BIOINFORMATICS ORIGINAL PAPER Vol. 28 no. 2 2012, pages 176–183
doi:10.1093/bioinformatics/btr635

Sequence analysis Advance Access publication November 26, 2011

Inferring sequence regions under functional divergence in duplicate genes

Yi-Fei Huang and G. Brian Golding*

Department of Biology, McMaster University, Hamilton, ON, Canada

Associate Editor: David Posada

ABSTRACT

Motivation: A number of statistical phylogenetic methods have been proposed to identify type-I functional divergence in duplicate genes by detecting heterogeneous substitution rates in phylogenetic trees. A common disadvantage of the existing methods is that autocorrelation of substitution rates along sequences is not modeled. This reduces the power of existing methods to identify regions under functional divergence.

Results: We design a phylogenetic hidden Markov model to identify protein regions relevant to type-I functional divergence. A C++ program, HMMDiverge, has been developed to estimate model parameters and to identify regions under type-I functional divergence. Simulations demonstrate that HMMDiverge can successfully identify protein regions under type-I functional divergence unless the discrepancy of substitution rates between subfamilies is very limited or the regions under functional divergence are very short. Applying HMMDiverge to G protein α subunits in animals, we identify a candidate region longer than 20 amino acids, which overlaps with the α-4 helix and the $\alpha4$-$\beta6$ loop in the GTPase domain with divergent rates of substitutions. These sites are different from those reported by an existing program, DIVERGE2. Interestingly, previous biochemical studies suggest the α-4 helix and the $\alpha4$-$\beta6$ loop are important to the specificity of the receptor–G protein interaction. Therefore, the candidate region reported by HMMDiverge highlights that the type-I functional divergence in G protein α subunits may be relevant to the change of receptor–G protein specificity after gene duplication. From these results, we conclude that HMMDiverge is a useful tool to identify regions under type-I functional divergence after gene duplication.

Availability: C++ source codes of HMMDiverge and simulation programs used in this study, as well as example datasets, are available at http://info.mcmaster.ca/yifei/software/HMMDiverge.html

Contact: golding@mcmaster.ca

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on September 6, 2011; revised on November 10, 2011; accepted on November 11, 2011

1 INTRODUCTION

An important challenge in the post-genomic era is the identification of biological sequences that contribute to functional divergence of duplicate genes. After gene duplication, homologous regions, e.g. protein motifs or protein domains, may evolve at different rates between two duplicates because of the discrepancy of functional constraints that results from functional divergence. Therefore, the difference of substitution rates between two duplicate subfamilies can be used as a proxy of functional divergence, which is referred to as type-I functional divergence (Gu, 1999) or rate shifting (Abhiman and Sonnhammer, 2005b). Alternatively, substitution rates in both duplicate genes may increase immediately after gene duplication due to relaxed functional constraints, but decrease at a late stage due to increased functional constraints. The sequence regions or sites that are conserved within subfamilies but diverged between them may be relevant to functional divergence, which is referred to as type-II functional divergence (Gu, 1999, 2006), conservation-shifting (Abhiman and Sonnhammer, 2005b) or 'constant but different' (Gribaldo *et al.*, 2003). A number of statistical models have been proposed to detect protein regions or amino acid sites relevant to functional divergence based on the heterogeneity of substitution rates in duplicate genes (Abhiman and Sonnhammer, 2005b; Arnau *et al.*, 2006; Bielawski and Yang, 2003; Blouin *et al.*, 2003; Dorman, 2007; Gu, 1999, 2001a, b, 2006; Knudsen and Miyamoto, 2001; Knudsen *et al.*, 2003; Marin *et al.*, 2001; Nam *et al.*, 2005; Neuwald, 2010; Pupko and Galtier, 2002; Susko *et al.*, 2002). The idea of these existing methods is to detect the discrepancy of substitution rates using an extended phylogenetic model in which the substitution rates could be different between different branches.

A common drawback of the existing methods is that any autocorrelation of substitution rates along sequences is not modeled. Most phylogenetic methods assume that every site evolves independently. However, this simple assumption is frequently violated. In a recent work, Callahan *et al.* (2011) performed a whole-genome level study on the correlated evolution of nearby residues in Drosophilid proteins. A strong autocorrelation was found between non-synonymous substitutions but not between synonymous substitutions, which suggests autocorrelation at protein level (Callahan *et al.*, 2011). In addition, it has been found that positive selection varies between protein secondary structures (Ridout *et al.*, 2010). Therefore, a number of neighboring pairs of sites may show correlated substitution patterns, such as the correlated substitution rates. Unfortunately, most existing methods for identifying functional divergence do not model the autocorrelation of substitutions. Instead, independence of substitution rates across sites is assumed in most of the existing methods (Abhiman and Sonnhammer, 2005a, b; Blouin *et al.*, 2003; Dorman, 2007; Gu, 1999, 2001a, b, 2006; Knudsen and Miyamoto, 2001; Knudsen *et al.*, 2003; Susko *et al.*, 2002). These methods

*To whom correspondence should be addressed.

may be useful to detect critical sites contributed to functional divergence, because these critical sites may evolve independently in terms of spatial distribution. However, if substitution rates are autocorrelated along sequences, these methods may be less powerful than a method which can model the autocorrelation correctly, because the evolutionary signals in individual sites are very limited. In addition, these methods may not be able to correctly infer the boundaries of regions under functional divergence. In a few studies, the autocorrelation of heterogeneous substitution rates along sequences are considered but are detected by heuristic methods, such as the sliding window method (Arnau *et al.*, 2006; Gao *et al.*, 2005; Nam *et al.*, 2005). It has been argued that the sliding window method is not a desired method to study the spatial distribution of evolutionary patterns. First, failure to correct for the multiple testing problem can lead to incorrect conclusions (Schmid and Yang, 2008). Second, the resolution of the sliding window method is coarse, since the patterns are averaged over multiple sites. Third, a predefined window size typically needs to be assigned before analyses and it is not clear how to define a universally optimized window size. A short window may not be suitable to detect long regions with weak signals in each site, whereas a long window may ignore short regions with strong signals in each site (Zhang and Townsend, 2009).

In this article, we propose a phylogenetic hidden Markov model (phylo-HMM) for identifying protein regions under type-I functional divergence, which explicitly models the autocorrelation of substitution rates along sequences by a hidden Markov model. A C++ program, HMMDiverge, has been developed to implement this phylo-HMM. Simulations suggest that HMMDiverge can efficiently identify protein regions under functional divergence unless the discrepancy of substitution rates between subfamilies is very weak or the regions relevant to functional divergence are very short. By applying this method to G protien α subunits, we identify a candidate region longer than 20 amino acids that may contribute to the diversity of receptor specificity in G protein α subunits.

2 MODEL AND IMPLEMENTATION

2.1 Motivation of the phylo-HMM

Consider a gene family in which the evolutionary relationships among members are known. If the root of the phylogenetic tree corresponds to a duplication event, we may divide the family into two subfamilies, i.e. Subfamily 1 and Subfamily 2, by removing the root. After gene duplication, some regions may evolve at different rates in the two subfamilies due to differentiated functional constraints. To detect this heterogeneous pattern, a model should be able to capture at least two features: the heterogeneity of substitution rates between two subfamilies and the autocorrelation of substitution rates along sequences. Phylo-HMM (Siepel and Haussler, 2004, 2005) is an extension of standard phylogenetic models, which can naturally capture both of these features (Siepel and Haussler, 2005; Yang, 1995). In phylo-HMM, the changes of evolutionary patterns along alignments are described by an unobserved Markov chain, which can be inferred from observed alignments. We design a simple phylo-HMM to identify protein regions under type-I functional divergence in duplicate genes. We focus on protein sequences rather than DNA sequences because a large number of duplicate genes are so old that it is difficult to infer nucleotide substitution rates accurately. Our phylo-HMM is similar to a phylo-HMM used for identifying DNA sequences under lineage-specific selection (Siepel *et al.*, 2006). However, there were only two discrete substitution rate categories in this model (Siepel *et al.*, 2006). This simple assumption may not be flexible enough to describe rate variation very well. Since the functional elements under diverged selection may be very short in proteins, it is desirable to model the substitution rates with a higher resolution so that short regions with a strong discrepancy of substitution rates can be detected. In our phylo-HMM, an arbitrary number of rate categories can be used by modeling the rate variation with a discrete Gamma distribution.

2.2 Notation of the phylo-HMM

To describe the phylo-HMM, we adopt a notation similar to that described by Siepel and Haussler (2005). Formally, we define the proposed phylo-HMM to be a four-tuple, $\theta = (R, \psi, \mathbf{A}, \mathbf{b})$, consisting of a set of hidden states, R, a set of associated phylogenetic models, ψ, a one-step state transition matrix, $\mathbf{A}$, and a vector of initial-state probabilities, $\mathbf{b}$ (Siepel and Haussler, 2005). ψ determines the emission probability, i.e. the probability that we observe a column in the alignment given a hidden state. $\mathbf{A}$ and $\mathbf{b}$ specify the transition probabilities among hidden states and the initial distribution of the hidden Markov chain.

2.3 Definition of hidden states and associated phylogenetic models

The first step in designing a phylo-HMM model is to define the set of hidden states, R, and the set of associated phylogenetic models, ψ. We assume the substitution process of amino acids can be described by a fixed continuous time-reversible Markov model. We also assume the phylogenetic tree with branch lengths is known. To fully define the phylogenetic models, we only need to know the relative substitution rates in branches, which are used as scale factors to rescale corresponding branches. We assume the substitution rate is a constant within each subfamily but the substitution rates can be different between two subfamilies. In addition, we assume the rate variation can be described by a discrete Gamma distribution with k substitution rate categories (Yang, 1994). We set the shape parameter, α, equal to the scale parameter, β, to ensure that branch lengths can be interpreted as the expected number of substitutions per site. We may define all the possible pairs of the k rate categories between the two subfamilies to be the members in R, and the corresponding phylogenetic models to be the members in ψ. Clearly, there are k^2 possible pairs of the k rate categories, so there are totally k^2 hidden states and k^2 associated phylogenetic models. We define r_{ij} as a hidden state, in which the substitution rate in Subfamily 1 is in the i-th category and that in Subfamily 2 is in the j-th category. If $i = j$, the substitution rates are equal between the two subfamilies. In this scenario, there is no difference in terms of evolutionary constraints, so type-I functional divergence is not relevant. If $i > j$, the substitution rate in Subfamily 1 is higher than that in Subfamily 2, which implies type-I functional divergence. If $i < j$, the substitution rate in Subfamily 1 is lower than that in Subfamily 2, which also implies type-I functional divergence but the divergence is in the opposite direction. Therefore, we divide the members in R into three state groups: $M_0 = \{r_{ij} : i = j\}$ in which there is no evidence of type-I functional divergence, $M_1 = \{r_{ij} : i > j\}$ and $M_2 = \{r_{ij} : i < j\}$ in which there is evidence of type-I functional

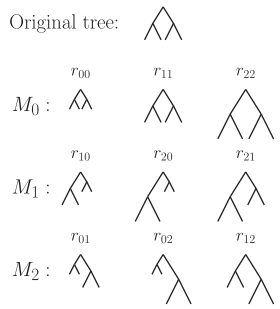

Fig. 1. The definition of hidden states and associated phylogenetic models. In this simple example, $k=3$. The tree topologies are exactly the same in all of the hidden states, in which the left subtree corresponds to Subfamily 1, whereas the right subtree corresponds to Subfamily 2. However, the two subtrees are rescaled by different factors (relative substitution rates). In model group M_0, there is no difference in terms of relative substitution rates. Therefore, there is no functional divergence in this case. In model group M_1, the substitution rate is higher in Subfamily 1. In model group M_2, the substitution rate is lower in Subfamily 1. There is functional divergence in the last two cases.

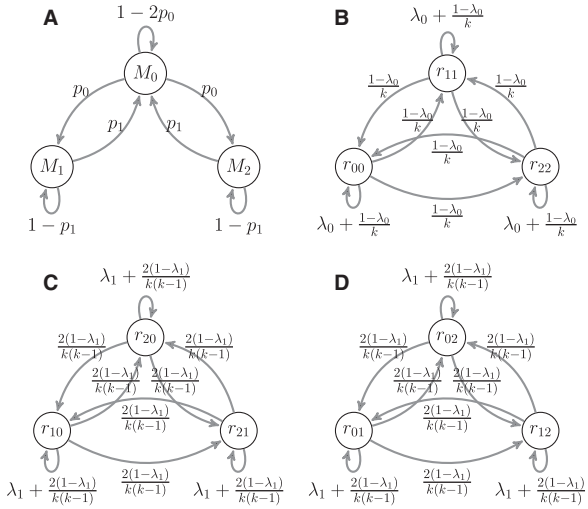

Fig. 2. The hierarchical parameterization of the one-step transition matrix. The number of rate categories, k, is set to 3 in this example, because it is the simplest non-trivial case. Nodes represent states or state groups while formulas beside arcs represent one-step transition probabilities. (**A**) transitions among three state groups (M_0, M_1, and M_2); (**B**) transitions conditional on staying in M_0; (**C**) transitions conditional on staying in M_1; (**D**) transitions conditional on staying in M_2.

divergence (Fig. 1). The key goal of the phylo-HMM is to infer the probability of each state group for each site.

2.4 Parameterization of state transition matrix and initial probability vector

To probabilistically describe the spatial distribution of hidden states in the alignment, we adopt a hierarchical framework to specify the one-step transition matrix, **A**. First, we can model the transitions among three state groups (M_0, M_1 and M_2) by a simple Markov chain (Fig. 2A). This Markov chain describes switches between 'type-I functional divergence relevant regions' and 'type-I functional divergence irrelevant regions'. We assume (i) both the one-step transition probability from M_0 to M_1 and that from M_0 to M_2 are p_0; (ii) both the one-step transition probability from M_1 to M_0 and that from M_2 to M_0 are p_1; (iii) one-step transitions between M_1 and M_2 are impossible and any transition between them must go through M_0.

The three assumptions define a symmetric Markov chain, but this symmetric Markov chain can only describe the transitions among three state groups. To fully define the transition probabilities among k^2 states, the transition process conditional on staying in each state group must be defined. We use a strategy similar to that described by Siepel and Haussler (2004). We introduce parameter λ_0 to describe the autocorrelation of states conditional on M_0 (Fig. 2B) and parameter λ_1 to describe the autocorrelation of states conditional on M_1 (Fig. 2C) or M_2 (Fig. 2D). The transition process within each state group is well defined by the two autocorrelation parameters. Conditional on staying in a state group, with probability λ_0 (or λ_1) the state in site i will be assigned to the same state in site $i-1$ and with probability $1-\lambda_0$ (or $1-\lambda_1$) it will be assigned to a state randomly drawn from all states in the same group with equal probabilities. By combining the transition probabilities among state groups and transition probabilities among states conditional on model groups, we can specify the unconditional transition probabilities between two states in the same group. For example, the transition probabilities between two states in M_0 can be defined to be

$$P(r_{i'j'}|r_{ij}) = \begin{cases} (1-2p_0)\cdot(\lambda_0+\frac{1-\lambda_0}{k}) \\ \quad \text{if } r_{ij}, r_{i'j'} \in M_0 \text{ and } r_{ij} = r_{i'j'}, \\ (1-2p_0)\cdot\frac{1-\lambda_0}{k} \\ \quad \text{if } r_{ij}, r_{i'j'} \in M_0 \text{ and } r_{ij} \neq r_{i'j'}. \end{cases} \quad (1)$$

In the products, the first components, $1-2p_0$, correspond to the probabilities of staying in M_0 and the second components, $\lambda_0+\frac{1-\lambda_0}{k}$ and $\frac{1-\lambda_0}{k}$, correspond to the transition probabilities between two states conditional on staying in M_0. Similarly, the transition probabilities between two states in M_1 (or two states in M_2) can be defined to be

$$P(r_{i'j'}|r_{ij}) = \begin{cases} (1-p_1)\cdot(\lambda_1+\frac{2(1-\lambda_1)}{k(k-1)}) \\ \quad \text{if } r_{ij}, r_{i'j'} \in M_1 \ (M_2) \text{ and } r_{ij} = r_{i'j'}, \\ (1-p_1)\cdot\frac{2(1-\lambda_1)}{k(k-1)} \\ \quad \text{if } r_{ij}, r_{i'j'} \in M_1 \ (M_2) \text{ and } r_{ij} \neq r_{i'j'}. \end{cases} \quad (2)$$

The first components, $1-p_1$, correspond to the probabilities of staying in M_1 or M_2, whereas the second components, $\lambda_1+\frac{2(1-\lambda_1)}{k(k-1)}$ and $\frac{2(1-\lambda_1)}{k(k-1)}$, correspond to the transition probabilities between two states conditional on staying in M_1 or M_2.

Based on this hierarchical structure, we can also specify the transition probabilities between two states in different state groups. It is easy to show the stationary probability of a state conditional on the corresponding state group is equal to one over the number of states in this group, i.e. $\frac{1}{k}$ for M_0 and $\frac{2}{k\cdot(k-1)}$ for M_1 and M_2 (Siepel and Haussler, 2004). Therefore, when the hidden Markov chain transits from one state group to another group, it is natural to draw one state from all the states in the new state group with equal probabilities as the new state. The transition probabilities between two states in

different state groups can be defined to be

$$P(r_{i'j'}|r_{ij}) = \begin{cases} p_0 \cdot \frac{2}{k(k-1)} \\ \quad \text{if } r_{ij} \in M_0 \text{ and } r_{i'j'} \in M_1 \cup M_2, \\ p_1 \cdot \frac{1}{k} \\ \quad \text{if } r_{ij} \in M_1 \cup M_2 \text{ and } r_{i'j'} \in M_0. \end{cases} \quad (3)$$

In the products, the first components, p_0 and p_1, correspond to the transition probabilities between two state groups, whereas the second components, $\frac{2}{k(k-1)}$ and $\frac{1}{k}$, correspond to the probabilities of randomly drawing a state from the new state group. Now, all the transition probabilities among k^2 states are fully defined.

We define the initial probability vector, **b**, to be the stationary distribution of the one-step transition matrix, **A**. As shown in the Supplementary Material, the stationary distribution is

$$\pi(r_{ij}) = \begin{cases} \frac{p_1}{(2p_0+p_1)k} & \text{if } r_{ij} \in M_0, \\ \frac{2p_0}{(2p_0+p_1)(k-1)k} & \text{if } r_{ij} \in M_1 \cup M_2. \end{cases} \quad (4)$$

In summary, the phylo-HMM is fully parameterized by five free parameters (p_0, p_1, λ_0, λ_1 and Gamma shape parameter, α).

2.5 Computational implementation

We used a model comparison method to test whether the alignment contains any sequence region under type-I functional divergence. In our phylo-HMM, if p_0 is equal to 0 and p_1 is a constant which is not equal to 0, the hidden Markov chain always stays in M_0 and our phylo-HMM degenerates to the model described by Siepel and Haussler (2004) with two parameters (λ_0 and Gamma shape parameter, α). This was the null model in which the duplicate genes are not under functional divergence. The full model with five parameters served as the alternative model. If the null model was rejected, we concluded that the two subfamilies evolved at different rates and might be relevant to functional divergence. We used a naïve empirical Bayesian framework to estimate how likely a site is relevant to type-I functional divergence (Yang, 2006). In this framework, parameters estimated in the full model were treated as true parameters and the posterior probability of each state group in each site was estimated using the forward–backward algorithm (Durbin *et al.*, 1998).

We have developed a C++ program, HMMDiverge, to implement the proposed phylo-HMM. HMMDiverge was based on Bio++ (Dutheil *et al.*, 2006), a set of libraries designed for phylogenetics and population genetics. In principle, the topology and branch lengths of the phylogenetic tree should be considered as free parameters and be estimated in the phylo-HMM. However, in practice it may be infeasible to estimate so many parameters. A preliminary simulation suggested standard phylogenetic software, such as PhyML (Guindon and Gascuel, 2003), could infer the tree topology and branch lengths with a high accuracy in the simulated data generated by HMMDiverge, if the regions under functional divergence are not very long (data not shown). Therefore, when we analyzed real data, we assumed that the phylogenetic trees estimated by PhyML (Guindon and Gascuel, 2003) were true trees and fixed them in HMMDiverge.

The JTT model (Jones *et al.*, 1992) was used to describe the transitions among amino acids and the number of rate categories, k, was set to 4. Maximum likelihood method was used to estimate parameters given a protein tree and an alignment. The emission probability, i.e. the probability of an observed column pattern in the alignment given r_{ij}, was calculated by the pruning algorithm proposed by Felsenstein (1981). The gaps were treated as 'missing data' or equivalently ambiguous amino acids (Felsenstein, 1981). Then, the likelihood of the observed alignment was calculated by the forward–backward algorithm (Durbin *et al.*, 1998). Parameters were estimated by maximizing the likelihood function using conjugate gradient method with multiple initial values (Press *et al.*, 1992), in which the derivatives are calculated numerically.

To identify regions under functional divergence, the marginal probability of each state in each site was calculated by the forward–backward algorithm (Durbin *et al.*, 1998) using parameters estimated in the full model. The probability of each state group was calculated by summing the probabilities of states in the group. We are especially interested in the sites in which the probabilities of M_1 or those of M_2 are very high, since these sites are likely to be located in regions under type-I functional divergence.

3 SIMULATION STUDY

3.1 Assumptions and implementation of simulations

To verify the usefulness and robustness of HMMDiverge, we performed a simulation study. In general, we do not assume the proposed phylo-HMM captures all aspects of functional evolution, because the real evolutionary process is too complicated to be fully described by any model. However, a useful model should be powerful enough to detect strong patterns even if the model itself is only a rough approximation of the true mechanism. Therefore, the reference simulation datasets are based on a set of assumptions, which are simple but very different from those in HMMDiverge:

(i) Lengths of 'type-I functional divergence relevant regions' and 'irrelevant regions' are both fixed rather than described by a Markov chain in each simulation. In the reference simulations, five lengths (5 amino acids, 10 amino acids, 20 amino acids, 50 amino acids and 100 amino acids) and three lengths (50 amino acids, 100 amino acids and 200 amino acids) were used for the 'type-I functional divergence relevant regions' and 'irrelevant regions', respectively.

(ii) 'Type-I functional divergence relevant regions' and 'irrelevant regions' are distributed alternatively in alignments while the first region is always irrelevant to functional divergence in every alignment. For a 'functional divergence relevant region', one subfamily is randomly selected to be the subfamily that evolves at lower rate.

(iii) In a 'type-I functional divergence relevant region', the branches in the slowly evolved subfamily are rescaled by a constant, ρ_1, and the branches in the rapidly evolved subfamily are rescaled by another constant, ρ_2 ($\rho_1 < \rho_2$). In the reference simulations, three pairs of scale factors were used. In the first pair, $\rho_1 = 0.5$ and $\rho_2 = 1.5$, which corresponds to a weak discrepancy of substitution rates between two subfamilies. In the second pair, $\rho_1 = 0.25$ and $\rho_2 = 1.75$, which corresponds to an intermediate discrepancy of substitution rates. In the third pair, $\rho_1 = 0.125$ and $\rho_2 = 1.875$, which corresponds to a strong discrepancy of substitution rates.

(iv) The standard discrete Gamma mixture model is used to describe rate variation across sites (Yang, 1994). We emphasize that the Gamma shape parameter, α, in the simulations has a different meaning from the α in the phylo-HMM. In the reference simulations, α was set to 0.5.

(v) The substitution process of amino acids is described by the JTT model (Jones *et al.*, 1992).

We have developed a C++ program to generate the simulation datasets. The protein phylogenetic tree of a set of 30 G protein α subunits (Supplementary Fig. S1) was used in the simulation, which will be described in more detail in the Section 4. To explore parameter space, we generated 20 alignments for each combination of the mentioned parameters in the reference simulations. The length of each alignment was set to 420 amino acids, which is the approximate length of the G protein α subunit alignment. Then, the simulated alignments and the true phylogenetic tree were fed to HMMDiverge to estimate parameters and the probabilities of state groups in all sites. If the probability of M_1 or that of M_2 is higher than a given probability cutoff, the site may be considered to be relevant to functional divergence. In this way, given a probability cutoff, we get a binary classification which indicates whether a given site is relevant to functional divergence. Comparing the classifications with the true states, we evaluated the performance of HMMDiverge. Since the probability cutoff could significantly influence true positive rates and false positive rates, we summarized the results by receiver operating characteristic (ROC) curves (Fig. 3) generated by the ROCR package (Sing *et al.*, 2005).

3.2 Performance of HMMDiverge in simulations

In the reference simulations, the performance of HMMDiverge is strongly influenced by the discrepancy of substitution rates and the lengths of 'functional divergence relevant regions' (Fig. 3). If the discrepancy of substitution rates between two subfamilies is very limited, i.e. the scale factors of branch lengths are 0.5 in the slowly evolved subfamily and 1.5 in the rapidly evolved subfamily, the performance of HMMDiverge is not very strong due to lack of sufficient signal, represented as ROC curves very close to the main diagonals (Fig. 3). However, if the discrepancy of substitution rates is intermediate, i.e. the scale factors of branch lengths are 0.25 in the slowly evolved subfamily and 1.75 in the rapidly evolved subfamily, the performance is fairly good unless the 'type-I functional divergence relevant regions' are very short, e.g. 5 amino acids (Fig. 3). If the discrepancy of substitution rates is very strong, i.e. the scale factors are 0.125 in the slowly evolved subfamily and 1.875 in the rapidly evolved subfamily, the performance is even better (Fig. 3). The lengths of 'functional divergence irrelevant regions' also influence the performance but are less important than the lengths of 'functional divergence relevant regions' and the discrepancy of substitution rates (Fig. 3). In summary, HMMDiverge can accurately identify regions under type-I functional divergence unless the rate shift is very limited or regions under functional divergence are very short. The results coincide with our intuition that it is easier to identify long regions in which substitution rates are very different between two subfamilies and highlight that HMMDiverge may be a useful tool to detect type-I functional divergence.

The reference simulations do not address whether the variability of substitution rates across sites influences the performance of HMMDiverge. Therefore, we performed two sets of additional simulations, in which all parameters were the same as these in the reference simulations except the shape parameter, α. In the first set of additional simulations, α was set to 0.2, which implied the substitution rates were highly variable across sites. In this

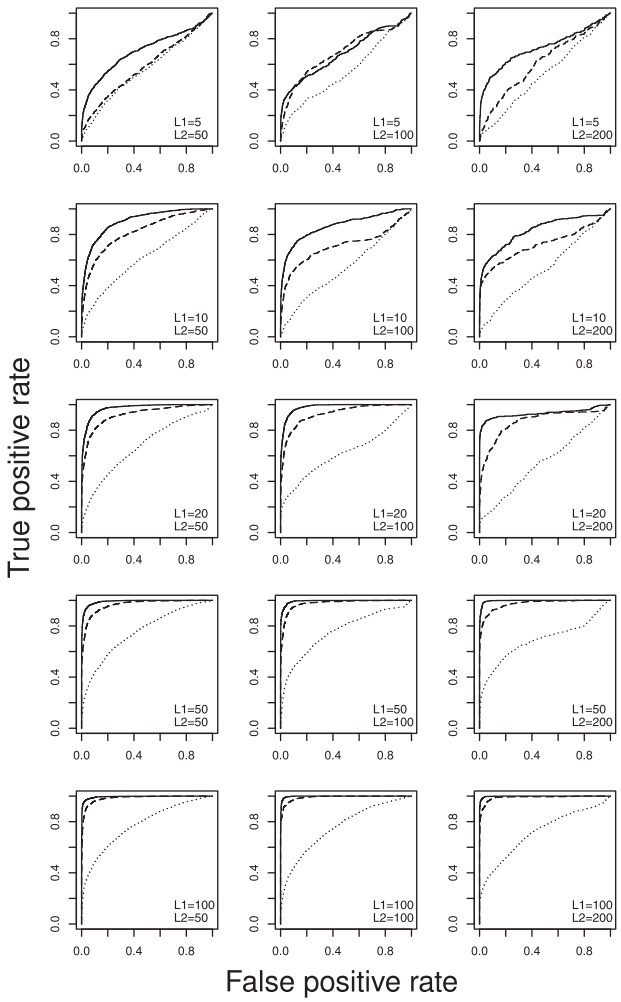

Fig. 3. The performance of HMMDiverge in the reference simulations. X axes represent false positive rates while Y axes represent true positive rates. Each row consists of the ROC curves of multiple simulations having the equal length of divergence relevant regions (L1), which are 5 aa, 10 aa, 20 aa, 50 aa, and 100 aa, increasing from top to bottom. Each column consists of the ROC curves of multiple simulations having the equal length of divergence irrelevant regions (L2), which are 50 aa, 100 aa, and 200 aa, increasing from left to right. Three types of curves represent three pairs of branch scale parameters. Dotted curves: the scale factor is 1.5 in the rapidly evolved subfamily and 0.5 in the slowly evolved subfamily. Dashed curves: the two scale factors are 1.75 and 0.25 respectively. Solid curves: the two scale factors are 1.875 and 0.125 respectively. The shape parameter α is 0.5 in all simulations.

scenario, the performance of HMMDiverge is quantitatively worse than that reported in the reference simulations (Supplementary Fig. S5). In contrast, in the second set of additional simulations, α was set to 1.0, which suggested the variability of substitution rates across sites was low. In this scenario, HMMDiverge performs better than that reported in the reference simulations (Supplementary Fig. S6). The results are fairly intuitive, because low variation means low noise, which in turn positively influences the performance. Thus, the variability of substitution rates does indeed influence the performance of HMMDiverge, but its influence is relatively small

compared with the discrepancy of substitution rates and the size of 'functional divergence relevant regions'.

DIVERGE2 (Zheng *et al.*, 2007) is an existing program to identify 'functional divergence relevant sites', in which independence of substitution rates across sites is assumed. If evolutionary signals in individual sites are strong, ignoring the autocorrelation of substitution rates along sequences may not significantly reduce the power to detect sequence regions under type-I functional divergence. To compare the power of DIVERGE2 with that of HMMDiverge in the context of detecting regions under functional divergence, we applied the 'Gu99' method in DIVERGE2 to the alignments in the reference simulations. The 'Gu99' method is a fast method to identify 'type-I functional divergence relevant sites', which gave similar results as the more advanced 'Gu2001' method (Gu, 2001a). As shown in Supplementary Figures S7–S9, seldom can DIVERGE2 identify sites in 'type-I functional divergence relevant regions', since the ROC curves of DIVERGE2 are very close to the main diagonals. Therefore, at least in the reference simulations, HMMDiverge is more powerful than DIVERGE2.

To compare the performance of HMMDiverge with that of DIVERGE2 in the context of identifying individual sites under functional divergence, we performed the third set of additional simulations. The simulations adopted the same set of assumptions as the reference simulations. However, the length of 'functional divergence relevant regions' was set to 1, which implies that individual sites rather than regions are units of functional divergence. Three lengths, 19, 9 and 4, were used for 'functional divergence irrelevant regions'. Besides, six pairs of branch scale factors were used (0.5 versus 1.5, 0.25 versus 1.75, 0.125 versus 1.875, 0.1 versus 5.0, 0.1 versus 10.0 and 0.1 versus 15.0). In the first three pairs, evolutionary signal in each site is weak while in the last three pairs it is strong. The Gamma shape parameter, α, was set to 0.5. In total, 18 combinations of parameters were examined. 20 alignments were generated for each combination of parameters and then both HMMDiverge and DIVERGE2 were used to identify sites under type-I functional divergence. As shown in Supplementary Figures S10 and S11, the power of HMMDiverge is very close to that of DIVERGE2 in the context of identifying individual sites under type-I functional divergence.

4 CASE STUDY OF G PROTEIN α SUBUNITS

4.1 Parameter estimation in G protein α subunits

Heterotrimeric guanine nucleotide-binding proteins (G proteins) are a family of protein complexes important to signal transduction (Kaziro *et al.*, 1991; Neer, 1995). There are three subunits in a typical G protein, a Gα subunit, a Gβ subunit and a Gγ subunit (Cabrera-Vera *et al.*, 2003; Lambright *et al.*, 1994). The Gα subunits, which have GTPase activity, are key factors in signal transduction pathways relevant to heterotrimeric G proteins (Kaziro *et al.*, 1991; Neer, 1995). Based on sequence similarities, Gα can be divided into four major subfamilies: Gs alpha, Gio alpha, Gq alpha and G12 alpha (Kaziro *et al.*, 1991; Simon *et al.*, 1991). Zheng *et al.* (2007) studied the functional divergence of Gα subunits in animals using their software, DIVERGE2, and detected a number of candidate sites under type-I or type-II functional divergence after the splitting of Gq alpha subunits and Gs alpha subunits.

Table 1. Estimation of parameters in G protein α subunits

Parameter	Estimation in null model	Estimation in alternative model
p_0	0[a]	0.0311
p_1	–[b]	0.153
λ_0	0.858	0.944
λ_1	–[b]	2.03×10^{-5}
α	0.808	0.771
Likelihood ratio	−4824.28	−4813.52

[a]Fixed parameters.
[b]Unused parameters.

However, DIVERGE2 assumes that substitution rates are not autocorrelated along sequences. Thus, it is highly desirable to reanalyze the functional divergence in G protein α subunits using HMMDiverge and check whether the phylo-HMM could uncover any new evidence on the functional divergence of G protein α subunits. We therefore reanalyzed the data provided by Zheng *et al.* (2007) and compared the results from HMMDiverge with the results reported by Zheng *et al.* (2007).

We downloaded the 16 Gq alpha protein sequences and 14 Gs alpha protein sequences analyzed by Zheng *et al.* (2007) from NCBI. To be consistent with the notation in the previous sections, Gq alpha class is labeled as Subfamily 1 while Gs alpha class is labeled as Subfamily 2. MUSCLE (Edgar, 2004) was used to align the 30 protein sequences. A maximum likelihood tree (Supplementary Fig. S1 in the Material) was reconstructed by PhyML (Guindon and Gascuel, 2003) with the JTT + Γ model. The maximum likelihood tree is essentially the same as the neighbor-joining tree reported by Zheng *et al.* (2007), which can be divided into two subfamilies, Gq alpha subunits and Gs subunits (Supplementary Fig. S1). We rooted the phylogenetic tree at the middle of the longest path. Then, the maximum likelihood tree and the alignment were fed to HMMDiverge to estimate parameters and log likelihoods in both the null and the alternative (full) model. As shown in Table 1, the log likelihood ratio of the alternative model and the null model is 21.5. Hypothesis testing was performed by a parametric bootstrap. We generated 1000 alignments based on the parameters estimated in the null model and HMMDiverge was applied to these alignments. We do not find any log likelihood ratios larger than 21.5 in the 1000 simulations (Supplementary Fig. S2). Therefore, the null model can be rejected and we conclude Gq subfamily and Gs subfamily are functionally diverged. The same conclusion is attained by performing a likelihood ratio test in which we assume the log likelihood ratio follows χ^2 distribution with 3 degrees of freedom ($p < 0.001$). We found that the χ^2 test is more conservative than the parametric bootstrap (data not shown).

4.2 Identification of regions under functional divergence

We can gain more insights on functional divergence by identifying the locations of the sequence regions relevant to functional divergence. The site-specific probabilities of the three model groups (M_0, M_1 and M_2) can be calculated by HMMDiverge (Fig. 4). To choose a reasonable cutoff for classifying sites, we generated 50 simulated alignments using parameters estimated in the full model and then applied HMMDiverge to these alignments. The ROC curve

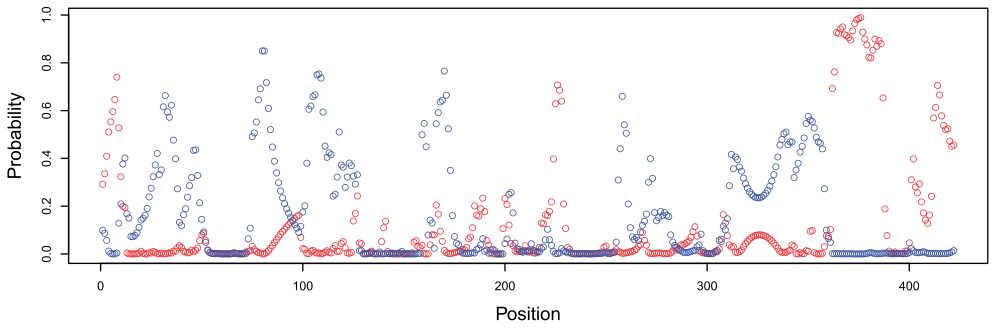

Fig. 4. The site-specific prediction for type-I functional divergence in G protein α subunits. The X-axis represents locations of sites, while the Y-axis represents the probability of each model group. Red dots represent model group M_1 and blue dots represent model group M_2.

is shown in Supplementary Fig. S3. We empirically choose 0.8 as the probability cutoff. In this case, the false positive rate is 1.6%, whereas the true positive rate is 48.2% (Supplementary Fig. S3). This cutoff is relatively conservative because typically we are less tolerant to false positives than false negatives.

As shown in Fig. 4, we do not find evidence of type-I functional divergence in most sites, since neither probabilities of M_1 nor those of M_2 are higher than the cutoff, 0.8, in most of sites. However, two regions do show evidence of functional divergence. The first region consists of site 80 and site 81. In this region, M_2 is strongly supported, which suggests the two sites evolve faster in Gs class. More interestingly, the second candidate region, in which M_1 is highly supported, is fairly long, ranging from site 364 to 386. As suggested by our simulations, HMMDiverge might not be able to identify short regions very well, so we focus on the second region.

Deeper insights on molecular adaptation can be gained by the combination of evolutionary evidence and structure information (Golding and Dean, 1998). We downloaded Protein Data Bank (PDB) entry 1AZS, which contains the Gq alpha subunit in *Bos taurus*, and then mapped the second candidate region onto chain C in PDB entry 1AZS using Jalview (Clamp *et al.*, 2004). The second candidate region overlaps with the α4-helix and the α4-β6 loop (Supplementary Fig. S4). Experimental studies have suggested both the α4-helix and the α4-β6 loop are critical to mediating receptor–G protein specificity (Bae *et al.*, 1997, 1999; Cabrera-Vera *et al.*, 2003; Lee *et al.*, 1995). The sequence region under functional divergence predicted by HMMDiverge may imply that functional divergence of receptor–G protein specificity after the splitting of Gq subfamily and Gs subfamily is related to the change of functional constraints in the α4-helix and the α4-β6 loop.

4.3 Comparison with previous studies

To gain some insights on how modeling the autocorrelation of evolutionary patterns along sequences influences prediction, we compared the sites predicted by HMMDiverge to those reported by DIVERGE2 (Zheng *et al.*, 2007). Both site 80 and site 81 in the first candidate region predicted by HMMDiverge were identified by DIVERGE2 as well. However, DIVERGE2 only identified sites 362, 374, and 376 close to the second candidate region. The inability of DIVERGE2 to identify most of the sites in the second candidate region reported by HMMDiverge might be due to the weak evolutionary signal per site in this region. In turn, for the 25 sites under type-I functional divergence reported by DIVERGE2, 20 sites are not related to the candidate regions reported by HMMDiverge. The 20 sites may contain strong evolutionary signal so that DIVERGE2 can detect them. However, these individual sites may be too short to be detected by HMMDiverge, because the parameters estimated by HMMDiverge may mostly reflect the patterns in the long regions under functional divergence. Therefore, DIVERGE2 and HMMDiverge may uncover different aspects of type-I functional divergence after duplication. DIVERGE2 may be more powerful to detect scattered critical amino acids relevant to type-I functional divergence. In contrast, HMMDiverge may be more powerful to detect regions under divergence, and may be able to find the boundaries of these regions. Nevertheless, the long regions reported by HMMDiverge, e.g. the second candidate region, may more likely be related to functional divergence, since the parallel shift of substitution rates in multiple sites in a region is strong evidence of functional divergence.

5 DISCUSSION

Here we report a customized phylo-HMM for identifying protein regions under type-I functional divergence. A C++ implementation of this phylo-HMM, HMMDiverge, has been developed. Given an alignment and a phylogenetic tree, HMMDiverge first estimates parameters by maximum likelihood estimation and then decodes the probabilities of underlying state groups by treating estimated parameters as true parameters. This is a naïve Bayesian method (Yang, 2006). In the case study of G protein α subunits, HMMDiverge needs about 1 cpu hour to finish the analysis. Therefore, it is fast enough to perform whole genomic analyses.

Extensive simulations have been performed to test HMMDiverge. As shown in Figure 3, HMMDiverge can identify candidate regions under type-I functional divergence unless the discrepancy of substitution rates between two subfamilies is very limited or the regions relevant to type-I functional divergence are very short, both of which suggest that the pattern of functional divergence is weak. Since the simulated datasets were generated by a set of assumptions different from the assumptions in the phylo-HMM, the phylo-HMM may be a robust method to identify regions under functional divergence. In the case study of G protein α subunits, HMMDiverge detected a long candidate region under type-I functional divergence. This long region may be important to the specific receptor–G protein interaction based on

existing biochemical evidence (Bae *et al.*, 1997, 1999; Cabrera-Vera *et al.*, 2003; Lee *et al.*, 1995). Most of the sites within this candidate region have not been identified by DIVERGE2, an existing program for functional divergence, which suggests HMMDiverge can identify some new candidates under functional divergence. In addition, the regions reported by HMMDiverge may not include the sites identified by DIVERGE2, because the former concentrates on regions while the latter examines only sites. We believe HMMDiverge is a useful supplement to existing methods for identifying regions under functional divergence. New insights can be gained by applying HMMDiverge to real data as we have shown in the case study of G protein α subunits.

ACKNOWLEDGEMENTS

We are grateful to Ben Evans, Jonathon Stone, Jonathan Dushoff and three anonymous reviewers for their helpful comments on the manuscript. We thank Julien Dutheil, Sylvain Gaillard, Wilson Sung and Hui Zhao for technical help.

Funding: National Sciences and Engineering Research Council of Canada (NSERC) (grant RGPIN140221-10) and Canada Research Chair (CRC) grant (to G.B.G.).

Conflict of Interest: none declared.

REFERENCES

Abhiman,S. and Sonnhammer,E.L. (2005a) FunShift: a database of function shift analysis on protein subfamilies. *Nucleic Acids Res.*, **33**, D197–D200.

Abhiman,S. and Sonnhammer,E.L. (2005b) Large-scale prediction of function shift in protein families with a focus on enzymatic function. *Proteins*, **60**, 758–768.

Arnau,V. *et al.* (2006) Uvpar: fast detection of functional shifts in duplicate genes. *BMC Bioinformatics*, **7**, 174.

Bae,H. *et al.* (1997) Molecular determinants of selectivity in 5-hydroxytryptamine1b receptor-g protein interactions. *J. Biol. Chem.*, **272**, 32071–32077.

Bae,H. *et al.* (1999) Two amino acids within the α4 helix of Gi1 mediate coupling with 5-Hydroxytryptamine1B receptors. *J. Biol. Chem.*, **274**, 14963–14971.

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

Blouin,C. *et al.* (2003) Inferring functional constraints and divergence in protein families using 3D mapping of phylogenetic information. *Nucleic Acids Res.*, **31**, 790–797.

Cabrera-Vera,T.M. *et al.* (2003) Insights into G protein structure, function, and regulation. *Endocrine Rev.*, **24**, 765–781.

Callahan,B. *et al.* (2011) Correlated evolution of nearby residues in drosophilid proteins. *PLoS Genet.*, **7**, e1001315.

Clamp,M. *et al.* (2004) The Jalview java alignment editor. *Bioinformatics*, **20**, 426–427.

Dorman,K. (2007) Identifying dramatic selection shifts in phylogenetic trees. *BMC Evol. Biol.*, **7**, S10.

Durbin,R. *et al.* (1998) *Biological Sequence Analysis*. Cambridge University Press, Cambridge, UK.

Dutheil,J. *et al.* (2006) Bio++: A set of C++ libraries for sequence analysis, phylogenetics, molecular evolution and population genetics. *BMC Bioinformatics*, **7**, 188.

Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.

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

Gao,X. *et al.* (2005) SplitTester: Software to identify domains responsible for functional divergence in protein family. *BMC Bioinformatics*, **6**, 137.

Golding,G.B. and Dean,A.M. (1998) The structural basis of molecular adaptation. *Mol. Biol. Evol.*, **15**, 355–369.

Gribaldo,S. *et al.* (2003) Functional divergence prediction from evolutionary analysis: a case study of vertebrate hemoglobin. *Mol. Biol. Evol.*, **20**, 1754–1759.

Gu,X. (1999) Statistical methods for testing functional divergence after gene duplication. *Mol. Biol. Evol.*, **16**, 1664–1674.

Gu,X. (2001a) Maximum-likelihood approach for gene family evolution under functional divergence. *Mol. Biol. Evol.*, **18**, 453–464.

Gu,X. (2001b) A site-specific measure for rate difference after gene duplication or speciation. *Mol. Biol. Evol.*, **18**, 2327–2330.

Gu,X. (2006) A simple statistical method for estimating Type-II (cluster-specific) functional divergence of protein sequences. *Mol. Biol. Evol.*, **23**, 1937–1945.

Guindon,S. and Gascuel,O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.*, **52**, 696–704.

Jones,D.T. *et al.* (1992) The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.*, **8**, 275–282.

Kaziro,Y. *et al.* (1991) Structure and function of signal-transducing GTP-binding proteins. *Annu. Rev. Biochem.*, **60**, 349–400.

Knudsen,B. and Miyamoto,M. M. (2001) A likelihood ratio test for evolutionary rate shifts and functional divergence among proteins. *Proc. Natl Acad. Sci. USA*, **98**, 14512–14517.

Knudsen,B. *et al.* (2003) Using evolutionary rates to investigate protein functional divergence and conservation: a case study of the carbonic anhydrases. *Genetics*, **164**, 1261–1269.

Lambright,D.G. *et al.* (1994) Structural determinants for activation of the α-subunit of a heterotrimeric g protein. *Nature*, **369**, 621–628.

Lee,C.H. *et al.* (1995) Multiple regions of G alpha 16 contribute to the specificity of activation by the C5a receptor. *Mol. Pharmacol.*, **47**, 218–223.

Marin,I. *et al.* (2001) Detecting changes in the functional constraints of paralogous genes. *J. Mol. Evol.*, **52**, 17–28.

Nam,J. *et al.* (2005) A simple method for predicting the functional differentiation of duplicate genes and its application to MIKC-type MADS-box genes. *Nucleic Acids Res.*, **33**, e12.

Neer,E.J. (1995) Heterotrimeric G proteins: organizers of transmembrane signals. *Cell*, **80**, 249–257.

Neuwald,A. (2010) Bayesian classification of residues associated with protein functional divergence: Arf and Arf-like GTPases. *Biol. Direct*, **5**, 66.

Press,W. *et al.* (1992) *Numerical Recipes in C*, 2nd edn. Cambridge University Press, New York, USA.

Pupko,T. and Galtier,N. (2002) A covarion-based method for detecting molecular adaptation: Application to the evolution of primate mitochondrial genomes. *Proc. R. Soc. Lond. B Biol. Sci.*, **269**, 1313–1316.

Ridout,K.E. *et al.* (2010). Positive selection differs between protein secondary structure elements in *Drosophila*. *Genome Biol. Evol.*, **2**, 166–179.

Schmid,K. and Yang,Z. (2008) The trouble with sliding windows and the selective pressure in BRCA1. *PLoS One*, **3**, e3746.

Siepel,A. and Haussler,D. (2004) Combining phylogenetic and hidden markov models in biosequence analysis. *J. Comput. Biol.*, **11**, 413–428.

Siepel,A. and Haussler,D. (2005) Phylogenetic hidden markov models. In Nielsen,R. (ed.) *Statistical Methods in Molecular Evolution*. Statistics for Biology and Health, Chapter 12. Springer, New York, pp. 325–351.

Siepel,A. *et al.* (2006) New methods for detecting lineage-specific selection. *Lect. Notes Comput. Sc.*, **3909**, 190–205.

Simon,M. *et al.* (1991) Diversity of G proteins in signal transduction. *Science*, **252**, 802–808.

Sing,T. *et al.* (2005) ROCR: visualizing classifier performance in R. *Bioinformatics*, **21**, 3940–3941.

Susko,E. *et al.* (2002) Testing for differences in rates-across-sites distributions in phylogenetic subtrees. *Mol. Biol. Evol.*, **19**, 1514–1523.

Yang,Z. (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *J. Mol. Evol.*, **39**, 306–314.

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

Yang,Z. (2006) *Computational Molecular Evolution*. Oxford University Press, New York, USA.

Zhang,Z. and Townsend,J.P. (2009) Maximum-likelihood model averaging to profile clustering of site types across discrete linear sequences. *PLoS Comput. Biol.*, **5**, e1000421.

Zheng,Y. *et al.* (2007) Functional divergence after gene duplication and sequence-structure relationship: a case study of G-protein alpha subunits. *J. Exp. Zoolog. B Mol. Dev. Evol.*, **308**, 85–96.