Available online at www.sciencedirect.com

SCIENCE DIRECT

Gene 347 (2005) 207–217

GENE
SECTION
EVOLUTIONARY GENOMICS

www.elsevier.com/locate/gene

Site interdependence attributed to tertiary structure in amino acid sequence evolution

Nicolas Rodrigue[a,*], Nicolas Lartillot[b], David Bryant[c], Hervé Philippe[a]

[a]*Canadian Institute for Advanced Research, Département de biochimie, Université de Montréal, Montréal, Canada*
[b]*Laboratoire d'informatique, de robotique et de microélectronique de Montpellier, Montpellier, France*
[c]*McGill Centre for Bioinformatics, McGill University, Montreal, Canada*

Received 21 September 2004; received in revised form 15 November 2004; accepted 2 December 2004

Received by U. Bastollla

Abstract

Standard likelihood-based frameworks in phylogenetics consider the process of evolution of a sequence site by site. Assuming that sites evolve independently greatly simplifies the required calculations. However, this simplification is known to be incorrect in many cases. Here, a computational method that allows for general dependence between sites of a sequence is investigated. Using this method, measures acting as sequence fitness proxies can be considered over a phylogenetic tree. In this work, a set of statistically derived amino acid pairwise potentials, developed in the context of protein threading, is used to account for what we call the *structural fitness* of a sequence. We describe a model combining statistical potentials with an empirical amino acid substitution matrix. We propose such a combination as a useful way of capturing the complexity of protein evolution. Finally, we outline features of the model using three datasets and show the approach's sensitivity to different tree topologies.

Keywords: Protein evolution; Phylogenetics; Bayesian Markov chain Monte Carlo; Statistical potentials

1. Introduction

Probabilistic approaches in molecular phylogenetics rest on explicit mathematical models of sequence evolution. Given such a model, a maximum likelihood (ML) or a Bayesian framework is adopted, requiring the computation of the likelihood, i.e. the probability of observing a particular dataset of sequences, conditional on the phylogenetic hypothesis. An important simplification–among others–is typically invoked to render this computation tractable: the assumption that evolutionary events at a particular site are independent from events at other sites. Accordingly, each site is considered separately, so as to define an independent Markov substitution process along the branches of a tree. The Markov process is specified by a rate matrix, the entries of which indicate the instantaneous rate of substitution from one nucleotide (or amino acid, or codon) to another. Rate matrices are either purely theoretical (e.g. Jukes and Cantor, 1969; Kimura, 1980), or based on generalized empirical measurements (e.g. Dayhoff et al., 1978; Jones et al., 1992a). Among the extensions of the single rate matrix approach, it has been proposed to draw the rate of each site from a gamma distribution (Yang, 1993, 1994). Such a 'rate across sites' model provides an implicit way of taking into account varying selective pressures occurring at different sites. Models with site-specific matrices have also been studied (Bruno, 1996; Halpern and Bruno, 1998), as well as various types of mixture models (e.g. Koshi and Goldstein, 1995; Thorne et al., 1996; Koshi and Goldstein, 1997; Goldman et al.,

Abbreviations: MCMC, Markov chain Monte Carlo; ML, Maximum likelihood; PDB, Protein data bank.

* Corresponding author. Département de biochimie, C.P. 6128, Succ. Centre-ville, Montréal, Québec, Canada H3C 3J7. Tel.: +1 514 343 6111x5091; fax: +1 514 343 2210.

E-mail address: nicolas.rodrigue@umontreal.ca (N. Rodrigue).

0378-1119/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2004.12.011

1998; Lartillot and Philippe, 2004; Pagel and Meade, 2004).

It is widely agreed that the assumption of site independence is not biologically sound. Consequently, means of relaxing this assumption have been pursued, usually with correlations or dependence introduced between a limited number of sites (e.g. Felsenstein and Chruchill, 1996; Siepel and Haussler, 2004), or considered for a limited number of sequences (Jensen and Pedersen, 2000; Pedersen and Jensen, 2001; Robinson et al., 2003). In particular, when considering pairs of coding nucleotide sequences, Robinson et al. (2003) have recently introduced a sampling technique with a model allowing for dependence between codons. With their sampling procedure, one can consider the stochastic process underlying the evolution of a sequence as a whole, so that the probability of a given substitution, at any given time and at any site, depends, in principle, on the state at all other positions. Robinson et al. measure the composition of site interdependencies using an empirical *energy function* (otherwise known as *statistical potentials*) derived in the context of protein threading (Jones et al., 1992b).

The central idea behind a wide variety of statistical potentials (e.g. Miyazawa and Jernigan, 1985; Hendlich et al., 1990; Jones et al., 1992b; Bastolla et al., 2001) is to associate pseudo-energy terms to the plausibility of having a given pair of amino acids at a particular spatial proximity, as learned from a database of protein sequences of known structure. Lower pseudo-energy values correspond to typical pairwise amino acid interactions, while higher values correspond to atypical interactions. The measure provided by statistical potentials can be said to capture the approximate *structural fitness* of a sequence, since it has been optimized to detect amino acid patterns with regards to the 3-dimensional structure of natural sequences (i.e. sequences that have been fixed by selection). In contrast with statistical potentials, empirical amino acid replacement matrices, such as the JTT matrix (Jones et al., 1992a), are optimized based on how frequently two given amino acids exchange with one another in homologous positions of a sequence database.

Here, we build on the ideas of Robinson et al. (2003) and propose a formulation at the amino acid level. We describe a model that allows for dependence between sites using statistical potentials (Bastolla et al., 2001) while also making use of the information available in an empirical amino acid replacement matrix (Jones et al., 1992a). We suggest that capturing the complexity of protein evolution may be best served by *layering* both varieties of empirical measurements. Furthermore, we generalize the sampling technique of Robinson et al. (2003) from two taxa to n taxa, thus introducing the approach in a broader phylogenetic context. In line with Robinson et al., we deal with the computational ramifications of structural fitness considerations by adopting a Bayesian Markov chain Monte Carlo (MCMC) approach to sample detailed *substitution histories* along a given phylogenetic tree from their posterior distribution. Following a similar framework to that used in previous studies (e.g. Parisi and Echave, 2001; Bastolla et al., 2003; Robinson et al., 2003), it is assumed that the tertiary structure of a protein is well conserved and, for our purposes, remains identical at all points along the tree. We illustrate features of the model when applied to three protein sequence alignments and prospect the possibility of applying the approach to the comparison of different tree topologies.

2. Material and methods

2.1. Datasets, trees and protein structures

For computational reasons, we have restrained our analyses to relatively small datasets, with a number of taxa ranging from 4 to 10. We have also focused on small monomeric proteins, both to lighten computational requirements and because the energy function used is known to perform better in such cases (Bastolla et al., 2001).

- *PPK10-158*: This dataset is composed of 10 amino acids sequences (with 158 positions) of bacterial 6-hydroxymethyl-7-8-dihydroxypterin pyrophosphokinase. The species (with GenBank accession numbers) are *Escherichia coli* (BAB9671), *Shigella flexneri* (AAP15678), *Salmonella typhimurium* (AAL19147), *Photorhabdus luminescens* (CAE13168), *Yersinia pestis* (AAS60560), *Erwina carotovora* (CAG76218), *Vibro vulnificus* (BAC95526), *Azotobacter vinelandii* (ZP_00091220), *Wigglesworthia glossinidia* (BAC24410) and *Coxiella burnetii* (AA089845).
- *MYO10-153*: This is a 10 species dataset of mammalian myoglobin amino acid sequences (with 153 positions). The species are *Physester catodon* (P02185), *Orcinus orca* (P02173), *Bos taurus* (BAA00311), *Rattus norvegicus* (AAF05848), *Mus musculus* (CAA27994), *Nannospalax ehrenbergi* (P04248), *Homo sapiens* (CAA25109), *Gorilla gorilla* (P02147), *Ornithorhynchus anatinus* (P02196) and *Tachyglossus aculeatus* (P02195).
- *MYO4-153*: This is also a dataset of myoglobin sequences, here taken from the 4 species *P. catodon* (P02185), *O. orca* (P02173), *Graptemys geographica* (P02201) and *Chelonia mydas caranigra* (MYTTG).

We worked under a fixed tree topology for all datasets. Topologies were obtained by ML, under a JTT+F model, with gamma+invariant distributed rates across sites, using the PhyML program (Guindon and Gascuel, 2003).

Protein structures are assumed constant throughout the tree. In practice, we used as a reference the structure of one of the sequences in the dataset, as determined by X-ray crystallography. The structure of *E. coli* (PDB code: 1HKA) was used as a reference for the *PPK10-158* dataset, and that of *P. catodon* (PDB code: 1MBD) for both *MYO10-153* and *MYO4-153* datasets. Imposing the structure of a reference sequence on other sequences is simplified when these are of identical length. Therefore, we constructed alignments without gaps. This was accomplished using sequences of the same length as the reference sequence–which was the case for *MYO10-153* and *MYO4-153*–or, in the case of *PPK10-158*, including only insertions with respect to the reference sequence (such positions were removed, leading to a gapless alignment).

2.2. General overview and notation

Standard likelihood computations in phylogenetics (Felsenstein, 1981), involving rate matrix exponentiation, correspond to the integral over all possible substitution histories, or mappings (Nielsen, 2002), over a given phylogenetic tree. We make use of this equivalence by *sampling substitution histories directly*. Within this framework, matrix exponentiation is avoided, allowing one to consider rate matrices of much higher order. In particular, when taking into account an entire amino acid sequence of length N, the rate matrix has an order of $20^N \times 20^N$ (clearly ruling out matrix exponentiation for all practical purposes).

We use b to index branches of the tree. By convention, a node has the same index as the branch that leads to it (except the root node, which has index 0). We refer to the set of branch lengths as β. A substitution history, denoted ω, includes the time and nature of each substitution event along branches. We index substitution events with z and symbolize the time of substitution event z as $t(z)$. The length of branch b is written as β_b, and its substitution history is denoted ω_b. Datasets consist of alignments of P amino acid sequences. An amino acid sequence is written as $s^{(\cdot)}$, using the superscripted parentheses to describe the context of a sequence as follows:

- the sequence at the root of the tree, written as $s^{(\text{root})}$;
- the sequences at other nodes, referring to the ancestral and descendant sequences found at the nodes at the ends of branch b respectively as $s^{(b\rightarrow \text{up})}$ and $s^{(b\rightarrow \text{down})}$;
- the sequence before and after a substitution event z, written as $s^{(z-1)}$ and $s^{(z)}$ respectively. When considering a series of q_b substitution events over branch b, we set $s^{(0)}=s^{(b\rightarrow \text{up})}$, and after the final substitution event along branch b, $s^{(q_b)}=s^{(b\rightarrow \text{down})}$.

Finally, we refer to the specific amino acid found at position i of sequence $s^{(x)}$ by writing $s_i^{(x)}$.

2.3. Estimating structural fitness

We used the knowledge-based protein energy function described in Bastolla et al. (2001) to estimate the structural fitness of a sequence in a given three-dimensional structure. Our use of the energy function is straightforward. Given a PDB file, one computes the distances between all atoms of all amino acids. As defined by Bastolla et al., two amino acids are said to be in contact if any of their heavy atoms (atoms other than hydrogen) are at a distance of 4.5 Å or less (contacts due to sequential proximity–within three positions or less–are ignored). As such, the structure of a protein can be represented as a contact map. The contact map of a protein structure of length N is an $N\times N$ matrix C with elements:

$$C_{i,j} = \begin{cases} 1 & \text{if amino acids at sites } i \text{ and } j \text{ are in contact,} \\ 0 & \text{if amino acids at sites } i \text{ and } j \text{ are not in contact, or if } |i-j|\leq 3. \end{cases} \quad (1)$$

Given a contact map, we evaluate the pseudo-energy of a protein as follows:

$$E\left(s^{(x)}\right) = \sum_{1\leq i\leq j\leq N} C_{i,j}\varepsilon_{s_i^{(x)},s_j^{(x)}}, \quad (2)$$

where the coefficients $\varepsilon_{lm,\ (1\leq l,m\leq 20)}$ is the *pair potential* matrix of Bastolla et al.

We impose the same structure throughout the phylogenetic tree by using the same contact map on all sequences, both observed and inferred.

2.4. Evolutionary model

Conventional models of protein evolution assume that an independent Markov process operates at each site, under an identical 20×20 rate matrix Q. The matrix Q is specified by 20 amino acid stationary probabilities (π_m), $1 \leq m \leq 20$, $\sum_{m=1}^{20} \pi_m = 1$, and exchangeability parameters (ρ_{lm}), $1 \leq l,m \leq 20$, such that

$$Q_{lm} = \frac{1}{Z} \rho_{lm} \pi_m, l \neq m \tag{3}$$

$$Q_{ll} = -\sum_{m \neq l} Q_{lm}, \tag{4}$$

where Z is a normalization constant so that branch lengths represent the expected number of substitutions per site:

$$Z = 2 \times \sum_{1 \leq l < m \leq 20} \rho_{lm} \pi_m. \tag{5}$$

One consequence of such models is that the probability that two sites undergo a substitution at the exact same moment vanishes. As a result, the instant rate of substitution between two sequences differing at more than one position is 0. Hence, site-specific Markov processes can, in principle, be combined to consider the evolution of the sequence as a whole, which therefore also operates under a Markov process, now described by a $20^N \times 20^N$ matrix R:

$$R_{s^{(x)},s^{(y)}} = \begin{cases} 0 & \text{if } s^{(x)} \text{ and } s^{(y)} \text{ differ at more than one position,} \\ Q_{lm} & \text{if } s^{(x)} \text{and } s^{(y)} \text{ differ only at site } i, s_i^{(x)} = l \text{ and } s_i^{(y)} = m, \\ -\sum_{s^{(y)} \neq s^{(x)}} R_{s^{(x)},s^{(y)}} & \text{if } s^{(x)} \text{ and } s^{(y)} \text{ are identical.} \end{cases} \tag{6}$$

Following Robinson et al. (2003), the underlying theme of our model is that amino acid substitution rates are likely to depend, at least in part, on how the amino acid substitutions in question affect the structural fitness of a sequence. The approach adopted is to complement current models of protein evolution, specified by Q, with a supplementary factor that takes into account the pseudo-energy difference before and after an amino acid substitution. This difference is weighted by a parameter p. Our rate matrix R becomes:

$$R_{s^{(x)},s^{(y)}} = \begin{cases} 0 & \text{if } s^{(x)} \text{ and } s^{(y)} \text{ differ at more than one position,} \\ Q_{lm} e^{p\left(E\left(s^{(x)}\right) - E\left(s^{(y)}\right)\right)} & \text{if } s^{(x)} \text{and } s^{(y)} \text{ differ only at site } i, s_i^{(x)} = l \text{ and } s_i^{(y)} = m, \\ -\sum_{s^{(y)} \neq s^{(x)}} R_{s^{(x)},s^{(y)}} & \text{if } s^{(x)} \text{ and } s^{(y)} \text{ are identical.} \end{cases} \tag{7}$$

Note that since the energy function considers the sequence as a whole, the process defined by $R(20^N \times 20^N)$ can no longer be decomposed into an independent process operating at each site, defined by Q (20×20), except in the case where p=0 (which simplifies to the model assuming independence in Eq. (6)). One of the goals of this work is to obtain probabilistic estimates of p. Positive values of p would have selection favoring substitution events that lead to typical amino acid interactions. Negative values of p would lead to the converse effect, increasing the rate of substitution toward atypical interactions, under the given structure.

The stationary probabilities of amino acids (π_m), $1 \leq m \leq 20$, included in Q, are free parameters of the model. The exchangeability terms (ρ_{lm}), $1 \leq l,m \leq 20$, could also be free parameters. However, for simplicity we have fixed these parameters to predefined values. In this study, we considered two cases for the exchangeability parameters:

- All exchangeability terms equal: the Poisson process;
- Exchangeability terms set to those of the JTT empirical matrix (Jones et al., 1992a).

2.5. Likelihood computations

For notational simplicity, the free parameters of the model are grouped into a vector denoted as θ (i.e. $\theta = \{p, \pi_{m,\ (1 \leq m \leq 20)}\}$). Likelihood computations require the probabilities of transition from one sequence state to the next. Over a branch b of length β_b, the probability of going from $s^{(b \to \text{up})}$ to $s^{(b \to \text{down})}$, given the evolutionary model parameters θ, can be calculated as

$$p\left(s^{(b \to \text{down})} | s^{(b \to \text{up})}, \beta_b, \theta\right) = \left[e^{\beta_b R}\right]_{s^{(b \to \text{up})}, s^{(b \to \text{down})}} = \int_{\Omega_b} p\left(s^{(b \to \text{down})}, \omega_b | s^{(b \to \text{up})}, \beta_b, \theta\right) d\omega_b. \tag{8}$$

As shown by Eq. (8), this probability is an integral over all possible substitution histories (Ω_b) having $s^{(b\rightarrow \text{up})}$ and $s^{(b\rightarrow \text{down})}$ as their initial and final states respectively. This equivalence suggests sampling substitution histories directly as a tractable alternative to matrix exponentiation. Robinson et al. (2003) have derived the calculations for evaluating the probability of a substitution history:

$$p\left(s^{(b\rightarrow \text{down})}, \omega_b | s^{(b\rightarrow \text{up})}, \beta_b, \theta\right) = \left(\prod_{z=1}^{q_b} R_{s^{(z-1)},s^{(z)}} e^{-R_{s^{(z-1)},s^{(\cdot)}}\left(t(z)-t(z-1)\right)} \right) e^{-R_{s(q_b),s^{(\cdot)}}(\beta_b - t(q_b))}, \tag{9}$$

where $e^{-R_{s(q_b),s^{(\cdot)}}(\beta_b - t(q_b))}$ accounts for the probability of no events occurring from the last event q to the end of the branch, and where $R_{s^{(x)},\ s^{(\cdot)}}$ represents the rate away from $s^{(x)}$. The rate away from $s^{(x)}$ can be calculated from

$$R_{s^{(x)},s^{(\cdot)}} = \sum_k R_{s^{(x)},s^{(k)}}, \tag{10}$$

the sum being over all sequences $s^{(k)}$ of length N different than $s^{(x)}$. Since $R_{s^{(x)},s^{(k)}}$ is equal to zero for all sequences $s^{(k)}$ that differ with $s^{(x)}$ at more than one position, only $19N$ non-zero elements need be considered in this summation.

Eq. (9) provides a means of evaluating probabilities of substitution histories along a branch. Using this in the context of a phylogenetic tree requires proposing sequences at internal nodes of the tree and applying Eq. (9) to each branch. Assuming that lineages evolve independently allows one to evaluate the substitution history of the entire tree by taking the product of Eq. (9) over all branches. To complete the likelihood computations, one must calculate the stationary probability of the sequence at the root of the tree, given the evolutionary model. This probability is given by Robinson et al. (2003):

$$p\left(s^{(\text{root})}|\theta\right) = \frac{e^{-2pE\left(s^{(\text{root})}\right)} \prod_{m=1}^{N} \pi_{s_m^{(\text{root})}}}{\sum_k e^{-2pE\left(s^{(k)}\right)} \prod_{n=1}^{N} \pi_{s_n^{(k)}}}, \tag{11}$$

where $\pi_{s_m^{(\text{root})}}$ is the stationary probability of the amino acid at position m of sequence $s^{(\text{root})}$, and where the sum in the denominator is over all possible sequences of length N. In practice, since the model is reversible, one can root the phylogenetic tree at any arbitrary point. We opted to root the tree at an external node (i.e. an observed sequence), which therefore remains fixed.

Combining Eqs. (9) and (11) yields the overall likelihood term:

$$p(D|\omega,\theta,\beta) = p\left(s^{(\text{root})}|\theta\right) \prod_{b=1}^{2P-2} p\left(s^{(b\rightarrow \text{down})}, \omega_b | s^{(b\rightarrow \text{up})}, \beta_b, \theta\right). \tag{12}$$

2.6. Monte Carlo sampling

This work adopts a Bayesian MCMC framework. Within this framework, ω, θ and β are sampled from their joint posterior distribution $p(\omega, \theta, \beta|D)$, where D is the dataset of P observed sequences. According to Bayes' theorem, this probability can be written as

$$p(\omega,\theta,\beta|D) = \frac{p(D|\omega,\theta,\beta)p(\theta,\beta)}{p(D)}, \tag{13}$$

where $p(\theta, \beta)$ is the joint prior probability density of model parameters and branch lengths, and $p(D)$ is a normalization factor such that the total probability equals 1. Our MCMC procedure then consists of defining a Markov chain, with state space the set of admissible values of (θ, β, ω), and having the posterior probability defined in Eq. (13) as its stationary distribution. This is done using the Metropolis–Hastings algorithm (Metropolis et al., 1953; Hastings, 1970): assuming the current state of the Markov chain is (θ, β, ω), a move to different model parameter values (θ'), branch lengths (β') or substitution histories (ω') is proposed (note that a substitution history move can also include proposing an alternative state, $s'^{(b\rightarrow up)}$, at an internal node). The move is accepted with a probability of r, where

$$r = \min\left(1, \frac{p(D|\omega',\theta',\beta')p(\theta',\beta')}{p(D|\omega,\theta,\beta)p(\theta,\beta)} \times h\right), \tag{14}$$

and where

$$h = \prod_{b=1}^{2P-2} \left[\frac{J\left(s^{(b\rightarrow \text{up})}, \omega_b, \theta, \beta_b | s'^{(b\rightarrow \text{up})}, \omega_b', \theta', \beta_b'\right)}{J\left(s'^{(b\rightarrow \text{up})}, \omega_b', \theta', \beta_b' | s^{(b\rightarrow \text{up})}, \omega_b, \theta, \beta_b\right)} \right] \tag{15}$$

is the ratio of the proposal densities (here denoted using J) known as the Hastings ratio (Hastings, 1970), which effectively corrects for biases in the proposal methods. If the move is accepted, θ, β and ω are set to those proposed; otherwise, they are left as-is. This procedure is reiterated a large number of times until convergence to the posterior distribution, from which we then sample.

2.7. Priors

We used uninformative priors for the amino acid stationary probabilities (π_m), $1 \leq m \leq 20$, $\sum_{m=1}^{20} \pi_m = 1$, branch lengths ($\beta_b$) and for the weight accorded to the protein structure considerations (p):

- $\pi_m \sim$Dirichlet(1,1,. . .,1);
- $\beta_b \sim$Uniform[0,100];
- $p \sim$Uniform[−5, 5]; the reasons for the narrowness of the interval are explained in Section 2.9.

2.8. Proposing substitution histories ω

Substitution histories are proposed site by site. The method used is an implementation of Nielsen's (2002) method for mapping substitutions along a tree using a model of evolution that assumes independence between sites, here denoted by a rate matrix Q^*. We set the exchangeability parameters of Q^* to those found in Eq. (7), while the amino acid stationary probabilities are fixed to the observed empirical frequencies (i.e. Q^*=JTT+F or Poisson+F). Two types of moves are used to propose new substitution histories ω'. For both moves, θ and β are kept constant.

•BRANCHHISTORY: this first type of move randomly selects a branch b on the tree and a site i among all sites N. A new substitution history for site i along branch b is re-sampled as in Nielsen's method, given the states $s_i^{(b\rightarrow \text{up})}$ and $s_i^{(b\rightarrow \text{down})}$. The move is then accepted with a probability given in Eq. (14). The corresponding Hastings ratio h is simply the probability of the substitution history according to the model that assumes independence before the move over that of the proposed substitution history, written as

$$h = \frac{p\left(s_i^{(b\rightarrow \text{down})}, \omega_b | s_i^{(b\rightarrow \text{up})}, \beta_b, Q^*\right)}{p\left(s_i^{(b\rightarrow \text{down})}, \omega_b' | s_i^{(b\rightarrow \text{up})}, \beta_b, Q^*\right)}. \tag{16}$$

•NODESTATE: the second move randomly selects an internal node b and a site i. The move then re-samples the amino acid state of site i at node b, again using Nielsen's method. Having re-sampled the state, the move also re-samples a substitution history for site i over the three branches connected to node b. The Hastings ratio for this move is the same as Eq. (16), but multiplied over the three branches in question.

2.9. Proposing θ and β

We used one move for proposing π_m', $1 \leq m \leq 20$, while keeping p, β and ω fixed, and one move for proposing p', while keeping π_m, $1 \leq m \leq 20$, β and ω fixed.

•STATIONARY: amino acid stationary probability moves are proposed according to a Dirichlet distribution centered on their current value and with a tuning parameter τ_A, as described in Larget and Simon (1999).

•STRUCTURE: proposing p' is accomplished by adding τ_S $(U-0.5)$ to p, where U is a random variable drawn from a uniform distribution in the interval [0,1] and $\tau_S > 0$ is a tuning parameter. The Hastings ratio is 1.

Evaluating the proposed set of stationary probabilities π_m', $1 \leq m \leq 20$, or a proposed p', raises a significant complication, since in these moves, $p(s^{(\text{root})}|\theta')/p(s^{(\text{root})}|\theta)$ does not cancel out in Eq. (14). This complication arises from the fact that in Eq. (11), the sum in the denominator is over all possible sequences of length N. Robinson et al. (2003) provide an approximation strategy re-implemented for this work. The strategy rests on sampling a group of M sequences, denoted as $\eta^{(1)}, \eta^{(2)}, \ldots, \eta^{(M)}$, from the stationary distribution of sequences for a third set of parameter values θ^*. For sufficiently large values of M, Robinson et al.'s importance sampling argument can be applied to this model to yield

$$\frac{p\left(s^{(\text{root})}|\theta'\right)}{p\left(s^{(\text{root})}|\theta\right)} \approx e^{-2(p'-p)E\left(s^{(\text{root})}\right)} \left(\prod_{m=1}^{N} \frac{\pi_{i_m}'}{\pi_{i_m}}\right) \left(\frac{\sum_{h=1}^{M} e^{-2(p-p^*)E(\eta^{(h)})} \prod_{n=1}^{N} \frac{\pi_{\eta_n^{(h)}}}{\pi^*_{\eta_n^{(h)}}}}{\sum_{h=1}^{M} e^{-2(p'-p^*)E(\eta^{(h)})} \prod_{n=1}^{N} \frac{\pi'_{\eta_n^{(h)}}}{\pi^*_{\eta_n^{(h)}}}}\right). \tag{17}$$

This approximation's quality depends on two factors: the value of M (high values improve the approximation) and the distance of θ^* to both θ and θ' (a θ^* at the midpoint between θ and θ' gives the best approximation). Robinson et al. opt to

partition their parameter space into a predefined grid. They then use the grid point θ^* that is nearest to the midpoint of θ and θ'. The strategy employed here is different. Our protocol creates new θ^*s dynamically, always at the midpoint of θ and θ'. A new θ^* is created whenever the distance (λ) between the midpoint of θ and θ', and the nearest θ^* is beyond a predefined threshold (λ_{max}). In practice, a limit is set on the number of θ^* stored in memory. Whenever this limit is reached, and a new θ^* is to be created, one simply writes over the θ^* (and the respective M sequences) that is the furthest away from the midpoint of θ and θ'. As such, one eventually has a 'hyper-cloud' of θ^*s following θ and θ' as the MCMC run progresses. We determined empirically the acceptable settings for this approximation procedure, fixing λ_{max}=0.01 and M=1000 (see Supplementary material). However, a larger λ_{max} and a lower M can be used to obtain faster rough estimates.

Restraining the interval of the uniform distribution used as the prior for p serves to increase the speed of convergence. An overly wide interval could lead to initial values that are very far from those at stationarity, which would require invoking the approximation procedure for $p(s^{(root)}|\theta')/p(s^{(root)}|\theta)$ many times before convergence.

Branch lengths are proposed one branch at a time, while keeping θ and ω constant.

•BRANCHLENGTH: a branch b is randomly selected from the tree. The length of branch b, as well as the times of each substitution event along b, is then multiplied by $\upsilon=e^{\tau_{BL}(U-0.5)}$, where $\tau_{BL}>0$ is another move-specific tuning parameter. The Hastings ratio in this case is υ^{df}, where degrees of freedom (df) is equal to 1 plus the number of substitution events along that branch.

2.10. General MCMC settings and implementation checks

In the course of an MCMC run, moves are called according to a set of weights (w_ψ), $1\leq\psi\leq\Psi$, where Ψ is the total number of possible moves. We thereby define a *cycle* as a set of W iterations, with $W=\sum_{\psi=1}^{\Psi} w_\psi$. We determined the weights (w_ψ) empirically, as well as all move-specific tuning parameters, so as to optimize mixing (here W=112, see Supplementary material). For the results presented here, we ran each chain for 100,000 cycles ($W\times$100,000=11,200,000 iterations), discarded the first 10,000 cycles as burn-in, and sub-sampled every 50 cycles from the remaining sample. The MCMC runs require 10–15 days of CPU time on a Xeon 2.4 GHz desktop computer.

When the parameter p=0, our model simplifies to the site independent JTT+π model (or Poisson+π, depending on the exchangeability parameters chosen). We tested our implementation with p=0 and compared it with the results of a standard JTT+π (Poisson+π) implementation (Lartillot and Philippe, 2004) and found both to converge to essentially identical parameters and branch lengths at stationarity (data not shown). We also verified that when p=0 and Q=Q^*, all substitution history moves are accepted (since in this case, Eq. (14) simplifies to 1), and the mean number of substitutions sampled per site indeed corresponds to the expected number of substitutions per site (data not shown, but see Nielsen, 2002).

3. Results and discussion

3.1. Exchangeability parameters in relation to structural fitness considerations

We applied our model to a first data set, *MYO10-153*, using the ML topology as a constraint (see Section 2.1). We performed several independent runs, starting from different initial values for the parameters, and found that the parameter p consistently stabilizes around the same value (Fig. 1). Additionally, p converges to positive values across all datasets, indicating that selection prefers sequence substitution histories that maintain a good structural fitness (Table 1). These results corroborate with those of Robinson et al. (2003). Interestingly, we note that p consistently stabilizes at higher values when the exchangeability parameters in Q are uniform (Poisson) than when they are set to those of the JTT empirical matrix (Table 1). For example, for the *MYO10-153* dataset, the mean posterior values obtained (with 95% credibility intervals) are p=0.7005 (0.5876, 0.8164) and p=0.6273 (0.5042, 0.7386) when using the Poisson and JTT exchangeability parameters respectively. A z-test indicates that this difference is significant, at a 99% confidence level.

Being empirically derived, the JTT matrix has a considerable amount of prior biochemical information regarding the amino acid substitution process. Accordingly, these results seem to indicate that, despite being a formally site independent model, the JTT matrix implicitly captures, to some extent, the average effects of dependencies between sites measured by the energy function. Hence, the weight accorded to structural fitness considerations need not be as high when using the JTT matrix in comparison to that when using the naive Poisson matrix.

3.2. Amino acid stationary probabilities and branch lengths

The substitution process, as specified in Eq. (7), can be viewed as a composition of two layered elements: (1) a process proposing substitutions, according to Q_{lm} and on the basis of the current branch length values; (2) a process selecting substitutions, by accepting or refusing according to $e^{p(E(s^{(x)})-E(s^{(y)}))}$. Consequently, the amino acid stationary probabilities are those of the substitution process *in the*

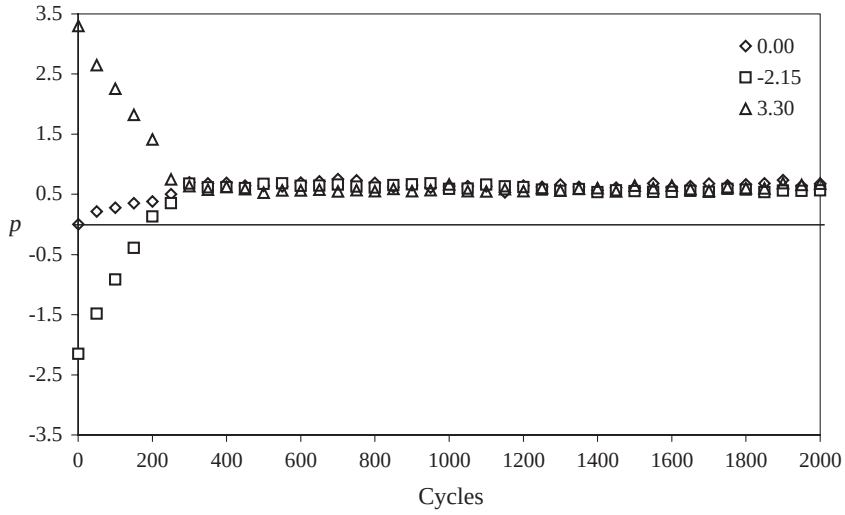

Fig. 1. Stabilization of p in three separate MCMC runs for the *MYO10-153* dataset with exchangeability parameters (ρ_{lm}), $1 \leq l,m \leq 20$ set those of the JTT matrix (first 2000 cycles, with points shown every 50). Three different initial values were used, indicated by the legend keys.

absence of the selection step. The energy function itself captures some elements of amino acid stationary probabilities, thereby creating an interplay between p and π_m, $1 \leq m \leq 20$. A better measure of the true (or actual) prevalence of each of the 20 amino acids is obtained by looking at the *induced* amino acid frequencies in a set of sequences sampled from θ. To monitor the induced frequencies, we found it convenient to simply look at the relative frequencies of amino acids in the sequences sampled from θ^* (see Section 2.9), as this parameter vector is always in the vicinity of the θ to θ' proposal. When p is fixed (p=0), sequences are directly sampled according to π_m, $1 \leq m \leq 20$, and the stationary probabilities and induced frequencies of amino acids are necessarily equivalent (see Table 2 of the Supplementary material). When p is a free parameter ($p \neq 0$), the stationary probabilities often differ widely with those when p=0 (Fig. 2). However, we found that the induced frequencies with $p \neq 0$ have only mild differences with those when p=0 (or with the empirical frequencies; see Fig. 2).

Likewise, branch lengths correspond to the expected number of substitutions per site proposed upstream of the selection step described above, and therefore, do not reflect the true branch lengths (i.e. the number of substitutions having actually occurred once the statistical potential has been taken into account, which we call the number of substitutions *induced* by the model). The sampling scheme makes it easy to look directly at the number of substitutions per site induced by the model. As would be expected, we found that these two measures of branch lengths do not correspond when $p \neq 0$, with the induced number of substitutions consistently lower. Using the same *MYO10-153* dataset as an example, we found that the tree length inferred with $p \neq 0$ was 1.0874 (0.8926, 1.3045) whereas the induced number of substitutions per site was 0.7779 (0.7255, 0.8434), a value only slightly higher to that with p=0, at 0.7678 (0.7190, 0.8235).

3.3. Sensitivity to tree topology

Using the *MYO4-153* dataset, we compared results of MCMC runs under each of the three possible topologies, focusing on the factor $\Pi_{b=1}^{2P-2} p\left(s^{(b\rightarrow \text{down})}, \omega_b | s^{(b\rightarrow \text{up})}, \beta_b, \theta\right)$ of Eq. (12). We refer to these comparisons as *likelihood comparisons*, although they are not true likelihood comparisons, since they ignore the factor $p(s^{(\text{root})}|\theta)$. We found, however, that the parameter estimates under each topology were essentially identical, so that the factor $p(s^{(\text{root})}|\theta)$ is effectively equivalent under each run. In any case, the comparisons provide an interesting contrast: the correct tree topology (grouping the two whales together and the two turtles together) is indeed the most favored tree (see Fig. 3). Although the correct topology is also the most favored tree when p=0 (data not shown), this demonstrates that the computational approach (i.e. sampling substitution histories) is indeed sensitive to tree topologies and could

Table 1
Mean values of p at stationarity for the three datasets under study

Exchangeability parameters in Q	Dataset		
	PPK10-158	*MYO10-153*	*MYO10-4*
Poisson	0.4207 (0.3300, 0.4994)	0.7005 (0.5876, 0.8164)	0.6901 (0.6141, 0.7913)
JTT	0.3613 (0.2759, 0.4358)	0.6273 (0.5042, 0.7386)	0.5717 (0.4804, 0.6555)

Parentheses indicate 95% credibility intervals.

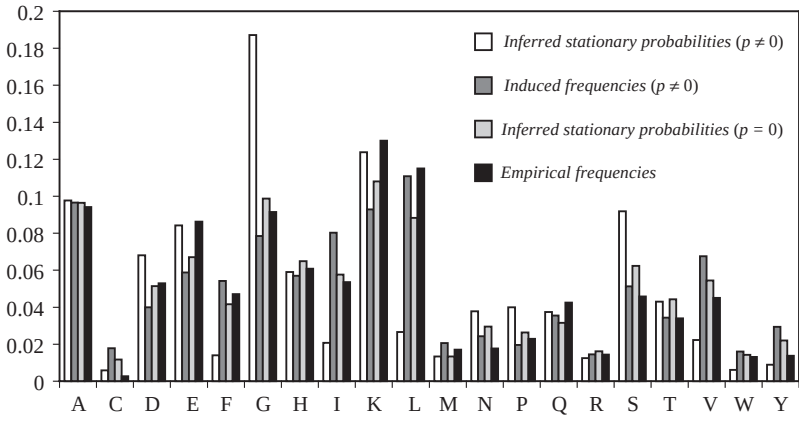

Fig. 2. Mean values inferred for π_m, $1 \leq m \leq 20$ at stationarity with $p=0$ and $p \neq 0$, as well as the induced frequencies with $p \neq 0$ for the *MYO10-153* dataset. Empirical frequencies are also shown. Exchangeability parameters (ρ_{lm}), $1 \leq l,m \leq 20$, are those of the JTT matrix (see Table 2 of the Supplementary material for numerical values and credibility intervals).

potentially be applied to address more difficult phylogenetic questions.

3.4. Perspectives

In this work, we have combined statistical potentials and an empirical amino acid replacement matrix in a model applied over phylogenetic trees. The model can be viewed as having two *layers*: one layer of underlying parameters that assume site-independence, specified by Q, and a second layer accounting for site interdependence, weighted using a parameter p. Here, we have found that the weight accorded to structural fitness considerations (p) is lower when using a more reasonable matrix Q (i.e. JTT+π) than when using a less reasonable matrix Q (Poisson+π). Conversely, an ideal measure of sequence fitness would render the use of underlying site-independent parameters inconsequential. Indeed, such an ideal measure would make the codon-based formulation of Robinson et al. (2003) a more appealing explicative model of the substitution process. In practical phylogenetics, however, using both layers may be valuable, since each is based on approximations, which may be capturing inherently different complexities of protein evolution.

It would be interesting to further explore the relation between these two layers. We hope to examine this relation by first using different underlying parameters, which could include a mixture of different matrices Q (e.g. Lartillot and Philippe, 2004), with, or without, gamma rate heterogeneity (Yang, 1993, 1994). In addition, we expect to study the impact of different existing statistical potentials (e.g. Miyazawa and Jernigan, 1985; Jones et al., 1992b; Singh et al., 1996) as well as developing a new set specifically tailored to our purposes. One particular problem we would like to address is the slight bias for hydrophobic amino acids in the induced frequencies with $p \neq 0$ with respect to the empirical frequencies (Fig. 2). This may be a result of the fact that we ignore the differential in stability with respect to alternative 'decoy' structures, as the energy function prescribes for the protein-threading context (Bastolla et al., 2001). We are currently constructing new potentials, optimized in an amino acid replacement context on predefined structures, rather than in the fold recognition framework.

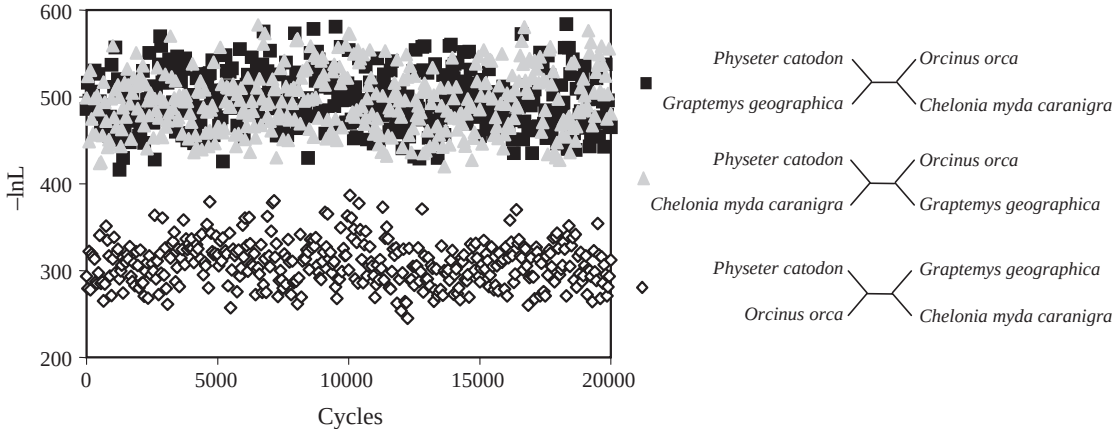

Fig. 3. Comparison of −lnL (where $L = \prod_{b=1}^{2P-2} p\left(s^{(b\rightarrow \text{down})}, \omega_b | s^{(b\rightarrow \text{up})}, \beta_b, \theta\right)$) for the three possible topologies of the *MYO4-153* dataset. Legend keys indicate the fixed topology used for each of the three MCMC runs. A window of 20,000 cycles is shown, with points every 50. Burn-in cycles were removed.

In all cases, it will be important to assess the performances of the model proposed here, as well as the possible combinations mentioned above. In the Bayesian framework, this is most often achieved by computing the Bayes factor (Jeffreys, 1935, 1961; Jaynes, 2003) between alternative models. For this purpose, we are currently adapting the technique of Bayes factor evaluation by thermodynamic integration described in Lartillot and Philippe (2004) to compare (two-layered) site-interdependent models to (single-layered) site-independent models. Using any particular site-independent model as a reference would further allow the assessment of different sequence fitness proxies applied to this context, and their relevance to the datasets of interest.

Finally, we have prospected the idea of using this approach to explore alternative topologies. Letting topologies be free parameters of the inference may be technically complex. The main complication arises from the fact that a rearrangement of the tree means that the current substitution history may not be compatible with the newly proposed topology. This raises the difficult problem of devising update mechanisms that simultaneously change the topology and the substitution history, while having a good acceptance rate; a task that would certainly be computationally very demanding. We are currently exploring a more reasonable alternative, which consists of using thermodynamic integration for evaluating the Bayes factor in support of one tree versus another. This would allow for the comparison of a set of pre-specified topologies, and in this way, assess the impact of the site-interdependent scheme proposed here on phylogenetic inference.

Acknowledgments

We thank Henner Brinkmann, Frédéric Delsuc and two anonymous referees for critical comments on the manuscript, as well as Ugo Bastolla for sharing potentials matrices and useful practical discussions. NR was supported by a bioinformatics grant from Génome Québec, and HP by the Canada Research Chair Program and the Université de Montréal.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.gene.2004.12.011.

References

Bastolla, U., Farwer, J., Knapp, E.W., Vendruscolo, M., 2001. How to guarantee optimal stability for most representative structures in the protein data bank. Proteins 44, 79–96.

Bastolla, U., Porto, M., Roman, E., Vendruscolo, M., 2003. Connectivity of neutral networks, overdispersion, and structural conservation in protein evolution. J. Mol. Evol. 56, 243–254.

Bruno, W.J., 1996. Modeling residue usage in aligned protein sequences via maximum likelihood. Mol. Biol. Evol. 13, 1368–1374.

Dayhoff, M.O., Schwartz, R.M., Orcutt, B.C., 1978. A model of evolutionary changes in proteins. Atlas of protein sequence and structure, vol. 5(3). National Biomedical Research Foundation, Washington, D.C.

Felsenstein, J., 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17, 368–376.

Felsenstein, J., Chruchill, G.A., 1996. A hidden Makov model approach to variation among sites in rate of evolution. Mol. Biol. Evol. 13, 93–104.

Goldman, N., Thorne, J.L., Jones, D.T., 1998. Assessing the impact of secondary structure and solvent accessibility on protein evolution. Genetics 149, 445–458.

Guindon, S., Gascuel, O., 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52, 606–704.

Halpern, A.L., Bruno, W.J., 1998. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. Mol. Biol. Evol. 15, 910–917.

Hastings, W.K., 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57, 97–109.

Hendlich, M., et al., 1990. Identification of native protein folds amongst a large number of incorrect models: the calculation of low energy conformations from potentials of mean force. J. Mol. Biol. 216, 167–180.

Jaynes, E., 2003. Probability theory: the logic of science. Cambridge University Press.

Jeffreys, H., 1935. Some tests of significance, treated by the theory of probability. Proc. Camb. Phil. Soc. 31, 203–222.

Jeffreys, H., 1961. Theory of probability. Oxford University Press.

Jensen, J.L., Pedersen, A.-M.K., 2000. Probabilistic models of DNA sequence evolution with context dependent rates of substitution. Adv. Appl. Prob. 32, 499–517.

Jones, D.T., Taylor, W.R., Thornton, J.M., 1992a. The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. 8, 275–282.

Jones, D.T., Taylor, W.R., Thornton, J.M., 1992b. A new approach to protein fold recognition. Nature 358, 86–89.

Jukes, T.H., Cantor, C.R., 1969. Evolution of protein molecules. In: Munro, H.N. (Ed.), Mammalian Protein Metabolism. Academic Press, New York, pp. 21–123.

Kimura, M., 1980. A simple method for estimating evolutionary of base substitution through comparative studies of nucleotide sequences. J. Mol. Evol. 16, 111–120.

Koshi, J.M., Goldstein, R.A., 1995. Context dependent optimal substitution matrices. Protein Eng. 8, 641–645.

Koshi, J.M., Goldstein, R.A., 1997. Mutation matrices and physical–chemical properties: correlations and implications. Proteins 27, 336–344.

Larget, B., Simon, D.L., 1999. Markov Chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. Mol. Biol. Evol. 16, 750–759.

Lartillot, N., Philippe, H., 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol. Biol. Evol. 21, 1095–1109.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E., 1953. Equation of state calculation by fast computing machines. J. Chem. Phys. 21, 1087–1092.

Miyazawa, S., Jernigan, R.L., 1985. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules 18, 534–552.

Nielsen, R., 2002. Mapping mutations on phylogenies. Syst. Biol. 51, 729–739.

Pagel, M., Meade, A., 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. Syst. Biol. 53, 571–581.

Parisi, G., Echave, J., 2001. Structural constraints and emergence of sequence patterns in protein evolution. Mol. Biol. Evol. 18, 750–756.

Pedersen, A.-M.K., Jensen, J.L., 2001. A dependent rates models and MCMC based methodology for the maximum likelihood analysis of sequences with overlapping reading frames. Mol. Biol. Evol. 18, 763–776.

Robinson, D.M., Jones, D.T., Kishino, H., Goldman, N., Thorne, J.L., 2003. Protein evolution with dependence among codons due to tertiary structure. Mol. Biol. Evol. 20, 1692–1704.

Siepel, A., Haussler, D., 2004. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. Mol. Biol. Evol. 21, 468–488.

Singh, R.K., Tropha, A., Vaisman, I.I., 1996. Delaunay tessellation of proteins: four body nearest neighbor propensities of amino acid residues. J. Comput. Biol. 3, 213–222.

Thorne, J.L., Goldman, N., Jones, D.T., 1996. Combining protein evolution and secondary structure. Mol. Biol. Evol. 13, 666–673.

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