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 α-β 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 α-β 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; Matin et al., 2001; Nam et al., 2005; Neuwald, 2010; Papko 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. Instead, independence of substitution 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
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; Guo 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 size 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 capture at least two features: the heterogeneity of substitution rates between two subfamilies due to differentiated functional divergence, and the autocorrelated along sequences, these methods may be less powerful than a method which can 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 to model the rate variation with a discrete Gamma distribution.

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 to model the rate variation with a discrete Gamma distribution.
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 \( (M_0, M_1, \text{and} M_2) \); (B) transitions conditional on staying in \( M_0 \); (C) transitions conditional on staying in \( M_1 \); (D) transitions conditional on staying in \( M_2 \).

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 \( (M_0, M_1, \text{and} M_2) \) 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 \( p_{01} \); (ii) both the one-step transition probability from \( M_1 \) to \( M_0 \) and that from \( M_2 \) to \( M_0 \) are \( p_{10} \); (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 \( \lambda_0 \) to describe the autocorrelation of states conditional on \( M_0 \) (Fig. 2B) and parameter \( \lambda_1 \) 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 \( \lambda_0 \) (or \( \lambda_1 \)) the state in site \( i \) will be assigned to the same state as 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

\[
P(\mathbf{r}_i | \mathbf{r}_j) = \begin{cases} 
(1 - 2p_{01}) (\lambda_0 + \frac{1}{\lambda_1}) & \text{if } r_i, r_j \in M_0 \text{ and } r_j = r_j', \\
(1 - 2p_{01}) \frac{1}{\lambda_1} & \text{if } r_i, r_j \in M_0 \text{ and } r_j \neq r_j'.
\end{cases}
\]
Detecting regions under functional divergence

2.5 Computational implementation

We used a model comparison method to test whether the alignment contains any sequence regions 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 ($\lambda$ and Gamma shape parameter, $\alpha$). 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 naive 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). A preliminary simulation suggested standard phylogenetic software, such as PhyML (Guindon and Gascuel, 2003), could infer the tree 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 powerful enough 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, 200 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, $p_1$, and the branches in the rapidly evolved subfamily are rescaled by another constant, $p_2$ ($p_1 < p_2$). In the reference simulations, three pairs of scale factors were used. In the first pair, $p_1 = 0.125$ and $p_2 = 1.875$, which corresponds to a strong discrepancy of substitution rates between two subfamilies. In the second pair, $p_1 = 0.25$ and $p_2 = 1.75$, which corresponds to an intermediate discrepancy of substitution rates. In the third pair, $p_1 = 0.125$ and $p_2 = 1.875$, which corresponds to a weak 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, $\alpha$, in the simulations has a different meaning from the $\alpha$ in the phylo-HMM. In the reference simulations, $\alpha$ was set to 0.5.

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) Each 'type-I functional divergence relevant region' and 'irrelevant region' is 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, 200 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, $p_1$, and the branches in the rapidly evolved subfamily are rescaled by another constant, $p_2$ ($p_1 < p_2$). In the reference simulations, three pairs of scale factors were used. In the first pair, $p_1 = 0.125$ and $p_2 = 1.875$, which corresponds to a strong discrepancy of substitution rates between two subfamilies. In the second pair, $p_1 = 0.25$ and $p_2 = 1.75$, which corresponds to an intermediate discrepancy of substitution rates. In the third pair, $p_1 = 0.125$ and $p_2 = 1.875$, which corresponds to a weak 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, $\alpha$, in the simulations has a different meaning from the $\alpha$ in the phylo-HMM. In the reference simulations, $\alpha$ was set to 0.5.
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, \( \alpha \). In the first set of additional simulations, \( \alpha \) was set to 0.2, which implied the substitution rates were highly variable across sites. In this 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, \( \alpha \) 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.

(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 \( \alpha \) 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 \( \alpha \) 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).
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 ‘Gun99’ method in DIVERGE2 to the alignments in the reference simulations. The ‘Gun99’ 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 relevant regions’. 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 relevant 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). 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 Gu subunit, a Gβ subunit and a Gγ subunit (Cabrera-Vera et al., 2003; Lambright et al., 1994). The Gu 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, Gu can be divided into four major subfamilies: Gs alpha, Gγ alpha, Gq alpha and G12 alpha (Kaziro et al., 1991; Simon et al., 1991). Zheng et al. (2007) studied the functional divergence of Gu 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

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimation in null model</th>
<th>Estimation in alternative model</th>
</tr>
</thead>
<tbody>
<tr>
<td>p_0</td>
<td>0</td>
<td>0.0311</td>
</tr>
<tr>
<td>p_1</td>
<td>b</td>
<td>0.153</td>
</tr>
<tr>
<td>λ</td>
<td>0.858</td>
<td>0.944</td>
</tr>
<tr>
<td>λ_2</td>
<td>b</td>
<td>2.03 × 10^{-5}</td>
</tr>
<tr>
<td>α</td>
<td>0.808</td>
<td>0.771</td>
</tr>
</tbody>
</table>

| Likelihood ratio | -4824.28 | -4813.52 |

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 (M₀, M₁ and M₂) 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

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) was reconstructed by PhyML (Guindon and Gascuel, 2003) with the JTT + F 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 a χ² distribution with 3 degrees of freedom (p < 0.001). We found that the χ² test is more conservative than the parametric bootstrap (data not shown).
To gain some insights on how modeling the autocorrelation of the probability of each model group. Red dots represent model group $M_1$ and blue dots represent model group $M_2$.

More interestingly, the second candidate region, in which supported, which suggests the two sites evolve faster in Gs class. region consists of site 80 and site 81. In this region, two regions do show evidence of functional divergence. The first region consists of site 80 and site 81. 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 $\alpha_4$-helix and the $\alpha_4$-$\beta_6$ loop (Supplementary Fig. S4). Experimental studies have suggested both the $\alpha_4$-helix and the $\alpha_4$-$\beta_6$ 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 $\alpha_4$-helix and the $\alpha_4$-$\beta_6$ 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 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 naive Bayesian method (Yang, 2006). In the case study of G protein $\alpha$ subunits, HMMDiverge needs about 1 cpu hour to finish the analysis.

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 $\alpha$ 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 HMMDDiverge can identify some new candidates under functional divergence. In addition, the regions reported by HMMDDiverge may not include the sites identified by DIVERGE2, because the former concentrates on regions while the latter examined only sites. We believe HMMDDiverge is a useful supplement to existing methods for identifying regions under functional divergence. New insights can be gained by applying HMMDDiverge to real data as we have shown in the case study of G protein α subunits. 

ACKNOWLEDGEMENTS

We are grateful to Ben Evans, Jonathan 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 GB 63).

Conflict of Interest: none declared.

REFERENCES


