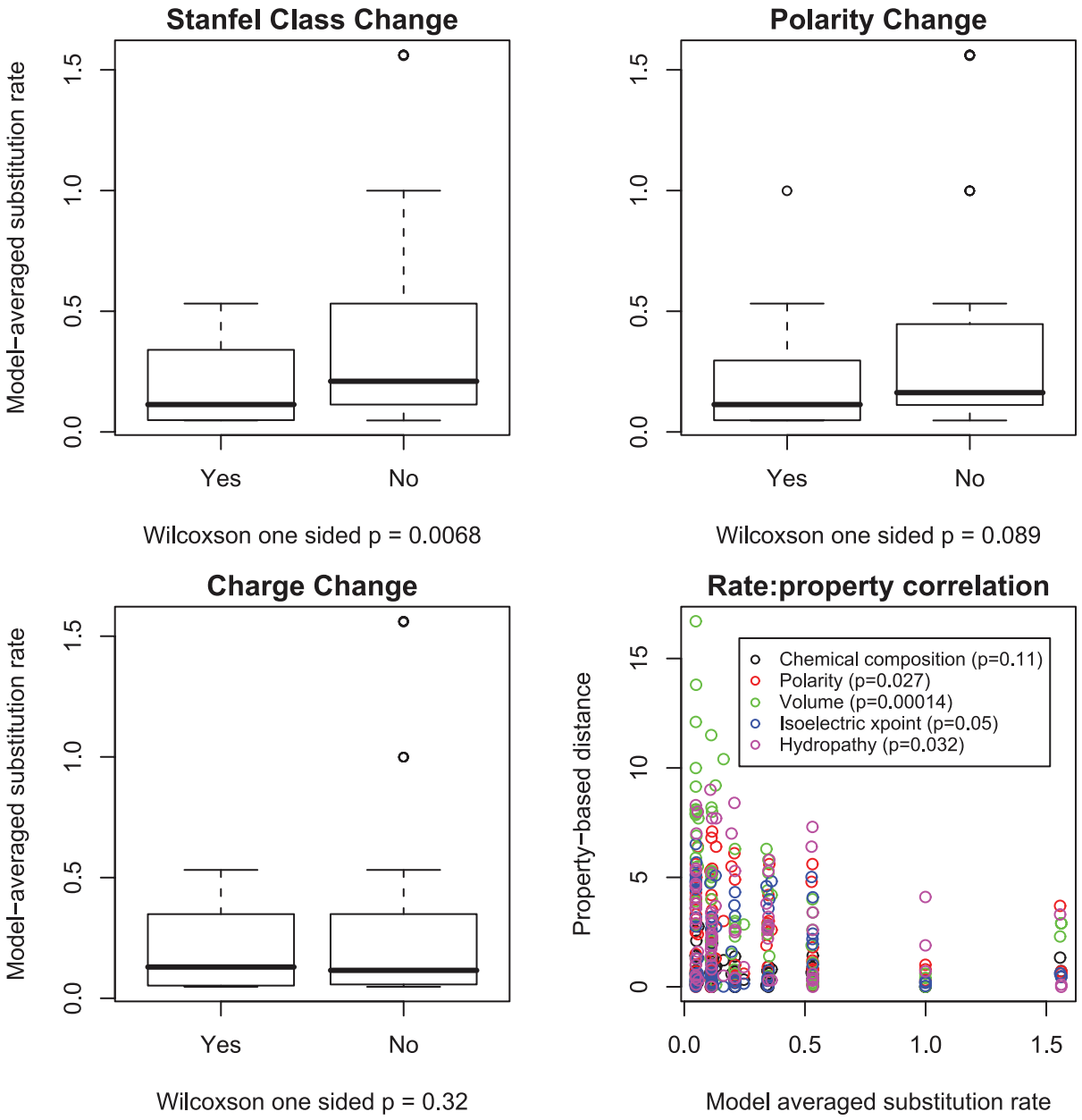


Figure 6. Correlations of lower substitution rates and property preservation in the HIV-1 group M *pol* alignment. Model-averaged GA_r rates were stratified by whether or not they involved a change in polarity, charge or Stanfel class, the medians of two rate distributions were compared using a one sided Wilcoxon rank-sum test. We further correlated the magnitude of substitution rates with one of five property-based distances between the corresponding residues (defined in [18]) using a one-sided (negative correlation) Pearson product-moment correlation test.
doi:10.1371/journal.pcbi.1000885.g006

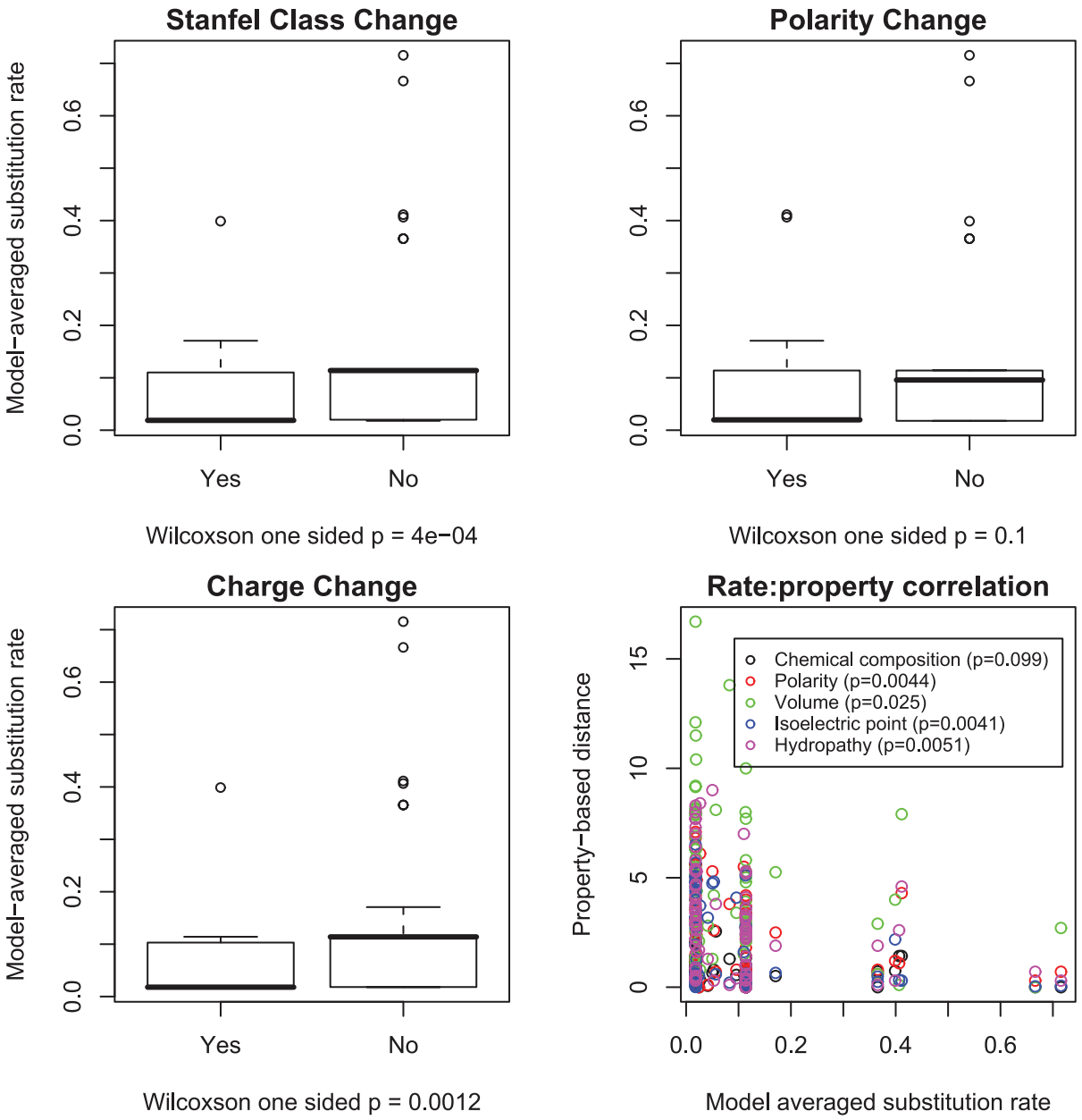


Figure 7. Correlations of lower substitution rates and property preservation in the vertebrate rhodopsin alignment. Model-averaged GA_r rates were stratified by whether or not they involved a change in polarity, charge or Stanfel class, the medians of two rate distributions were compared using a one sided Wilcoxon rank-sum test. We further correlated the magnitude of substitution rates with one of five property-based distances between the corresponding residues (defined in [18]) using a one-sided (negative correlation) Pearson product-moment correlation test.
doi:10.1371/journal.pcbi.1000885.g007

linear regression slope of 0.97, intercept of 0.0002 and $R^2 = 0.9996$.

Ancestral character reconstruction is considerably more sensitive to the substitution model. In the vertebrate rhodopsin data set, for example, the joint maximum likelihood ancestral reconstruction [49] under SR and GA_s models differed in the number of inferred non-synonymous substitutions at 10/330 sites, with more non-synonymous substitutions in 7 cases under GA_s. At 20/330 sites substitutions were mapped to a different set of branches.

Site-specific diversifying selection screens are likely to be profoundly affected by a switch from single- to multi-rate models. Consider the FEL method [50], where the SR model is fitted site-by-site and a likelihood ratio test (LRT) is used to test whether $\omega \neq 1$. First, because GA_s defines multiple substitution classes, one can now apply a variety of tests to see *which* non-synonymous rates at a given site exceed the baseline synonymous rate. To explore this approach for a 4−rate GA_s multi-class model applied to the vertebrate rhodopsin alignment, we performed 4 LRT tests, where we independently constrained each non-synonymous rate parameter (C_k, $k = 1 \ldots 4$, Table 1) to be equal to 1 at a site (neutral evolution in class k), vs an unconstrained 4−parameter alternative. This is analogous to performing a test for selection at a site by constraining the non-synonymous rate to be equal to the synonymous rate, and comparing the fit to the unconstrained model (FEL), except that we only place the constraint on one rate class at a time. At $p = 0.05$, the standard (SR) FEL reported 1/330

(codon 54) sites as being under diversifying selection (positively selected). However, for the GA_s model, there were 0, 1, 8 and 8 positively selected sites for the four substitution classes (Figure 5, increasing rate magnitude), respectively at the Bonferroni corrected p of 0.0125. Codon 54 was selected only with the fastest rate class ($r=0.72$), because the signal of selection is driven by a large number of $I\leftrightarrow V$ substitutions. Only one codon (198) was selected with two or more different tests (rate classes 0.116 and 0.72).

The effect of site-to-site rate variation

We remark that the effects of site-to-site rate variation and multiple non-synonymous rates appear to be largely additive, and not confounded. This is a critical observation: if the effects are confounded, then we cannot justify inferring the multi-rate model independently assuming no site-to-site rate variation, as is done in this manuscript for computational expedience. To illustrate, we fitted both a constant rate model and the general bivariate distribution [27], with and without accounting for multiple non-synonymous rate classes (Table 5). The constant rate model assumes all sites share the same rate of substitution, whereas a general bivariate distribution infers the number of site-to-site variation classes from the data [27]. These models were fitted to the vertebrate rhodopsin alignment, which exhibits extensive site to site rate heterogeneity. The GA_s inferred 4 non-synonymous rate classes for the rhodopsin alignment, whereas the single ω has one, resulting in three degrees of freedom for the comparison of these models. When the general bivariate model was fitted with a single ω or GA_s, 6 and 7 site classes were inferred, respectively, resulting in 4 degrees of freedom for the comparison of single ω and GA_s models (3 rate and 1 site class are added to the GA_s model). The important observation is that the addition of site-to-site rate variation component resulted in a significant improvement in log likelihood scores, regardless of the underlying substitution model (single ω or GA_s). This suggests that by allowing multiple rate classes, we are not merely fitting variability in site-to-site selective constraints. However, as the cost of computing cores in clusters decreases, we expect that it will become practical to infer GA_s models with the site-to-site rate variation component included directly in the search procedure.

The effect of allowing multiple instantaneous nucleotide substitutions

Recent extensions of codon models which permit multiple instantaneous nucleotide substitutions [9,10,48] tend to fit the data better than their traditionally parameterized counterparts. We explored whether this observation held for GA_s models using a straightforward extension of the rate matrix in Equation (1), following the ideas of [9]. We introduce four new independently estimated parameters to model the relative rates of synonymous (α_2, α_3) and non-synonymous (β_2, β_3) substitutions which replace two or three nucleotides, and modulate them by the product of the corresponding nucleotide rates θ_{ij} and the target codon frequency π (assuming the GY parameterization with the F61 estimator). For instance the rate of synonymous substitution (Serine) from AGT to TCT is $\alpha_2\theta_{AT}\theta_{CG}\pi_{TCT}$, while the rate of non-synonymous substitution $AAA\rightarrow CCC$ (Lysine to Proline) is $\beta_3\theta_{AC}^3\pi_{CCC}$.

Table 6 summarizes the effect of adding multi-step substitutions to SR and GA_s models for the vertebrate rhodopsin alignment. Much as was the case for site-to-site rate variation, the effects of multiple single-step non-synonymous rates and the non-zero rates of two or three nucleotide substitutions are additive at the $\log L$ level, and the estimates of single-step substitution rates were minimally influenced by the presence of the multi-step component (results not shown). The GA_s model augmented to allow multi-step substitutions can be directly compared to the Mechanistic-Empirical codon (MEC) model [9] coupled with the LG [51]empirical amino-acid substitution model (selected as the best fitting empirical model using the procedure implemented on http://www.datamonkey.org. Assuming no site-to-site rate variation, BIC of the MEC model is 27393.4, while that of the GA_s+multi-step model using the HKY85 nucleotide component (a direct analog to the MEC model) is 26800.6, once again highlighting how strongly the substitution process in an individual gene appears to deviate from the "average" encoded by empirical protein models.

The GA could be modified to search for optimal partitions among all 190 pairs of rates, for example using the above parameterization, but as the rhodopsin example indicates, the single-step and multi-step rate rate components appear to be effectively independent. We will explore this option in future versions of the model selection GA.

Discussion

In this manuscript we have developed, validated and benchmarked a procedure to quickly and reliably infer a multi-rate model from the combinatorially large class of general time-reversible codon substitution models. Using extensive simulations, we demonstrated that our conservative *mBIC* model selection criterion controls over-fitting and has excellent power on data sets of biologically realistic size, inferring the exact model simulated given sufficient sequence divergence and length. We have previously argued against using the single rate model as a

Table 5. The effects of modeling site-to-site rate variation and multiple non-synonymous rates in the vertebrate rhodopsin alignment using the MG frequency parameterization.

	Single ω	GA_s (+3 df)	$\Delta\log L$
Constant rates	−13382.6	−12954.2	428.4
General bivariate rates (+4 df)	−12780.8	−12500.4	280.4
$\Delta\log L$	601.8	453.8	

The entry for joint effect was obtained by running the general bivariate model fit using the GA_s model obtained under the assumption of constant site-to-site rates. df = degrees of freedom.

doi:10.1371/journal.pcbi.1000885.t005

Table 6. The effects of modeling multi-nucleotide instantaneous substitutions and multiple non-synonymous rates in the vertebrate rhodopsin alignment using the F61 frequency parameterization.

	Single ω	GA_s (+3 df)	$\Delta\log L$
Single-nucleotide substitutions only	−13317.6	−13005.5	312.1
Single and multi-nucleotide substitutions (+4 df)	−13033.4	−12712.5	320.9
$\Delta\log L$	284.2	293	

The entry for joint effect was obtained by augmenting the GA_s model with non-zero rates for substitutions requiring two or three nucleotide changes. df = degrees of freedom.

doi:10.1371/journal.pcbi.1000885.t006

benchmark against which multi-rate models should be compared, since it is trivial to improve upon using a random assignment of substitutions to rate classes [19]. We reiterate this argument here, and suggest we should rather consider how well a multi-rate model approximates the REV model (Figure 2), given the limitations posed by the information content in an alignment. On a diverse collection of biological data, GA_s models consistently outperform the best-in-class empirical and mechanistic models, and match the performance of fully parameterized general time reversible models with only a few biologically relevant rate parameters (Table 4). Therefore, the GA_s provides goodness of fit matching or exceeding that of REV, with substantially fewer parameters and is thus computationally and statistically feasible for downstream analyses.

ModelTest [22] has been universally adopted to mitigate the effect of model misspecification on statistical inference from nucleotide data, and we posit that a robust codon model selection procedure, for example the one offered in this paper, will play a similar role for codon data. In the same vein as ModelTest, we infer the best model (which we term the GA_s) for an alignment, and also utilize model averaging [26] to achieve more robust estimates of biologically relevant parameters. Certain applications of codon models, such as divergence estimation, appear unaffected by the gross biological over-simplification of single-rate models, because they are only influenced by the mean of substitution rates. Others, including ancestral sequence reconstruction (e.g. for guided site directed mutagenesis, [38]), substitution mapping (e.g. for co-evolutionary analysis, [52]) and character sampling (e.g. for data augmentation modeling approaches, [53]) can see moderate effects. Applications which are tightly integrated with the substitution model and the interpretation of its parameters, such as site-by-site positive selection detection (e.g. [50,54]), will be profoundly affected by the introduction of multiple rates. Our results strongly argue against the prospect of deriving a single "generalist" model of codon evolution, that is capable of fitting most protein alignments well. Hence we should strive to fit both gene and taxonomy specific models of codon evolution. We further hypothesize that independent alignments representing a gene or a protein family will share most of the model structure and confirm this with HIV-1 polymerase and Influenza A virus hemagglutinin examples. While significant further validation is required and is currently underway, we assert that a collection of substitution models inferred from carefully selected training datasets can provide a useful library of organism and gene-specific models to be used in inference on codon sequences. This is conceptually similar to a library of Hidden Markov profile models, inferred from seed alignments, used for detecting protein domain homology in the Pfam database [55]. In order to facilitate the process of generating gene and taxonomic specific multi-rate codon models we have implemented the GA on our free analysis webserver (http://www.datamonkey.org, [46]), and have begun to assemble a library of representative multi-rate substitution models that are needed to reduce biases in those procedures that are sensitive to model misspecification.

The inference of the multi-rate codon models should be considered more than just a necessary step for downstream applications. By examining the structure of inferred rate classes, we argue that the *GA* captures the *a priori* expectation that radical changes in one or more biochemical properties of a residue happen relatively infrequently, but also that a mere reliance on such data-abstract mechanistic properties misses out important gene and organism specific peculiarities of the evolutionary process. For instance the elevation of substitution rates between amino acids that do not preserve physicochemical properties may be indicative of selective pressures which promote property changes. These selective pressures are of crucial importance in understanding evolution in viruses, such as HIV-1, known to evade host immune response [56]. We anticipate that considering specific substitution types when estimating selective pressures will improve power, as demonstrated with our multi-rate FEL analysis of vertebrate rhodopsin. However, this may also increase the rate of false positives, a conjecture that can be evaluated with straightforward, but laborious simulations.

Finally, we demonstrate how simple metrics on GA_s models inferred from different (e.g. non-homologous) alignments can be used to obtain an objective measure of similarity and disparity in substitutional preferences in different proteins and thus improve the resolution in evolutionary fingerprinting of genes [27].

Supporting Information

Table S1 Difference in mean (standard deviation) model *mBIC* scores for multi-taxon simulations. *D* is the average pairwise divergence; $mBIC_n$ is the difference in model *mBIC* score between the model with $n-1$ rates and a more complex with n rates; P is the proportion of correctly identified models for 100 simulations. Positive *mBIC* scores indicate preference for the more complex model with n rates, i.e. $mBIC_n = mBIC_{n-1} - mBIC_n$.
Found at: doi:10.1371/journal.pcbi.1000885.s001 (0.04 MB PDF)

Table S2 Randomly selected Pandit data model comparisons using BIC. In each case we fitted the ECM, LCAP and GA_s models to each of four randomly selected Pandit datasets. Model ranks (BIC/difference in BIC score relative to the best model) are shown.
Found at: doi:10.1371/journal.pcbi.1000885.s002 (0.03 MB PDF)

Table S3 Qualitative comparison of structured GA models.
Found at: doi:10.1371/journal.pcbi.1000885.s003 (0.02 MB PDF)

Acknowledgments

We thank Associate Editor, Wen-Hsiung Li, Tal Pupko and an anonymous reviewer for insightful comments on an earlier draft of this manuscript.

Author Contributions

Conceived and designed the experiments: KS MBG SVM SLKP. Performed the experiments: WD GB SLKP. Analyzed the data: WD SLKP. Contributed reagents/materials/analysis tools: MBG SLKP. Wrote the paper: WD KS SVM SLKP.

References

1. Felsenstein J (1981) Evolutionary trees from DNA-sequences – a maximum-likelihood approach. J Mol Evol 17: 368–376.
2. Muse SV, Gaut BS (1994) A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol Biol Evol 11: 715–724.
3. Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11: 725–736.
4. Anisimova M, Kosiol C (2009) Investigating protein-coding sequence evolution with probabilistic codon substitution models. Mol Biol Evol 26: 255–271.
5. Delport W, Scheffler K, Seoighe C (2009) Models of coding sequence evolution. Brief Bioinform 10: 97–109.
6. Dayhoff MO, Eck EV, Park CM (1972) A model of evolutionary change in proteins. In: Dayhoff MO, ed. Atlas of protein sequence and structure, National Biomedical Research Foundation, Washington D.C., volume 5. pp 89–99.

7. Jones D, Taylor W, Thornton J (1992) The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 8: 275–82.
8. Whelan S, Goldman N (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol 18: 691–699.
9. Doron-Faigenboim A, Pupko T (2007) A combined empirical and mechanistic codon model. Mol Biol Evol 24: 388–397.
10. Kosiol C, Holmes I, Goldman N (2007) An empirical codon model for protein sequence evolution. Mol Biol Evol 24: 1464–1479.
11. Tavaré S (1986) Some probabilistic and statistical problems in the analysis of DNA sequences. Lectures on Mathematics in the Life Sciences 17: 57–86.
12. Whelan S, de Bakker PIW, Quevillon E, Rodriguez N, Goldman N (2006) Pandit: an evolution-centric database of protein and associated nucleotide domains with inferred trees. Nucleic Acids Res 34: D327–31.
13. Adachi J, Hasegawa M (1996) Model of amino acid substitution in proteins encoded by mitochondrial DNA. J Mol Evol 42: 459–468.
14. Adachi J, Waddell P, Martin W, Hasegawa M (2000) Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. J Mol Evol 50: 348–358.
15. Dimmic MW, Rest JS, Mindell DP, Goldstein RA (2002) rtREV: an amino acid substitution matrix for inference of retrovirus and reverse transcriptase phylogeny. J Mol Evol 55: 65–73.
16. Nickle DC, Heath L, Jensen MA, Gilbert PB, Mullins JI, et al. (2007) HIV-specific probabilistic models of protein evolution. PLoS ONE 2: e503.
17. Conant GC, Wagner GP, Stadler PF (2007) Modeling amino acid substitution patterns in orthologous and paralogous genes. Mol Phylogenet Evol 42: 298–307.
18. Conant GC, Stadler PF (2009) Solvent exposure imparts similar selective pressures across a range of yeast proteins. Mol Biol Evol 26: 1155–1161.
19. Delport W, Scheffler K, Muse SV, Kosakovsky Pond S (in press) Benchmarking multi-rate codon models. PLoS One.
20. Sainudiin R, Wong WSW, Yogeeswaran K, Nasrallah JB, Yang Z, et al. (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.
21. Huelsenbeck JP, Joyce P, Lakner C, Ronquist F (2008) Bayesian analysis of amino acid substitution models. Philos Trans R Soc Lond B Biol Sci 363: 3941–3953.
22. Posada D, Crandall K (1998) Modeltest: testing the model of DNA substitution. Bioinformatics 14: 817–818.
23. Kosakovsky Pond SL, Frost SDW (2005) A genetic algorithm approach to detecting lineage-specific variation in selection pressure. Mol Biol Evol 22: 478–485.
24. Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SD (2006) Automated phylogenetic detection of recombination using a genetic algorithm. Mol Biol Evol 23: 1891–1901.
25. Kosakovsky Pond SL, Mannino FV, Gravenor MB, Muse SV, Frost SD (2007) Evolutionary model selection with a genetic algorithm: a case study using stem RNA. Mol Biol Evol 24: 159–170.
26. Posada D (2008) jModelTest: phylogenetic model averaging. Mol Biol Evol 25: 1253–1256.
27. Kosakovsky Pond S, Scheffler K, Gravenor M, Poon A, Frost S (2009) Evolutionary fingerprinting of genes. Mol Biol Evol 27: 520–536.
28. Kosakovsky Pond S, Delport W, Muse SV, Scheffler K (in press) Correcting the bias of empirical frequency parameter estimators in codon models. PLoS One.
29. Kosakovsky Pond SL, Muse SV (2005) Site-to-site variation of synonymous substitution rates. Mol Biol Evol 22: 2375–2385.
30. Stanfel L (1996) A new approach to clustering the amino acids. J Theor Biol 183: 195–205.
31. Hasegawa M, Kishino H, Yano T (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. Mol Biol Evol 21: 160–174.
32. Kosakovsky Pond SL, Frost SDW, Muse SV (2005) HyPhy: hypothesis testing using phylogenies. Bioinformatics 21: 676–9.
33. Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6: 461–464.
34. Atkinson AC (1980) A note on the generalized information criterion for choice of a model. Biometrika 67: 413–418.
35. Rand WM (1971) Objective criteria for the evaluation of clustering methods. J Amer Statist Assoc 66: 846–850.
36. Kosakovsky Pond SL, Posada D, Stawiski E, Chappey C, Poon AFY, et al. (2009) An evolutionary model-based algorithm for accurate phylogenetic breakpoint mapping and subtype prediction in HIV-1. PLoS Comput Biol 5: e1000581.
37. Bao Y, Bolotov P, Dernovoy D, Kiryutin B, Zaslavsky L, et al. (2007) The influenza virus resource at the national center for biotechnology information. J Virol 82: 596–601.
38. Yokoyama S, Tada T, Zhang H, Britt L (2008) Elucidation of phenotypic adaptations: Molecular analyses of dim-light vision proteins in vertebrates. Proc Natl Acad Sci U S A 105: 13480–13485.
39. Brumme ZL, Brumme CJ, Heckerman D, Korber BT, Daniels M, et al. (2007) Evidence of differential HLA class I-mediated viral evolution in functional and accessory/regulatory genes of HIV-1. PLoS Pathog 3: e94.
40. Rousseau CM, Daniels MG, Carlson JM, Kadie C, Crawford H, et al. (2008) HLA class I-driven evolution of human immunodeficiency virus type 1 subtype C proteome: immune escape and viral load. J Virol 82: 6434–6446.
41. Russell CA, Jones TC, Barr IG, Cox NJ, Garten RJ, et al. (2008) The global circulation of seasonal influenza A (H3N2) viruses. Science 320: 340–346.
42. Burnham K, Anderson D (2003) Model selection and multimodel inference. New York: Springer, 2nd ed. edition.
43. Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52: 696–704.
44. Tavaré S (1986) Some probabilistic and statistical problems in the analysis of DNA sequences. In: Miura RM, ed. Lectures on Mathematics in the Life Sciences. Providence, R.I.: Amer. Math. Soc. pp 57–86.
45. Yang Z (1993) Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol Biol Evol 10: 1396–1401.
46. Kosakovsky Pond SL, Frost SDW (2005) Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. Bioinformatics 21: 2531–2533.
47. Wong W, Sainudiin R, Nielsen R (2006) Identification of physicochemical selective pressure on protein encoding nucleotide sequences. BMC Bioinformatics 7: 148–158.
48. Whelan S, Goldman N (2004) Estimating the frequency of events that cause multiple-nucleotide changes. Genetics 167: 2027–2043.
49. Pupko T, Pe'er I, Shamir R, Graur D (2000) A fast algorithm for joint reconstruction of ancestral amino acid sequences. Mol Biol Evol 17: 890–896.
50. Kosakovsky Pond SL, Frost SDW (2005) Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol Biol Evol 22: 1208–1222.
51. Le SQ, Gascuel O (2008) An improved general amino acid replacement matrix. Mol Biol Evol 25: 1307–20.
52. Poon AFY, Lewis FI, Pond SLK, Frost SDW (2007) An evolutionary-network model reveals stratified interactions in the V3 loop of the HIV-1 envelope. PLoS Comput Biol 3: e231.
53. Rodrigue N, Kleinman CL, Philippe H, Lartillot N (2009) Computational methods for evaluating phylogenetic models of coding sequence evolution with dependence between codons. Mol Biol Evol 26: 1663–76.
54. Yang Z, Nielsen R, Goldman N, Pedersen AM (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155: 431–449.
55. Sonnhammer EL, Eddy SR, Durbin R (1997) Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins 28: 405–20.
56. Leslie A, Pfafferott K, Chetty P, Draenert R, Addo M, et al. (2004) HIV evolution: CTL escape mutation and reversion after transmission. Nat Med 10: 282–9.
57. Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4: 406–25.