# STATISTICAL PHYLOGENETIC MODELS FOR THE INFERENCE OF FUNCTIONALLY IMPORTANT REGIONS IN PROTEINS

# STATISTICAL PHYLOGENETIC MODELS FOR THE INFERENCE OF FUNCTIONALLY IMPORTANT REGIONS IN PROTEINS

BY

YIFEI HUANG, M.Sc.

A THESIS

SUBMITTED TO THE DEPARTMENT OF BIOLOGY

AND THE SCHOOL OF GRADUATE STUDIES

OF MCMASTER UNIVERSITY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

Doctor of Philosophy (2014) McMaster University

(Biology) Hamilton, Ontario, Canada

| | |
|---|---|
| TITLE: | STATISTICAL PHYLOGENETIC MODELS FOR THE INFERENCE OF FUNCTIONALLY IMPORTANT REGIONS IN PROTEINS |
| AUTHOR: | Yifei Huang<br>M.Sc., (Beijing Normal University) |
| SUPERVISOR: | Dr. G. Brian Golding |
| NUMBER OF PAGES: | xiv, 141 |

*To my parents and Bie Bie*

# Abstract

An important question in biology is the identification of functionally important sites and regions in proteins. A variety of statistical phylogenetic models have been developed to predict functionally important protein sites, e.g. ligand binding sites or protein-protein interaction interfaces, by comparing sequences from different species. However, most of the existing methods ignore the spatial clustering of functionally important sites in protein tertiary/primary structures, which significantly reduces their power to identify functionally important regions in proteins. In this thesis, we present several new statistical phylogenetic models for inferring functionally important protein regions in which Gaussian processes or hidden Markov models are used as prior distributions to model the spatial correlation of evolutionary patterns in protein tertiary/primary structures. Both simulation studies and empirical data analyses suggest that these new models outperform classic phylogenetic models. Therefore, these new models may be useful tools for extracting functional insights from protein sequences and for guiding mutagenesis experiments. Furthermore, the new methodologies developed in these models may also be used in the development of new statistical models to answer other important questions in phylogenetics and molecular evolution.

# Acknowledgements

Firstly, I would like to thank my parents, Linqian Huang and Xianwen Chen, for their endless love and supports. Without them, it is impossible for me to persist and finish my Ph.D. study. Secondly, I would like to thank Wilson Sung, Hermina Ghenu, Dr. Melanie Lou, Dr. Terri Porter, and Dr. Wilfried Haerty for encouragements and helps. Thirdly, I would like to thank Dr. Ben Evans, Dr. Jianping Xu, Dr. Jonathan Dushoff, Dr. Ben Bolker, and Dr. Jonathon Stone for teaching me valuable knowledge and providing insightful comments on my research. Finally, I would like to thank my supervisor, Dr. Brian Golding, who is the best supervisor ever in the world. My Ph.D. research was financially supported by NSERC and CRC grants to Dr. Brian Golding as well as McMaster University.

# Contents

# List of Figures

# List of Tables

# Acronyms

**BMP** Bone Morphogenetic Protein

**ERK** Extracellular Signal-regulated Kinase

**MAPK** Mitogen-activated Protein Kinase

**MCMC** Markov Chain Monte Carlo

**PDB** RCSB Protein Data Bank

**Phylo-GPM** Phylogenetic Gaussian Process Model

**Phylo-HMM** Phylogenetic Hidden Markov Model

**ROC** Receiver Operating Characteristic

**RSA** Relative Solvent Accessibility

**SMAD** Sma and Mad Related Proteins

# Chapter 1

# Introduction

Because of the fast development of DNA sequencing technologies, the number of sequenced genomes is increasing exponentially. How to interpret massive sequence data and extract useful information from them becomes a very important question in the post-genomic era. Statistical models have been shown to be very powerful for inferring functional information from biological sequences, because statistical principles naturally model the uncertainty of the relationships between biological sequences and functions. Among these statistical models for the inference of biological functions, evolution based methods, e.g. statistical phylogenetic models, are particularly interesting, because these methods are not only useful for predicting functions but also provide a unified framework for understanding the relationships among sequences, functions, and evolution. In this chapter, we will firstly introduce the existing works which apply classic statistical phylogenetic models to infer functionally important sites in coding sequences. Thereafter, we will discuss the common drawbacks of the classic phylogenetic methods which may reduce their power of inferring functionally important coding regions. Finally, we will briefly introduce several new statistical

phylogenetic models which were designed to overcome the drawbacks of the classic phylogenetic methods. These new models will be described in detail in later chapters.

In 1981, Joseph Felsenstein published a landmark paper (Felsenstein, 1981) in which the framework of statistical phylogenetics was developed to infer phylogenies using the maximum likelihood principle. In Felsenstein's paper, a continuous-time discrete-state Markov model was used to describe the evolutionary processes of DNA sequences and a phylogenetic likelihood function was constructued by combining the Markov model with a phylogenetic tree. In addition, an efficient algorithm, the pruning algorithm (Felsenstein, 1981), also known as the sum-product (belief propagation) algorithm in the machine learning literature (Bishop, 2007), was developed to calculate the phylogenetic likelihood function. Parameters in the Markov model, banch lengthes, and the topology of the phylogenetic tree can then be estimated by maximizing the phylogenetic likelihood function (Felsenstein, 1981). Even though the statistical phylogenetic models were originally designed for inferring phylogenies, it is relatively straightforward to apply similar models to infer natural selection acted on biological sequences (Yang, 1998; Yang *et al.*, 2000; Yang, 2006; Mayrose *et al.*, 2004; Glaser *et al.*, 2003; Gu, 1999, 2001a, 2006). Because natural selection is related to biological functions, the inferred selection pressures provide insights on the functionally important sites and regions in sequences. For example, if a protein region is under strong purifying selection, most of the substitutions in the region are deleterious and the region may be essential for the biological activity of the protein. In contrast, if a protein region is under strong positive selection, many substitutions in the region may convey fitness benefits and the region may contribute to the adaptation of the organisms in a changing environment.

Most of the existing phylogenetic methods for inferring functional coding sites and regions are based on a series of highly cited papers published by Ziheng Yang and colleagues (Yang, 1994; Yang *et al.*, 2000; Yang and Nielsen, 2002). The basic idea of these models is that we could firstly design a categorical distribution which consists of multiple categories each of which describes the type and the magnitude of natural selection potentially acting on a number of sites in an alignment. Then, we could design a mixture model based on the categorical distribution and Felsenstein's phylogenetic likelihood function (Felsenstein, 1981) to model the evolution of the observed alignment. Finally, we could use standard statistical inference machineries designed for mixture models to infer the type and the magnitude of natural selection acting on each site in the observed alignment. For example, if we use an amino acid substitution model, e.g. the JTT model (Jones *et al.*, 1992; Kosiol and Goldman, 2005), to describe the substitution processes of amino acids and a discrete Gamma distribution (Yang, 1994) to model the variation of substitution rates across sites, we could calculate the posterior distribution of substitution rate at each amino acid site given a protein alignment, which in turn provides an approximate estimation of the magnitude of purifying selection acting on each site. The Rate4Site program (Mayrose *et al.*, 2004) and the ConSurf web server (Glaser *et al.*, 2003) use this idea to find highly conserved protein sites which may have important functions. The idea can be further generalized by assuming that substitution rates may be different in different subfamilies in the phylogeny, which has been implemented in the DIVERGE program to infer amino acids sites under functional divergence after gene duplication (Gu, 1999, 2001a, 2006).

While these existing models are useful, they inherit a number of drawbacks from

the framework of mixture models. The biggest drawback is that they assume that each site in the alignment is independent and identically distributed (i.i.d.). While the i.i.d. assumption significantly simplifies the parameterizations and computational implementations of the models, it is not compatible with the current knowledge in biology. For example, in the context of inferring conserved sites in a protein alignment, the i.i.d. assumption implies that slowly evolved functionally sites are randomly distributed and do not form any spatial pattern in either the protein primary structures or the protein tertiary structures. However, it is well known that functionally important sites tend to be clustered together in the protein tertiary/primary structures and form functional regions instead of random sites. Therefore, the existing methods based on the i.i.d. assumption have weak statistical power to identify functional regions. The problem is further aggravated by the fact that homologous sequences are typically very similar, which reduces the effective sample sizes in datasets. Therefore, methods based on the i.i.d. assumption are vulnerable to over-fitting and may not be able to infer selection pressures acting on sites accurately.

In this thesis, we will present several new statistical phylogenetic models for the Bayesian inference of functionally important regions in protein tertiary/primary structures. These models combine prior distributions which can naturally capture the spatial correlation of evolutionary patterns, e.g. substitution rates, in protein tertiary/primary structures with Felsenstein's phylogenetic likelihood function (Felsenstein, 1981) to infer functionally important protein regions. Both simulations and focused case studies suggest that these new models are more powerful than the classic phylogenetic models based on the i.i.d. assumption in the context of inferring functionally important regions in protein tertiary/primary structures. Therefore, we

believe that these new models are useful tools for the *in silico* inference of functionally important protein regions and several techniques developed in these models, e.g. the framework of phylogenetic Gaussian process models, may be used to study other important questions in evolutionary biology and bioinformatics.

# Chapter 2

# Inferring Sequence Regions under Functional Divergence in Duplicate Genes

Huang, Y.-F, and Golding, G. B. (2012) Inferring sequence regions under functional divergence in duplicate genes. *Bioinformatics* **28**: 176–183.

## 2.1 Abstract

After gene duplication, some protein sites or regions may evolve at different substitution rates in the two duplicate genes due to different natural selection pressures. This phenomenon is known as type-I functional divergence. 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. 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 $\alpha$ subunits in animals, we identify a candidate region longer than 20 amino acids which overlaps with the $\alpha$-4 helix and the $\alpha$4-$\beta$6 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 $\alpha$-4 helix and the $\alpha$4-$\beta$6 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 $\alpha$ 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. 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.

## 2.2 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 acted on the two duplicates. 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, 2005a). 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, 2005a), 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 (Gu, 1999, 2001b,a, 2006; Knudsen and Miyamoto, 2001; Marin *et al.*, 2001; Susko *et al.*, 2002; Bielawski and Yang, 2003; Blouin *et al.*, 2003; Knudsen *et al.*, 2003; Abhiman and Sonnhammer, 2005a; Nam *et al.*, 2005; Arnau *et al.*, 2006; Dorman, 2007; Neuwald, 2010; Pupko and Galtier, 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 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 (Gu, 1999, 2001a,b; Knudsen and Miyamoto, 2001; Susko *et al.*, 2002; Blouin *et al.*, 2003; Knudsen *et al.*, 2003; Abhiman and Sonnhammer, 2005a,b; Gu, 2006; Dorman, 2007). 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 (Gao *et al.*, 2005; Nam *et al.*, 2005; Arnau *et al.*, 2006). It

has been argued that the sliding window method is not a desired method to study the spatial distribution of evolutionary patterns. Firstly, failure to correct for the multiple testing problem can lead to incorrect conclusions (Schmid and Yang, 2008). Secondly, the resolution of the sliding window method is coarse, since the patterns are averaged over multiple sites. Thirdly, 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 while a long window may ignore short regions with strong signals in each site (Zhang and Townsend, 2009).

In this paper, 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 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 $\alpha$ subunits, we identify a candidate region longer than 20 amino acids which may contribute to the diversity of receptor specificity in G protein $\alpha$ subunits.

## 2.3 Model and Implementation

### 2.3.1 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 different 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, 2005, 2004) is an extension of standard phylogenetic models which can naturally capture both of these features (Yang, 1995; Siepel and Haussler, 2005). 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. Because 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.

### 2.3.2 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, $\psi$, a one-step state transition matrix, $\mathbf{A}$, and a vector of initial-state probabilities, $\mathbf{b}$ (Siepel and Haussler, 2005). $\psi$ 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.

### 2.3.3 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, $\psi$. 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

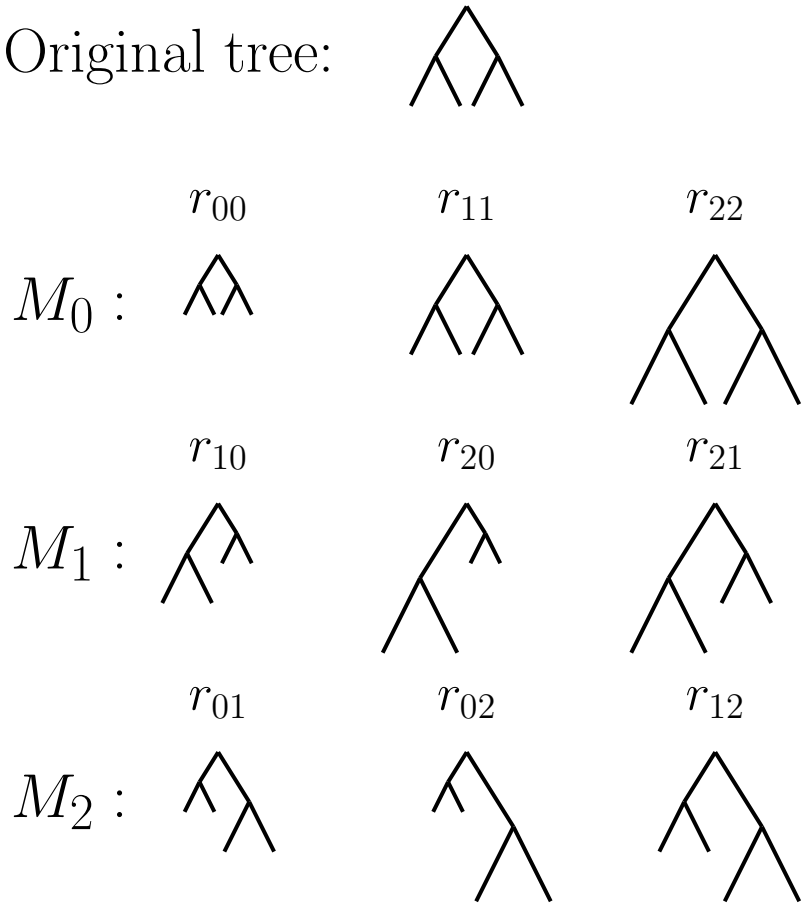


Figure 2.1: The definition of hidden states and associated phylogenetic models. In this simple example, $k = 3$. The tree topologies are exactly the same in all of the hidden states, in which the left subtree corresponds to subfamily 1 while the right subtree corresponds to subfamily 2. However, the two subtrees are rescaled by different factors (relative substitution rates). In model group $M_0$, there is no difference in terms of relative substituion 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.

described by a discrete Gamma distribution with $k$ substitution rate categories (Yang, 1994). We set the shape parameter, $\alpha$, equal to the scale parameter, $\beta$, 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 $\psi$. 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 divergence (Figure 2.1). The key goal of the phylo-HMM is to infer the probability of each state group for each site.

### 2.3.4 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,

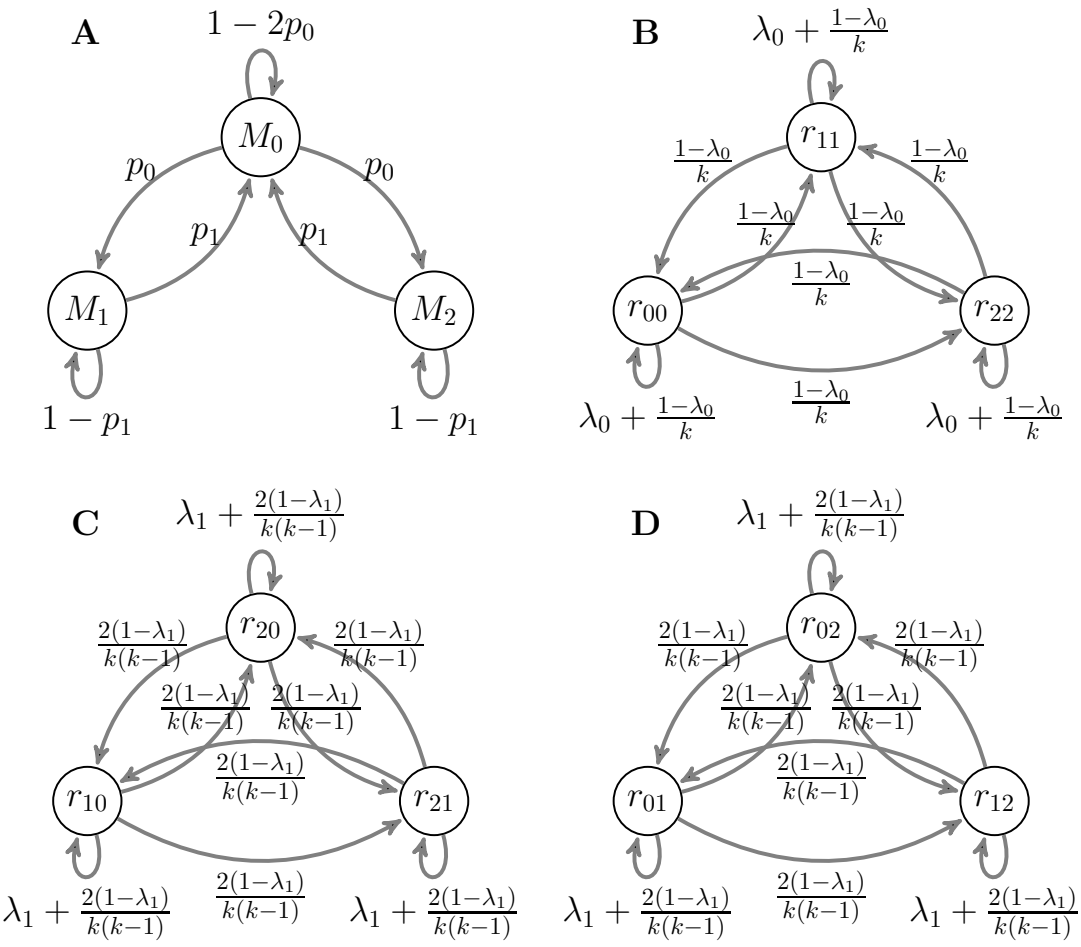


Figure 2.2: The hierarchical parameterization of the one-step transition matrix. The number of rate categories, $k$, is set to 3 in this example, because it is the simplest non-trivial case. Nodes represent states or state groups while formulas beside arcs represent one-step transition probabilities. **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 $M_2$.

**A**. Firstly, we can model the transitions among three state groups ($M_0$, $M_1$, and $M_2$) by a simple Markov chain (Figure 2.2A). This Markov chain describes switches between 'type-I functional divergence relevant regions' and 'type-I functional divergence irrelevant regions'. We assume (a) both the one-step transition probability from $M_0$ to $M_1$ and that from $M_0$ to $M_2$ are $p_0$; (b) both the one-step transition probability from $M_1$ to $M_0$ and that from $M_2$ to $M_0$ are $p_1$; (c) 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 $M_0$ (Figure 2.2B) and parameter $\lambda_1$ to describe the autocorrelation of states conditional on $M_1$ (Figure 2.2C) or $M_2$ (Figure 2.2D). 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 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 of 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

$$P(r_{i'j'}|r_{ij}) = \begin{cases} (1-2p_0)\cdot(\lambda_0 + \frac{1-\lambda_0}{k}) \\ \quad \text{if } r_{ij}, r_{i'j'} \in M_0 \text{ and } r_{ij} = r_{i'j'}, \\ (1-2p_0)\cdot\frac{1-\lambda_0}{k} \\ \quad \text{if } r_{ij}, r_{i'j'} \in M_0 \text{ and } r_{ij} \neq r_{i'j'}. \end{cases} \tag{2.1}$$

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

$$P(r_{i'j'}|r_{ij}) = \begin{cases} (1-p_1)\cdot(\lambda_1 + \frac{2(1-\lambda_1)}{k(k-1)}) \\ \quad \text{if } r_{ij}, r_{i'j'} \in M_1 \ (M_2) \text{ and } r_{ij} = r_{i'j'}, \\ (1-p_1)\cdot\frac{2(1-\lambda_1)}{k(k-1)} \\ \quad \text{if } r_{ij}, r_{i'j'} \in M_1 \ (M_2) \text{ and } r_{ij} \neq r_{i'j'}. \end{cases} \tag{2.2}$$

The first components, $1-p_1$, correspond to the probabilities of staying in $M_1$ or $M_2$, while 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 of 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

$$P(r_{i'j'}|r_{ij}) = \begin{cases} p_0 \cdot \frac{2}{k(k-1)} \\ \quad \text{if } r_{ij} \in M_0 \text{ and } r_{i'j'} \in M_1 \cup M_2, \\ p_1 \cdot \frac{1}{k} \\ \quad \text{if } r_{ij} \in M_1 \cup M_2 \text{ and } r_{i'j'} \in M_0. \end{cases} \tag{2.3}$$

In the products, the first components, $p_0$ and $p_1$, correspond to the transition probabilities between two state groups, while 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 of the transition probabilities among $k^2$ states are fully defined.

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

$$\pi(r_{ij}) = \begin{cases} \frac{p_1}{(2p_0+p_1)k} & \text{if } r_{ij} \in M_0, \\ \frac{2p_0}{(2p_0+p_1)(k-1)k} & \text{if } r_{ij} \in M_1 \cup M_2. \end{cases} \tag{2.4}$$

In summary, the phylo-HMM is fully parameterized by 5 free parameters ($p_0$, $p_1$, $\lambda_0$, $\lambda_1$, and Gamma shape parameter, $\alpha$).

### 2.3.5 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 ($\lambda_0$ 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 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 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.

## 2.4 Simulation Study

### 2.4.1 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:

(a) 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.

(b) '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.

(c) In a 'type-I functional divergence relevant region', the branches in the slowly evolved subfamily are rescaled by a constant, $\rho_1$, and the branches in the rapidly evolved subfamily are rescaled by another constant, $\rho_2$ ($\rho_1 < \rho_2$). 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 a 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.

(d) 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.

(e) 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 (see Figure A.1 in the Supplementary Material) was used in the simulation, which will be described in more detail in the section of Case Study of G Protein $\alpha$ Subunits. 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. Because the probability cutoff could significantly

influence true positive rates and false positive rates, we summarized the results by receiver operating characteristic (ROC) curves (Figure 2.3) generated by the ROCR package (Sing *et al.*, 2005).

### 2.4.2 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' (Figure 2.3). 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 (Figure 2.3). 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 (Figure 2.3). 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 (Figure 2.3). 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 (Figure 2.3). 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 (see Figure A.5 in the Supplementary Material). 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 (see Figure A.6 in the Supplementary Material). 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 to 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

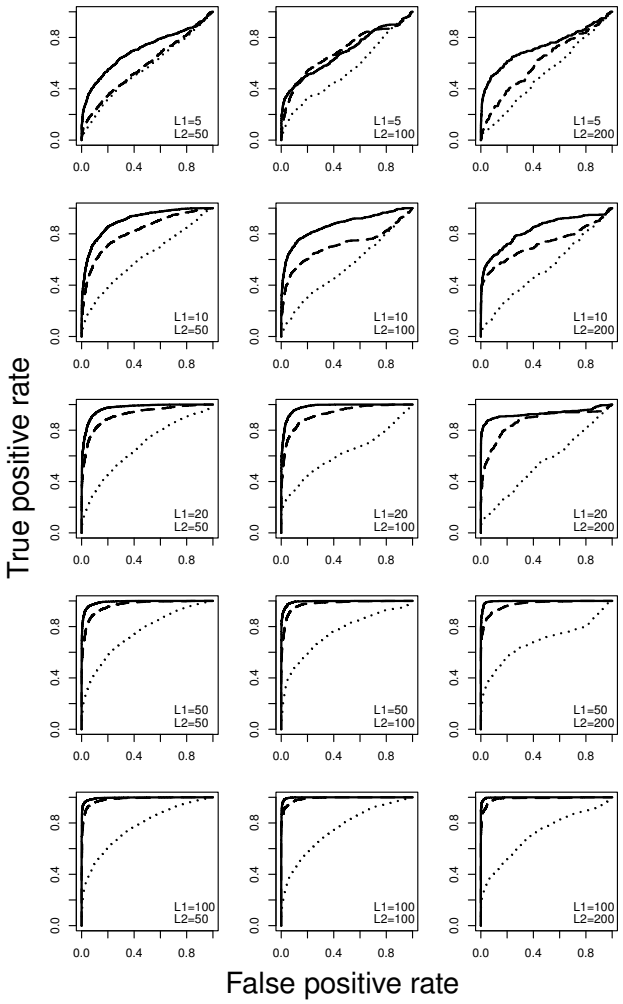


Figure 2.3: The performance of HMMDiverge in the reference simulations. X axes represent false positive rates while Y axes represent true positive rates. Each row consists of the ROC curves of multiple simulations having the equal length of divergence relevant regions (L1), which are 5 aa, 10 aa, 20 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 $\alpha$ is 0.5 in all simulations.

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 Figure A.7, A.8, and A.9 in the Supplementary Material, 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 vs 1.5, 0.25 vs 1.75, 0.125 vs 1.875, 0.1 vs 5.0, 0.1 vs 10.0, and 0.1 vs 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, $\alpha$, 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 Figure A.10 and A.11 in the Supplementary Material, the power of HMMDiverge is very close to that of DIVERGE2 in the context of identifying individual sites under type-I functional divergence.

## 2.5 Case Study of G Protein $\alpha$ Subunits

### 2.5.1 Parameter Estimation in G Protein $\alpha$ 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 G$\alpha$ subunit, a G$\beta$ subunit, and a G$\gamma$ subunit (Lambright *et al.*, 1994; Cabrera-Vera *et al.*, 2003). The G$\alpha$ 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, G$\alpha$ can be divided into four major subfamilies: Gs alpha, Gio alpha, Gq alpha, and G12 alpha (Simon *et al.*, 1991; Kaziro *et al.*, 1991). Zheng *et al.* (2007) studied the functional divergence of G$\alpha$ 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.

However, DIVERGE2 assumes substitution rates are not autocorrelated along sequences. Thus, it is highly desirable to reanalyze the functional divergence in G protein $\alpha$ subunits using HMMDiverge and check whether the phylo-HMM could uncover any new evidence on the functional divergence of G protein $\alpha$ 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

Table 2.1: Estimation of parameters in G protein $\alpha$ subunits.

| Parameter | Est. in null model | Est. in alternative model |
|---|---|---|
| $p_0$ | $0*$ | 0.0311 |
| $p_1$ | - | 0.153 |
| $\lambda_0$ | 0.858 | 0.944 |
| $\lambda_1$ | - | $2.03 \times 10^{-5}$ |
| $\alpha$ | 0.808 | 0.771 |
| Log likelihood | $-4824.28$ | $-4813.52$ |

$*$: fixed parameters; -: unused parameters.

protein sequences. A maximum likelihood tree (see Figure A.1 in the Supplementary Material) was reconstructed by PhyML (Guindon and Gascuel, 2003) with the JTT + $\Gamma$ 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 (see Figure A.1 in the Supplementary Material). 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 2.1, 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 (see Figure A.2 in the Supplementary Material). 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 $\chi^2$ distribution with 3

degrees of freedom ($p < 0.001$). We found that the $\chi^2$ test is more conservative than the parametric bootstrap (data not shown).

### 2.5.2 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 (Figure 2.4). 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 is shown in Figure A.3 in the Supplementary Material. We empirically choose 0.8 as the probability cutoff. In this case, the false positive rate is 1.6% while the true positive rate is 48.2% (see Figure A.3 in the Supplementary Material). This cutoff is relatively conservative because typically we are less tolerant to false positives than false negatives.

As shown in Figure 2.4, 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.

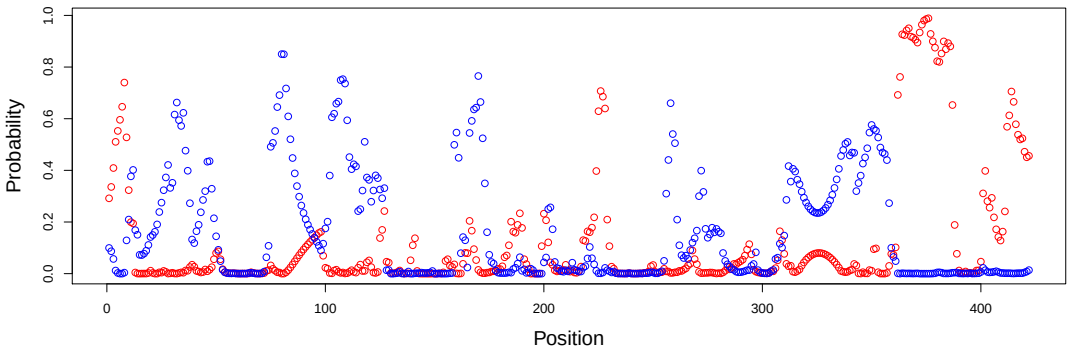


Figure 2.4: The site-specific prediction for type-I functional divergence in G protein $\alpha$ subunits. The X axis represents locations of sites, while the Y axis represents the probability of each model group. Red dots represent model group $M_1$ and blue dots represent model group $M_2$.

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 Gs 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 (see Figure A.4 in the Supplementary Material). Experimental studies have suggested both the $\alpha$4-helix and the $\alpha$4-$\beta$6 loop are critical to mediating receptor-G protein specificity (Cabrera-Vera *et al.*, 2003; Bae *et al.*, 1997, 1999; 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.

### 2.5.3 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.

## 2.6 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 firstly 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 $\alpha$ 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 2.3, 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. Because 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 (Cabrera-Vera *et al.*, 2003; Bae *et al.*, 1997, 1999; 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 $\alpha$ subunits.

## 2.7 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.

# Chapter 3

# Phylogenetic Gaussian Process Model for the Inference of Functionally Important Regions in Protein Tertiary Structures

Huang, Y.-F, and Golding, G. B. (2014) Phylogenetic Gaussian Process Model for the Inference of Functionally Important Regions in Protein Tertiary Structures. *PLoS Computational Biology* **10**: e1003429.

## 3.1 Abstract

A critical question in biology is the identification of functionally important amino acid sites in proteins. Because functionally important sites are under stronger purifying selection, site-specific substitution rates tend to be lower than usual at these sites.

A large number of phylogenetic models have been developed to estimate site-specific substitution rates in proteins and the extraordinarily low substitution rates have been used as evidence of function. Most of the existing tools, e.g. Rate4Site, assume that site-specific substitution rates are independent across sites. However, site-specific substitution rates may be strongly correlated in the protein tertiary structure, since functionally important sites tend to be clustered together to form functional patches. We have developed a new model, GP4Rate, which incorporates the Gaussian process model with the standard phylogenetic model to identify slowly evolved regions in protein tertiary structures. GP4Rate uses the Gaussian process to define a nonparametric prior distribution of site-specific substitution rates, which naturally captures the spatial correlation of substitution rates. Simulations suggest that GP4Rate can potentially estimate site-specific substitution rates with a much higher accuracy than Rate4Site and tends to report slowly evolved regions rather than individual sites. In addition, GP4Rate can estimate the strength of the spatial correlation of substitution rates from the data. By applying GP4Rate to a set of mammalian B7-1 genes, we found a highly conserved region which coincides with experimental evidence. GP4Rate may be a useful tool for the *in silico* prediction of functionally important regions in the proteins with known structures.

## 3.2 Introduction

An important question in biology is the identification of functional residues in proteins. This information can help us understand the relationship between protein

structures and functions as well as guide us to design new proteins by genetic engineering. However, experimental techniques for identifying functional sites, e.g. mutagenesis, are time consuming and expensive, which prohibits the brute force scanning of functional sites by experiments. Therefore, bioinformatics tools are useful, because they can narrow down the candidate sites for experimental investigation. Evolution operates similar to a high-throughput mutagenesis experiment: spontaneous mutations introduce protein variants in each generation and then the functional effects of the spontaneous mutations are "measured" by natural selection (Kumar *et al.*, 2011). Therefore, protein sequences contain signatures of natural selection which reflect the functions of amino acid residues. For example, mutations at the functionally important sites tend to disrupt the proteins' normal functions, so these sites usually are more conserved than unimportant ones. If the sequences of a family of homologous proteins can be collected from multiple species, we may compare these sequences to infer which sites are more important than others.

A number of bioinformatics tools based on phylogenetics have been developed to infer functional sites by the simple idea that functionally important amino acid sites tend to be more conserved than unimportant ones (Lichtarge *et al.*, 1996; Dean and Golding, 2000; Madabushi *et al.*, 2002; Simon *et al.*, 2002; Innis *et al.*, 2004; Mayrose *et al.*, 2004; Nimrod *et al.*, 2005; Capra and Singh, 2007; Goldenberg *et al.*, 2009; Ashkenazy *et al.*, 2010). Given the multiple sequence alignment and the phylogenetic tree of a protein family, these phylogenetic methods can infer the amino acid substitution rate at each site in the alignment and an unusually low substitution rate implies that the site is functionally important. It has been shown that the predicted conserved sites coincide with experimental evidence, which confirms that these

bioinformatics tools are useful.

However, these existing methods are far from flawless. Most of the popular methods, e.g. Rate4Site (Mayrose *et al.*, 2004) used in the ConSurf web server (Ashkenazy *et al.*, 2010), assume that the substitution rates are independent across sites. In statistical terms, this means that the sites in the alignment are independent and identically distributed (i.i.d.). The i.i.d. assumption simplifies the statistical modeling, but it is unrealistic from the viewpoint of biology. The i.i.d. assumption implies that the slowly evolved functional sites are randomly distributed in the protein tertiary structure. In contrast, it is well known that functionally important sites tend to be close to each other in the protein tertiary structure and form functional regions, e.g. ligand binding sites or catalytic active sites. Clearly the i.i.d. assumption is inappropriate if a functional region consists of a number of sites.

Several methods have been developed to incorporate the spatial correlation of evolutionary patterns, e.g. substitution rates at the protein level or $dN/dS$ ratios at the codon level, to overcome the drawbacks of the i.i.d. assumption (Dean and Golding, 2000; Simon *et al.*, 2002; Suzuki, 2004; Berglund *et al.*, 2005; Nimrod *et al.*, 2005; Liang *et al.*, 2006; Tusche *et al.*, 2012; Watabe and Kishino, 2013). Most of these methods use a sliding window framework, in which the amino acid substitution rate or the $dN/dS$ ratio at a focal site is approximated by the average substitution rate in a set of neighbor sites in the protein tertiary structure (Dean and Golding, 2000; Suzuki, 2004; Berglund *et al.*, 2005). A site is considered to be a neighbor of the focal site if the Euclidean distance between the two sites is smaller than a predefined window size. Unfortunately, these sliding window methods also have intrinsic drawbacks. Firstly, in most, if not all, of sliding window methods the neighbor sites, including the focal site

itself, are weighted equally in the inference of the substitution rate. However, clearly the focal site itself contains more information on its substitution rate than the sites near the boundary of the sliding window. Secondly, it is unclear how to determine the optimal window size (Huang and Golding, 2012; Zhang and Townsend, 2009). If the window size is too large, there will be too many distant sites in the window, which could bias the estimation at the focal site. In contrast, if the window size is too small, the sliding window methods will not be able to capture the spatial correlation of substitution rates and may lead to overfitting. Furthermore, there is evidence that the optimal window sizes may vary among different protein families (Suzuki, 2004).

Very recently, a Bayesian model which combines the Potts model in statistical physics and the phylogenetic model has been proposed by Watabe and Kishino to infer protein patches under positive selection in protein tertiary structures (Watabe and Kishino, 2013). In Watabe and Kishino's model, the Potts model is used to define a prior distribution of $dN/dS$ ratios over a protein tertiary structure. This model solved many problems of the sliding window framework. However, the prior distribution in Watabe and Kishino's model is unnormalized (Watabe and Kishino, 2013), which makes it difficult to design efficient algorithms to estimate hyperparameters. An advanced algorithm, thermodynamic integration (Lartillot and Philippe, 2006), was used in Watabe and Kishino's model to infer hyperparameters. However, the algorithm may be very inefficient, especially if there are many hyperparameters in the Potts model.

Here we propose to incorporate a Gaussian process with the phylogenetic model to overcome the drawbacks of the existing methods. The Gaussian process has been

widely applied in geostatistics and machine learning to capture the spatial correlation of interesting features (Banerjee *et al.*, 2004; Rasmussen and Williams, 2005). Here we will briefly introduce the basic idea of the Gaussian process. More details of the Gaussian process and its applications can be found in the geostatistics and machine learning literature, e.g. (Banerjee *et al.*, 2004). A Gaussian process defines a probability distribution over functions, namely that a single sample point of the Gaussian process is a function over a space, e.g. a 3D space. Because the sample points of the Gaussian process are "smooth" functions, the Gaussian process encodes an intrinsic spatial correlation. Thus physically closely located points in the space are more likely to have similar function values. Therefore, the Gaussian process is very useful for defining prior distributions over spatially correlated patterns. For example, in this paper we are interested in modeling the spatial correlation of site-specific substitution rates in protein tertiary structures. If we image each residue in a protein tertiary structure as a single point in the 3D space, the Gaussian process can be used to define a prior distribution of site-specific log substitution rates over these points (residues). The "smoothness" property of Gaussian process prior suggests that two physically closely located sites are more likely to have similar site-specific log substitution rates than two distantly located sites. Then, the Gaussian process prior can be combined with standard phylogenetic likelihood functions (Felsenstein, 1981) to infer site-specific substitution rates from real data.

We name this kind of hybrid model of Gaussian processes and phylogenetics as a phylogenetic Gaussian process model (Phylo-GPM). In the Phylo-GPM framework, the spatial correlation of substitution rates can be naturally described and the strength of spatial correlation can be learned from the data. Therefore, it overcomes

the common drawback of the sliding window methods that the window size must be manually specified. Unlike Watabe and Kishino's model (Watabe and Kishino, 2013), the phylogenetic Gaussian process model uses a normalized prior, so simple algorithms, i.e. the widely used Metropolis algorithm (Metropolis *et al.*, 1953; Hastings, 1970), can be used to efficiently infer hyperparameters. We have developed software, GP4Rate, based on the Phylo-GPM framework. In both simulated and real datasets, GP4Rate outperforms Rate4Site, a widely used tool based on the i.i.d. assumption. Therefore, GP4Rate may be a useful tool for the identification of functionally important sites.

## 3.3 Results

### 2D Toy Protein Simulations

Simulations were implemented to evaluate the performance of GP4Rate and to compare it with the widely used software, Rate4Site (Mayrose *et al.*, 2004). In the comparisons, Rate4Site is used as a representative of the classic phylogenetic models which use the discrete Gamma distribution to describe the variation of substitution rates across sites (Yang, 1994) but do not consider the spatial correlation of site-specific substitution rates in the protein tertiary structure. Because the true site-specific substitution rates are known in the simulated alignments, the estimated site-specific substitution rates can be compared with the true rates to evaluate the performance of the two methods. We generated two sets of simulated alignments based on different assumptions. In this and the next section, we will describe the first set of simulations which were based on a 2D toy protein structure. Thereafter we will describe the

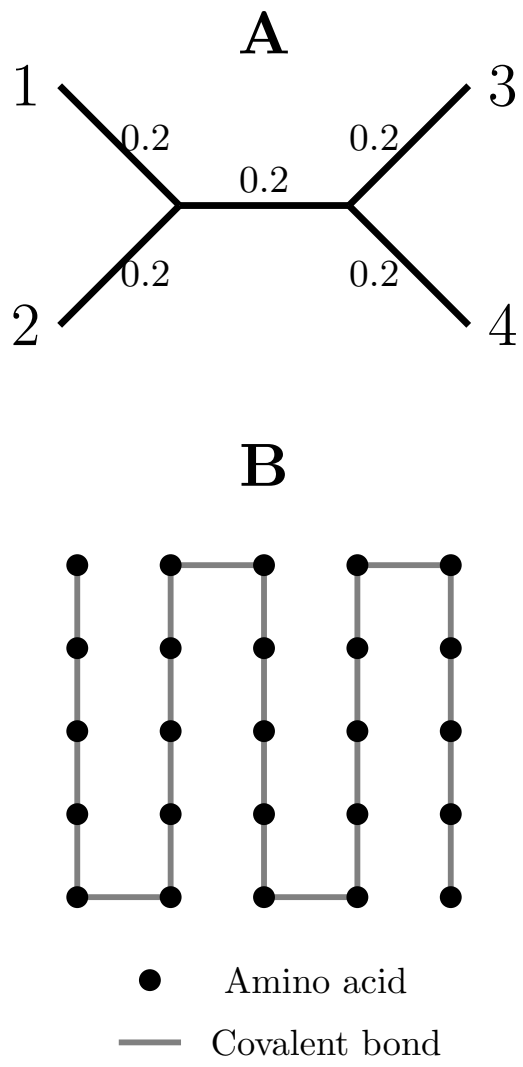


Figure 3.1: The phylogenetic tree used in all simulations and an example of 2D toy protein structure. (A) the phylogenetic tree used in all simulations; (B) a 5 by 5 2D toy protein. In the phylogenetic tree, there are 4 species and all branch lengths are equal to 0.2 substitution per site. In the example of 2D toy protein, there are 25 amino acids which are dots in a 5 by 5 2D grid. Lines between dots correspond to the "covalent bounds" between amino acid residues. A larger 20 by 20 2D toy protein with 400 residues is used in the 2D toy protein simulations.

second set of simulations which were based on more realistic assumptions.

To generate simulated alignments, we need a phylogenetic tree to describe the evolutionary relationship between simulated sequences, a protein structure to calculate the pairwise Euclidean distances between sites, a substitution model, and a vector of substitution rates. Note that the following discussions will be mainly based on the substitution rates rather than their log values. A simple phylogenetic tree was used in all simulations (Figure 3.1A). The tree consisted of four sequences and all the branch lengths were equal to 0.2 substitution per site. Because the total branch length was equal to 1 substitution per site, on average an amino acid site only contained a single substitution. Therefore, the accurate estimation of substitution rate at a single site is challenging. The JTT substitution model (Jones *et al.*, 1992; Kosiol and Goldman, 2005) was used in all simulations. Note that the protein tertiary structure and the vectors of substitution rates used in the two sets of simulated alignments were different and will be described in detail below.

In the 2D toy protein model, the protein tertiary structure was described by a 20 by 20 regular 2D grid, in which each dot corresponds to an amino acid in the toy protein structure (Figure 3.1B). In addition, we assumed that the distance between adjacent sites in the 2D grid is equal to 5 Å. This distance is comparable to the average distance between $\alpha$-carbon atoms of the physically interacting residues in real proteins. Even though the 2D toy protein model is artificial and no real protein has a similar structure, it is useful because the estimated site-specific substitution rates can be easily visualized by a heatmap (Figure 3.2). Therefore, we used the 2D toy protein model to check the correctness of the program and to get insights on the performance of GP4Rate.

Two different spatial configurations of site-specific substitution rates were used in

the 2D toy protein simulations. In the first configuration, the 20 by 20 grid was divided into 4 non-overlapping blocks, each of which was a 10 by 10 grid (Figure 3.2A). Sites within a block had the same substitution rates but different blocks could have different substitution rates. Two substitution rates, 0.2 and 1.8, were used for simulations and the substitution rates of blocks were alternatively arranged in the 2D protein structure (Figure 3.2A). Therefore, the toy proteins consisted of two conserved blocks with low substitution rates (0.2) and two variable blocks with high substitution rates (1.8). The second configuration was similar to the first one, but the sizes of non-overlapping blocks were 5 by 5 instead of 10 by 10 (Figure 3.2B). Twenty simulated alignments were generated for each configuration of site-specific substitution rates. It is easy to notice that the average site-specific substitution rate is equal to 1 in both configurations.

A program based on Bio++ (Dutheil *et al.*, 2006; Gueguen *et al.*, 2013) was developed to implement the simulations. For each simulated alignment, we ran two separate MCMC chains using GP4Rate to estimate site-specific substitution rates. For each MCMC chain, $10^6$ iterations were implemented and the trace plots of the MCMC outputs were monitored to ensure the convergence of the MCMC chains. The first 30% of the iterations were discarded as burn-in. Then, the two chains were combined to calculate the average substitution rate at each site. To compare the performance of GP4Rate with that of Rate4Site, we also used Rate4Site to estimate the substitution rates. To make the results of GP4Rate and Rate4Site more comparable, the phylogenetic tree and branch lengths were fixed to the true values in both GP4Rate and Rate4Site.

We firstly randomly sampled two simulated alignments, one for each configuration,

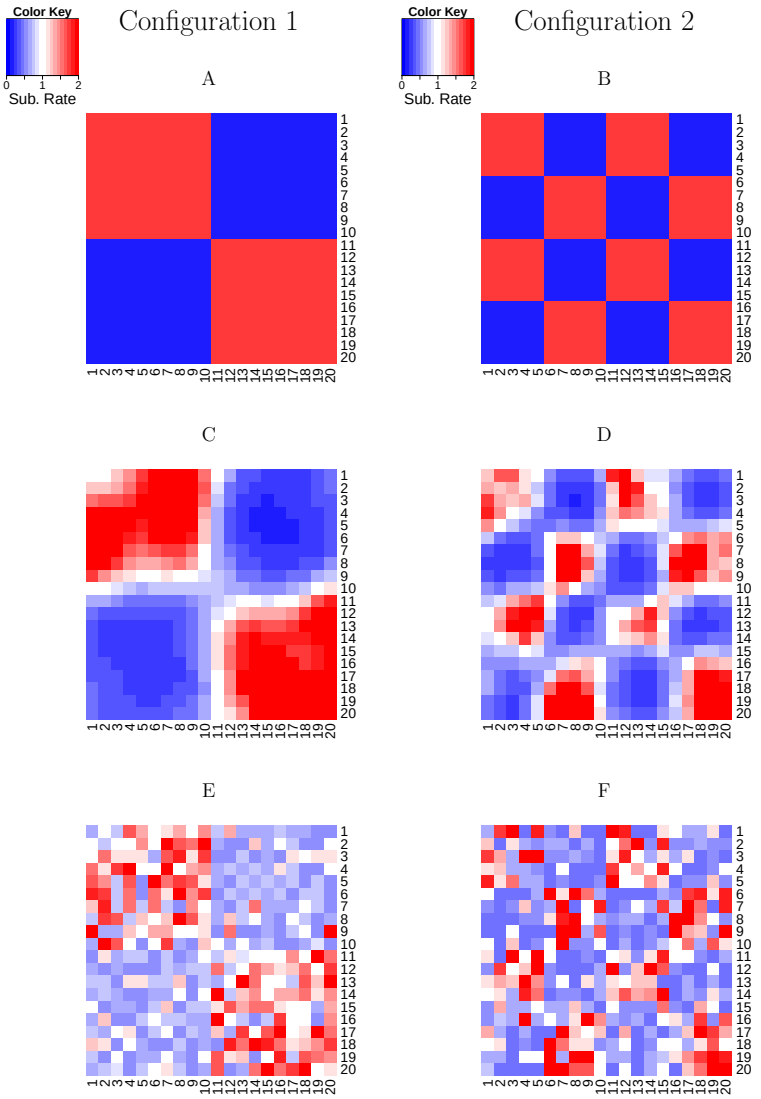


Figure 3.2: The visualization of the estimated site-specific substitution rates in the 2D toy protein simulations. The heatmaps are based on two randomly selected alignments, one for each configuration. The substitution rates in the heatmaps are arranged according to the toy 2D protein structure. (A, B) the true substitution rates in the first and second configurations, respectively; (C, D) the substitution rates estimated by GP4Rate in the first and second configurations, respectively; (E, F) the substitution rates estimated by Rate4Site in the first and second configurations, respectively.

as examples to get insights on the performances of GP4Rate and Rate4Site. As shown in Figure 3.2C and 3.2D, the site-specific substitution rates estimated by GP4Rate are smoothly distributed within the 2D protein structures. In addition, GP4Rate segments the 2D protein structures into blocks which correspond to the true patches with different substitution rates. In contrast, the spatial distributions of substitution rates estimated by Rate4Site are far from smooth. The sites with similar substitution rates are not clustered together and do not form clearly bounded patches (Figure 3.2E and 3.2F). Thus, GP4Rate can capture the spatial correlation of substitution rates but Rate4Site cannot.

### 3.3.1 Quantitative Evaluation of Different Models

To quantitatively evaluate the performance of GP4Rate and Rate4Site, we used receiver operating characteristic (ROC) curves to measure the power of the two methods. ROC curves are widely used to evaluate the accuracy of binary classifiers. The area under a ROC curve is usually used as a measure of the power of the corresponding method. To apply ROC curves to the simulated datasets, we must divide the amino acid sites into two categories, functional sites and nonfunctional sites, before generating simulated alignments. The functional sites are used as true positives while the nonfunctional sites are used as true negatives. In the 2D toy protein simulations, functional sites evolved at the lower rate (0.2) while nonfunctional sites evolved at the higher rate (1.8). Then, the ROC curves were created by plotting the average true positive rates *versus* the average false positive rates using the ROCR library in R (Sing *et al.*, 2005). As shown in Figure 3.3A and 3.3B, the areas under the ROC curves generated by GP4Rate are larger than those generated by Rate4Site, which

suggests that GP4Rate outperforms Rate4site.

ROC curves measure whether a model can distinguish slowly evolved functional sites from the other sites. If a model can assign relatively low substitution rates to slowly evolved sites and relatively high rates to the other sites, it will perform well in the evaluations based on ROC curves. However, ROC curves cannot capture potential systematic biases of the model. For example, if the model adds a constant bias to the site-specific substitution rates, its ROC curves will be exactly the same regardless of the magnitude of the constant bias. Therefore, we used a simple loss function complementary with the ROC curves to capture any potential systematic biases of the estimated site-specific substitution rates. The loss function is defined by the following formula

$$\text{Loss}(\hat{\mathbf{\Phi}}, \mathbf{\Phi}^{\text{True}}) = \Sigma_{i=1}^{N}(\hat{\Phi}_i - \Phi_i^{\text{True}})^2, \tag{3.1}$$

in which $N$ is the total number of sites in the alignment, while $\Phi_i^{\text{True}}$ and $\hat{\Phi}_i$ are the true and estimated log substitution rates at site $i$, respectively. The log values of site-specific substitution rates are used in the right-hand side of Equation 3.1, since we want to emphasize the differences between low substitution rates. It is desirable because both GP4Rate and Rate4Site were designed to detect conserved regions with low substitution rates. Unlike ROC curves, a model which introduces a larger systematic bias will have a higher average loss than a model which introduces a smaller bias.

We plotted the losses of both GP4Rate and Rate4Site in the 2D toy protein simulations. As shown in Figure 3.3C and 3.3D, GP4Rate outperforms Rate4Site, as evident by the lower losses produced by GP4Rate (paired Wilcoxon test, $p$ values $<$

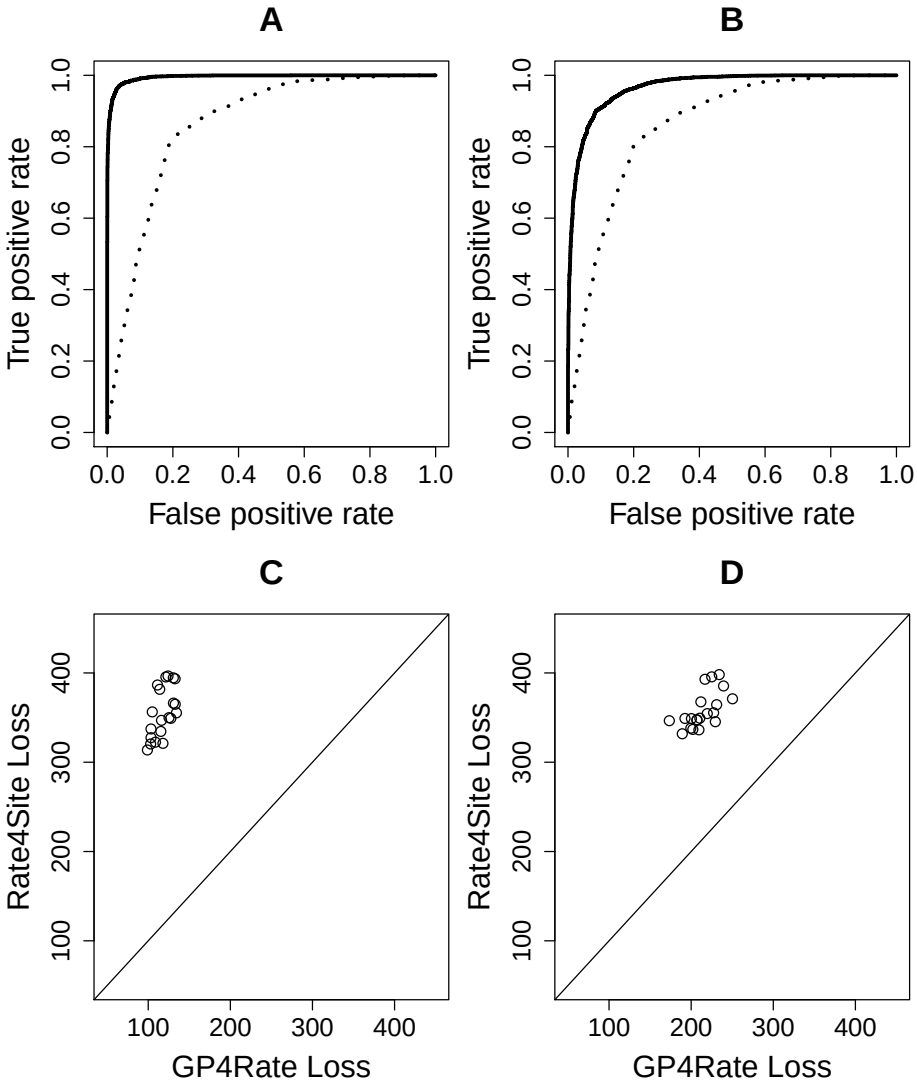


Figure 3.3: The quantitative comparison of GP4Rate and Rate4Site in the 2D toy protein simulations. (A) the ROC curves of GP4Rate and Rate4Site in the first configuration; (B) the ROC curves of GP4Rate and Rate4Site in the second configuration; (C) the losses of GP4Rate and Rate4Site in the first configuration; (D) the losses of GP4Rate and Rate4Site in the second configuration. In the ROC curves, the solid lines correspond to the performance of GP4Rate while the dotted lines correspond to the performance of Rate4Site. In the plots of losses, each point corresponds to a simulated alignment. The losses of the two methods are calculated by Equation 3.1.

$10^{-6}$ for both of the two configurations). The improved accuracy originates from GP4Rate's ability to model the spatial correlation of site-specific substitution rates, since the performance gap between GP4Rate and Rate4Site becomes smaller in the second configuration which consists of smaller conserved and variable patches.

GP4Rate has two hyperparameters, i.e. the characteristic length scale $l$ and the signal standard deviation $\sigma$, which model the strength of spatial correlation of substitution rates and the marginal variation of substitution rate at a single site, respectively. An advantage of GP4Rate over the sliding window methods is that the hyperparameters can be learned from the data. In contrast, the window size of the sliding window methods must be predefined before analyses. To show that GP4Rate can learn the hyperparameters from the data, we plotted the estimated median hyperparameters of the simulated alignments. As shown in Figure 3.4A, the characteristic length scales $l$ estimated in the first configuration are about 3 fold larger than those estimated in the second configuration. Because the patches are much larger in the first configuration, the result suggests that GP4Rate can learn the magnitude of the spatial correlation of substitution rates from the data. The estimated signal standard deviations $\sigma$ in the two configurations are similar, which matches the intuition that the two configurations are similar except in the strength of spatial correlations of substitution rates.

In summary, when spatial correlation of substitution rates exists in proteins, GP4Rate always outperforms Rate4Site. However, the spatial correlation of site-specific substitution rates may be insignificant in some proteins. Therefore, we also evaluated both GP4Rate and Rate4Site in simulated alignments in which the spatial correlation of site-specific substitution rates is absent. These simulated alignments

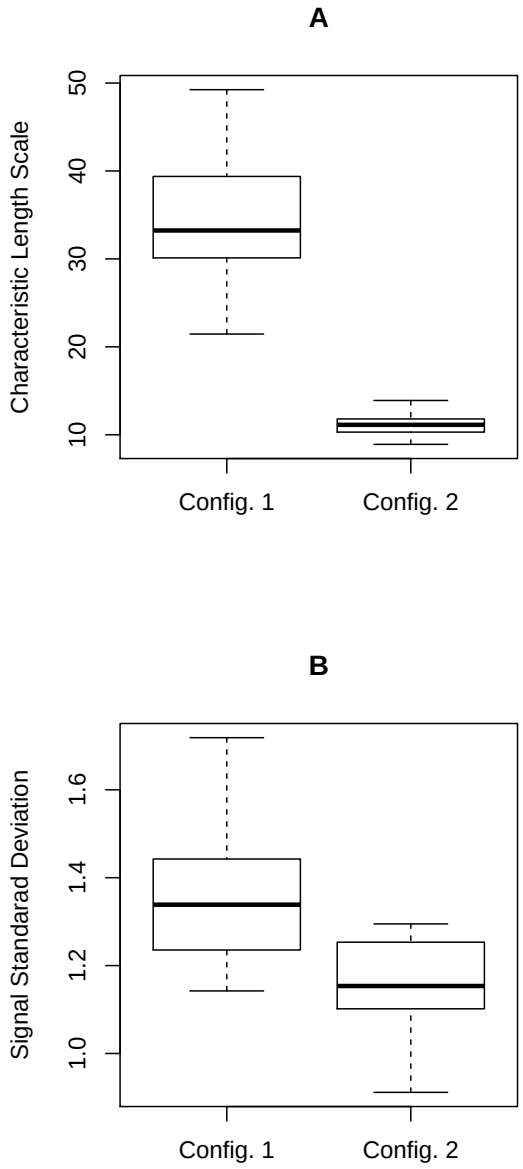


Figure 3.4: The hyperparameters estimated by GP4Rate in the 2D toy protein simulations. The unit of the characteristic length scale is Å while the signal standard deviation is unitless. (A) the estimated characteristic length scale; (B) the estimated signal standard deviation.

were generated by randomly shuffling the columns in each alignment in the first spatial configuration of substitution rates (Figure 3.2A). The permutations of alignments destroyed the spatial patten of site-specific substitution rates. Here we only summarize the performance of GP4Rate and Rate4Site in the permuted alignments and more details can be found in the Supplementary Material. The absence of spatial correlation results in close-to-zero characteristic length scales in GP4Rate, which confirms that GP4Rate can detect the absence of spatial correlation when there is none. Plots of ROC curves show that GP4Rate and Rate4Site have effectively the same power to distinguish slowly evolved sites from the other sites. In contrast, when we use the loss function (Equation 3.1) to measure the accuracy of estimated substitution rates, GP4Rate is less accurate than Rate4Site. Nevertheless, GP4Rate and Rate4Site have similar power to find slowly evolved functional sites, since in practice it is the relative rankings of sites instead of their absolute substitution rates tell us which sites may be more likely to be functional.

### 3.3.2 Realistic Simulations

We generated a second set of simulated alignments based on more realistic assumptions. The basic idea is that if we have a large number of highly diverged sequences, a simple method which does not consider the spatial correlation of substitution rates may accurately estimate the site-specific substitution rates because of the rich information in a very large dataset. We may generate simulated alignments based on the real protein tertiary structure and the presumably accurately estimated site-specific substitution rates. These simulated alignments may have similar features as real proteins.

In this set of simulations, we used the same phylogenetic tree (Figure 3.1A) and the JTT substitution model (Jones *et al.*, 1992; Kosiol and Goldman, 2005) used in the 2D toy protein simulations. The protein tertiary structure and the site-specific substitution rates were based on a real protein, B-cell lymphoma extra large (Bcl-xL). This protein has been studied using Rate4Site and the two predicted conserved patches coincide with the regions with known functions (Glaser *et al.*, 2003). We downloaded the protein tertiary structure of Bcl-xL from Protein Data Bank (PDB ID: 1MAZ (Muchmore *et al.*, 1996)). The site-specific substitution rates estimated by Rate4Site were obtained from the ConSurf-DB database (Goldenberg *et al.*, 2009). In ConSurf-DB, 131 unique homologs of Bcl-xL were automatically collected and then Rate4Site was applied to estimate the site-specific substitution rates. Because of the very large number of sequences in the dataset, the estimation of site-specific substitution rates may be relatively accurate. We generated 20 simulated alignments based on the above assumptions and both GP4Rate and Rate4Site were applied to the simulated alignments using the same setting described in the 2D toy protein simulations.

To evaluate the performance of GP4Rate and Rate4Site by ROC curves, we divided the sites into two categories before generating simulated alignments: slowly evolved functional sites and others. Based on the site-specific substitution rates reported by ConSurf-DB, the 10 percent most slowly evolved sites were considered to be functional while the others were not. As shown in Figure 3.5A, GP4Rate is more powerful to distinguish slowly evolved sites from the other sites, since the area under the ROC curve of GP4Rate is larger than that of Rate4Site. In addition, based on the loss function defined by Equation 3.1, GP4Rate produces lower losses in 18

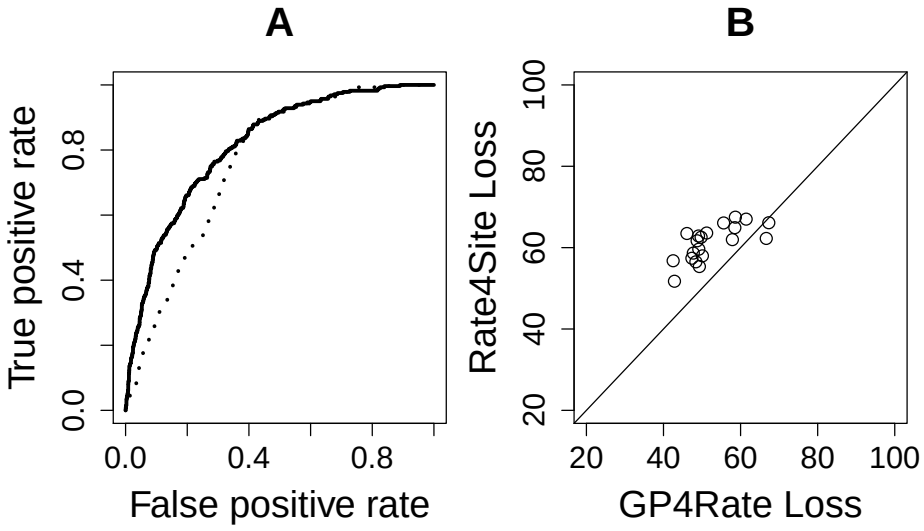


Figure 3.5: The quantitative comparison of GP4Rate and Rate4Site in the realistic simulations. (A) the ROC curves of GP4Rate and Rate4Site in the realistic simulations; (B) the losses of GP4Rate and Rate4Site in the realistic simulations. In the ROC curves, the solid line corresponds to the performance of GP4Rate while the dotted line corresponds to the performance of Rate4Site. In the plot of losses, each point corresponds to a simulated alignment. The losses of the two methods are calculated by Equation 3.1.

out of the 20 simulated alignments (Figure 3.5B) and the median loss of GP4Rate is significantly smaller than that of Rate4Site (paired Wilcoxon test, $p$ value $< 10^{-4}$). Therefore, GP4Rate still outperforms Rate4Site in the realistic simulations.

### 3.3.3 Case Study of B7-1 Genes

The B7-1 (CD80) family is a member of the immunoglobulin superfamily (IgSF) and is critical for the regulation of immune responses (Collins *et al.*, 2005). The protein tertiary structure of the human B7-1 protein has been determined (Ikemizu *et al.*, 2000; Stamper *et al.*, 2001). The human B7-1 protein consists of two IgSF domains (IgV and IgC), each of which shows an anti-parallel $\beta$ sandwich structure (Ikemizu *et al.*, 2000). We applied GP4Rate and Rate4Site to 7 mammalian B7-1 sequences downloaded from the NCBI HomoloGene database (Sayers *et al.*, 2012) and compared their performances. The N-terminal and C-terminal sequences were trimmed in the alignment, because the corresponding atoms are absent in the X-ray crystal structure. The resulting alignment consists of 199 amino acid sites. Then the phylogenetic tree was inferred by PhyML with the JTT+$\Gamma$ model (Guindon and Gascuel, 2003). The protein sequences in the alignment are very similar to each other as evident by the lack of gaps in the alignment (data not shown). Therefore, the information in each site in the alignment is very limited and it is hard to infer site-specific substitution rates accurately.

We used the human B7-1 protein structure (PDB ID: 1I8L (Stamper *et al.*, 2001)) to calculate the pairwise Euclidean distances between the $\alpha$-carbon atoms of amino acids. Then, we applied GP4Rate to the B7-1 alignment to infer site-specific substitution rates. We ran two independent MCMC chains for $10^6$ iterations, and the

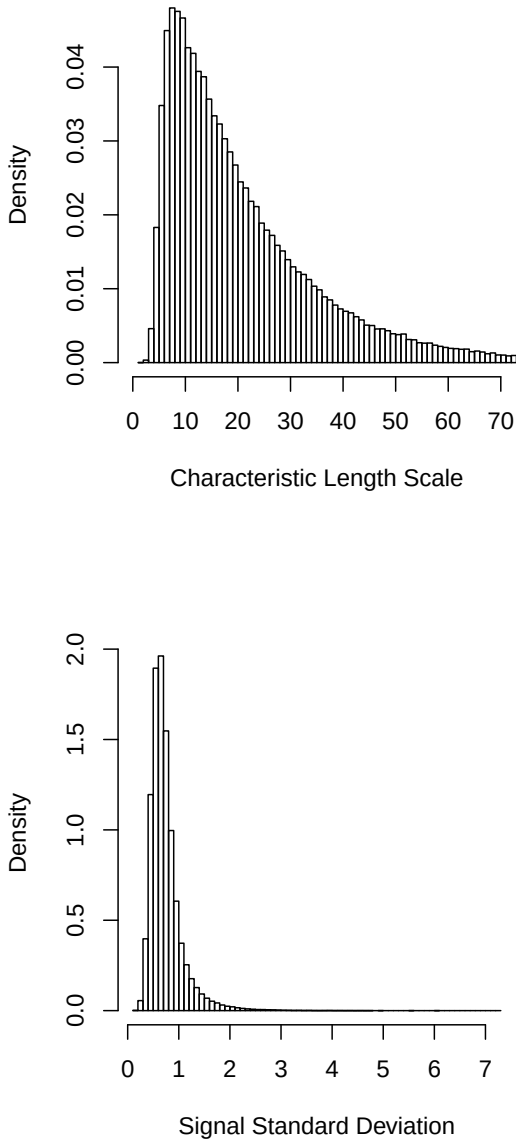


Figure 3.6: The empirical marginal density functions of the hyperparameters in the case study of B7-1 genes. The unit of the characteristic length scale is Å while the signal standard deviation is unitless.

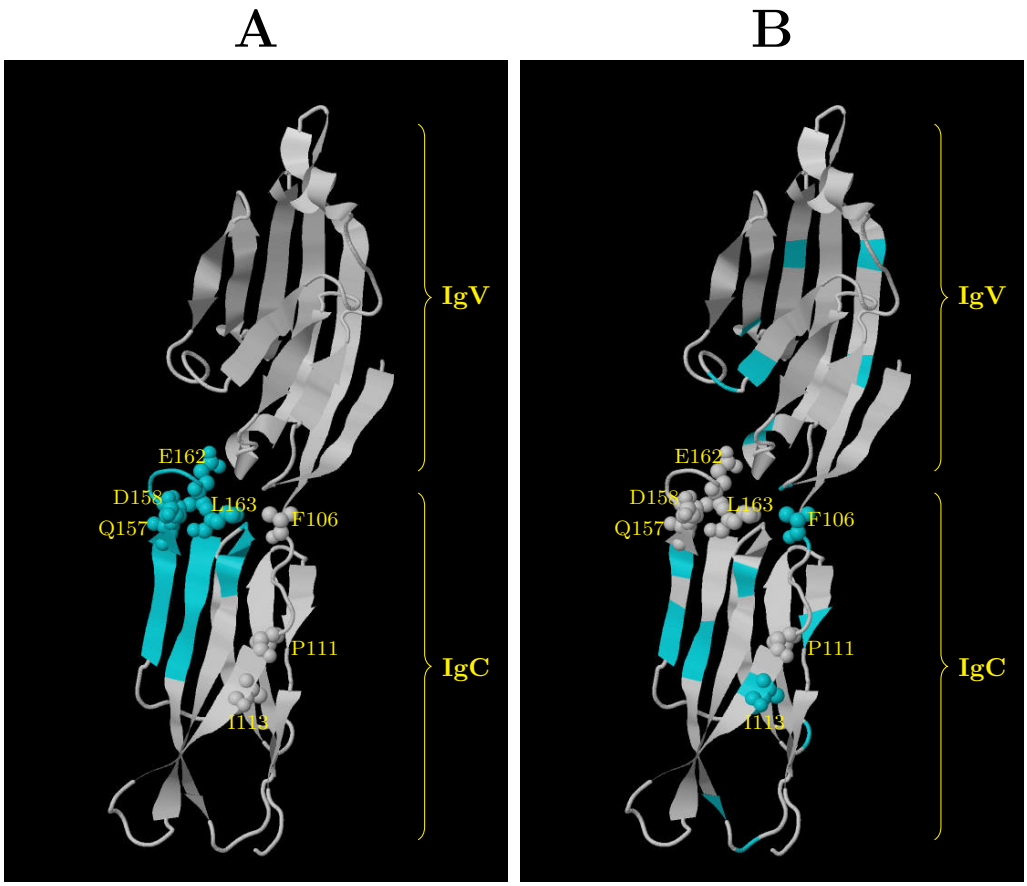


Figure 3.7: The locations of the 20 most conserved sites in the protein tertiary structure of the human B7-1 protein (PDB ID: 1I8L). The blue sites are the 20 most conserved sites and the space-filled atoms correspond to the experimentally verified functional sites in the IgC domain (Peach *et al.*, 1995). The experimentally verified functional sites in the IgV domain are not shown. The protein structures are visualized by Jmol (Willighagen and Howard, 2007). A list of the most conserved sites can be found in the Supplementary Material. (A) the 20 most conserved sites predicted by GP4Rate; (B) the 20 most conserved sites predicted by Rate4Site.

first 30% of the iterations were discarded as burn-in. We first estimated the posterior marginal distributions of hyperparameters based on the MCMC samples. As shown in Figure 3.6, the estimated characteristic length scale $l$ is significantly higher than 0, which confirms that the substitution rates are correlated in real proteins. The presence of spatial correlation of substitution rates may facilitate the discovery of slowly evolved functional regions. To test this hypothesis, the mean site-specific substitution rates of the MCMC samples were calculated and the 20 most slowly evolved sites were considered to be functional. Then, the 20 most slowly evolved sites were superimposed onto the protein tertiary structure (PDB ID: 1I8L (Stamper *et al.*, 2001)). As shown in Figure 3.7A, the slowly evolved sites predicted by GP4Rate are not randomly distributed and instead form a single large region in the IgC domain. A systematic mutagenesis study has suggested that the IgC domains are important for binding CTLA-4 and CD28 (Peach *et al.*, 1995), even though the effects of the IgC domain may be indirect (Stamper *et al.*, 2001). To test whether the predicted slowly evolved sites overlap with the experimentally verified functional sites (Peach *et al.*, 1995), the 7 experimentally verified functional sites in the IgC domain were mapped onto the human B7-1 structure (Figure 3.7A). Clearly 4 experimentally verified functional sites in the IgC domain, i.e. Q157, D158, E162, and L163, are within the slowly evolved patch predicted by GP4Rate, which highlights the potential usefulness of GP4Rate.

To compare GP4Rate with Rate4Site, we also applied Rate4Site to the same dataset. The superimposition of the 20 most slowly evolved sites predicted by Rate4Site is shown in Figure 3.7B. The sites predicted by Rate4Site are present in both the IgV

and IgC domains and do not form clearly bounded regions. Even though 2 experimentally verified functional sites in the IgC domain, i.e. F106 and I113, overlap with the sites predicted by Rate4Site, the 4 experimentally verified functional sites detected by GP4Rate do not overlap with the sites predicted by Rate4Site. Therefore, GP4Rate and Rate4Site can provide complementary insights to real data.

To investigate which model, GP4Rate or Rate4Site, fits the B7-1 dataset better, we performed a Bayesian model comparison. The direct comparison between GP4Rate and Rate4Site is impractical, because Rate4Site is based on the maximum likelihood principle instead of the Bayesian principle. However, it is not very difficult to develop a Bayesian version of Rate4Site by specifying a prior distribution over parameters. Therefore, we developed a Bayesian version of Rate4Site and compared it with GP4Rate. Details of the Bayesian model comparison can be found in the Supplementary Material and we only summarize the results here. We compared the site-specific substitution rates estimated by the original Rate4Site and its Bayesian version and found that the two programs produced essentially the same result. Therefore, the marginal likelihood estimated by the Bayesian version of Rate4Site may be used to evaluate how good the original Rate4Site fits the B7-1 dataset. The log marginal likelihood of GP4Rate is equal to $-1705.1$ while the log marginal likelihood of the Bayesian Rate4Site is equal to $-1710.9$, which suggests a very large Bayes factor of GP4Rate compared with the Bayesian Rate4Site ($\mathcal{BF} = e^{-1705.1+1710.9} = 330.3$). Therefore, GP4Rate fits the B7-1 dataset much better than the Bayesian Rate4Site.

## 3.4 Discussion

Many phylogenetic methods have been developed to identify slowly evolved amino acid sites which may be functional. However, the most widely used methods, e.g. Rate4Site, ignore the spatial correlation of site-specific substitution rates. Some other methods use the sliding-window framework to capture the spatial correlation of substitution rates, but the statistical method for choosing the optimal window size is largely unknown. Since the strength of the spatial correlation of substitution rates is unknown in most of proteins, the sliding window methods are problematic in real data analyses. In GP4Rate, both of the two issues are solved under a Bayesian statistical framework. By using the Gaussian process to define the prior distribution of the site-specific log substitution rates, GP4Rate can naturally model the spatial clustering of functionally important sites and the hyperparameters which measure the strength of spatial correlation can be inferred from the data instead of being manually specified before the analyses.

In simulated datasets, GP4Rate significantly outperforms Rate4Site. The power of GP4Rate is mainly derived from the fact that GP4Rate has the added ability to model the spatial correlation of substitution rates. By borrowing statistical information from neighbor sites with similar substitution rates, GP4Rate can estimate the site-specific substitution rates with a much higher accuracy than Rate4Site. In the case study of B7-1 genes, GP4Rate predicted a slowly evolved functional patch in the protein tertiary structure and 4 sites within the region are well supported by experimental evidence. In contrast, the slowly evolved sites predicted by Rate4Site are scattered and do not form clearly bounded regions. In addition, we have shown that GP4Rate fits the B7-1 dataset much better than Rate4Site based on Bayesian

model comparison.

The performance gap between GP4Rate and Rate4Site will be maximized when the protein sequences are very similar to each other and the spatial correlation is strong. Therefore, GP4Rate is most suitable to analyze small gene families, e.g. new genes or small gene families derived from recent gene duplication events. When the spatial correlation of substitution rates is weak, GP4Rate and Rate4Site may generate similar results. For example, we applied GP4Rate to 38 RH1 genes (Yokoyama *et al.*, 2008) and found that the spatial correlation of substitution rates is much weaker in the RH1 dataset than that in the B7-1 dataset (data not shown). In this case, the difference between GP4Rate and Rate4Site is subtle. Therefore, a rigorous model comparison as shown in the case study of B7-1 genes may be important in data analyses.

Because GP4Rate is based on MCMC simulations, it is slower than Rate4Site. For example, it took about 1 CPU day for GP4Rate to analyze the B7-1 dataset. However, GP4Rate is still very useful for small scale problems, e.g. guiding mutagenesis experiments, since the experimental time is much longer than the execution time of GP4Rate. The time cost of GP4Rate can be reduced in the future using advanced algorithms, e.g. more efficient MCMC sampling algorithms or sparse approximations of the Gaussian process (Vanhatalo and Vehtari, 2007). The most time consuming step of GP4Rate is the Cholesky decomposition whose time complexity is a cubic function of the number of sites in the alignment. In practice, a simple method to reduce the computational time is to perform the analyses based on a selected subset of amino acid sites. For example, it is well known that surface residues are more likely to be involved in interactions with other proteins or ligands. If these interactions are

most interesting to users, a fast analysis based only on the surface residues may be appropriate.

In addition to modeling the spatial correlation of site-specific substitution rates, protein tertiary structures have been used to improve phylogenetic models and the estimation of site-specific substitution rates in a few other studies (Robinson *et al.*, 2003; Rodrigue *et al.*, 2005, 2006; Conant and Stadler, 2009; Meyer and Wilke, 2013; Meyer *et al.*, 2013). These methods can be roughly divided into two categories. The first category of models assumes that the fixation probability of new mutations is determined by how the mutations influence the stability of the protein (Robinson *et al.*, 2003; Rodrigue *et al.*, 2005, 2006). Typically it is assumed that mutations which stabilize the protein structure are more likely to be fixed than mutations which destabilize the protein structure. Unlike this category of models, the Phylo-GPM framework does not provide a mechanistic interpretation for the estimated substitution rates. However, GP4Rate may be more powerful to identify functional regions which are not directly relevant to the stability of proteins. The second category of models assumes that the site-specific substitution rates or $dN/dS$ ratios are influenced by the local environment of the focal site in the protein tertiary structure (Conant and Stadler, 2009; Meyer and Wilke, 2013; Meyer *et al.*, 2013). For example, it has been shown that the $dN/dS$ ratio of a site is influenced by its relative solvent accessibility (RSA) (Conant and Stadler, 2009; Meyer and Wilke, 2013; Meyer *et al.*, 2013). It is relatively straightforward to combine the Phylo-GPM framework with local features of amino acid sites. For example, in this study we assume that the site-specific log substitution rates follow a zero-mean Gaussian distribution. We may replace the zero-mean rate vector by a new one in which the mean of log substitution

rate at a site is a linear function of its local features, e.g. RSA. It is very interesting to investigate whether adding local features to the Phylo-GPM framework improves model fitting in the future.

The Phylo-GPM framework proposed in this paper may be used as a general tool to model the spatial correlation of patterns in the protein tertiary structure. The phylogenetic hidden Markov model (Phylo-HMM) is a popular method which combines the hidden Markov model and statistical phylogenetics (Siepel and Haussler, 2004). It has been used to model the spatial correlation of evolutionary patterns along primary sequences (Yang, 1995; Felsenstein and Churchill, 1996; Siepel *et al.*, 2005, 2006; Mayrose *et al.*, 2007; Huang and Golding, 2012; De Maio *et al.*, 2013). The Phylo-GPM framework may be viewed as an extension of the Phylo-HMM to the protein tertiary structures. In the future, new methods based on the Phylo-GPM framework may be developed to identify functional divergence or positive selection in proteins.

## 3.5 Models

### 3.5.1 Overall Design of the Phylogenetic Gaussian Process Model

GP4Rate is an open-source software application written in C++ and its source code is freely available from http://info.mcmaster.ca/yifei/software.html. GP4Rate combines the protein alignment and the protein tertiary structure to infer groups of close-located functional sites evolved at low rate. We assume that the protein alignment, the phylogenetic tree, and the tertiary structure of one protein in the alignment

are provided by users. In GP4Rate, both the topology and the branch lengths of the phylogenetic tree are fixed to improve the speed of the program. In addition, we assume that the protein sequences in the alignment belong to the same gene family and have very similar functions, which implies that the functionally important sites do not vary among sequences and the site-specific substitution rates do not change over time. However, we do assume that the substitution rates can vary across different sites. The site-specific rates are used as proxies of functionality: very low substitution rates suggest the corresponding sites are functionally important.

In most molecular phylogenetic programs, e.g. Rate4Site (Mayrose *et al.*, 2004), PAML (Yang, 2007), and PhyML (Guindon and Gascuel, 2003), the site-specific substitution rates are assumed to be i.i.d. and follow a simple discrete distribution, usually the discrete Gamma distribution (Yang, 1994). Recently, Dirichlet process pirors have been used to model the variable substitution rates over sites to overcome the inflexibility of the simple discrete distributions (Huelsenbeck and Suchard, 2007), but it is still assumed that the site-specific substitution rates are i.i.d. The i.i.d. assumption implies that slowly evolved functional sites are scattered in the protein tertiary structure. The major contribution of this paper is to relax the i.i.d. assumption using the Gaussian process (Rasmussen and Williams, 2005) which can naturally capture the spatial correlation of site-specific substitution rates in the protein tertiary structure.

In GP4Rate, the parameters are estimated using the Bayesian principle. In Bayesian statistics, the parameters are random variables and the conditional distribution of parameters given data, i.e. the posterior distribution, gives us an estimation of parameters. For simplicity of presentation, first we focus on the vector of site-specific

log substitution rates, which is the collection of log values of substitution rates at all amino acid sites, and defer the discussions on the other parameters. The posterior distribution of the vector of log site-specific substitution rates can be defined by the following equation,

$$P(\mathbf{\Phi}|\mathbf{X},\mathcal{T}) \propto P(\mathbf{\Phi}) \prod_{i=1}^{N} \mathcal{L}_i(\Phi_i;\mathbf{X}_i,\mathcal{T}). \tag{3.2}$$

In the equation, $\mathbf{\Phi}$ is the vector of site-specific log substitution rates, $\mathbf{X}$ is the protein alignment while $\mathbf{X}_i$ is its $i$-th column, and $\mathcal{T}$ is the phylogenetic tree with the associated branch lengths. $\mathcal{L}_i(\Phi_i;\mathbf{X}_i,\mathcal{T})$ is the site-specific likelihood at site $i$, which is a function of the site-specific log substitution rate at site $i$. $P(\mathbf{\Phi})$ is the fundamentally important prior distribution of site-specific log substitution rates.

A realistic $P(\mathbf{\Phi})$ should be able to describe the spatial correlation of site-specific substitution rates. In GP4Rate, $P(\mathbf{\Phi})$ is specified by a zero-mean Gaussian process. A Gaussian process is a probability measure defined over a function space. In the statistical modeling of site-specific substitution rates, we are only interested in the marginal distribution of the Gaussian process over a finite set of spatial locations which correspond to the locations of residues in the protein tertiary structure. By the definition of Gaussian processes, the marginal distribution of a zero-mean Gaussian process is a zero-mean multivariate Gaussian distribution (Rasmussen and Williams, 2005). Therefore, $P(\mathbf{\Phi})$ may be rewritten in the following format,

$$P(\mathbf{\Phi}|\mathbf{D},l,\sigma) = \frac{1}{(2\pi)^{\frac{N}{2}}|\mathbf{\Sigma}(\mathbf{D},l,\sigma)|^{\frac{1}{2}}} \exp(-\frac{\mathbf{\Phi}^T\mathbf{\Sigma}(\mathbf{D},l,\sigma)^{-1}\mathbf{\Phi}}{2}). \tag{3.3}$$

The correlation of site-specific substitution rates is determined by the covariance matrix $\mathbf{\Sigma}(\mathbf{D}, l, \sigma)$, in which $\mathbf{D}$ is the pairwise distance matrix which measures the Euclidean distance between the $\alpha$-carbon atoms of amino acids in the protein tertiary structure. Furthermore, the covariance function is parameterized by two hyperparameters, $l$ and $\sigma$, which measure the strength of spatial correlation and the variation of substitution rates across sites, respectively. By plugging $P(\mathbf{\Phi}|\mathbf{D}, l, \sigma)$ and $P(l, \sigma)$, the prior distribution of the hyperparameters, into Equation 3.2, it can be expanded to the following format,

$$P(\mathbf{\Phi}, l, \sigma|\mathbf{X}, \mathbf{D}, \mathcal{T}) \propto P(l, \sigma)P(\mathbf{\Phi}|\mathbf{D}, l, \sigma)\prod_{i=1}^{N}\mathcal{L}_i(\Phi_i; \mathbf{X}_i, \mathcal{T}). \tag{3.4}$$

In the following sections, we will provide more details on the specifications of the right-hand side terms of Equation 3.4 and the MCMC algorithm for the sampling of parameters, i.e. $\mathbf{\Phi}$, $l$, and $\sigma$.

### 3.5.2 Gaussian Process as a Prior Distribution of Site-specific Log Substitution Rates

As mentioned above, $\mathbf{\Phi}$ follows a zero-mean multivariate Gaussian distribution (Equation 3.3). In the multivariate Gaussian distribution, the covariance matrix $\mathbf{\Sigma}$ is specified by a covariance function. By default, GP4Rate uses the Matérn 1.5 covariance function,

$$\Sigma_{ij} = \sigma^2(1 + \frac{\sqrt{3}d_{ij}}{l})\exp(-\frac{\sqrt{3}d_{ij}}{l}) + \mathbb{I}_{i=j}(i, j)J^2. \tag{3.5}$$

In the equation, $\Sigma_{ij}$ is an element in the covariance matrix $\mathbf{\Sigma}(\mathbf{D}, l, \sigma)$ while $d_{ij}$ is an element in the distance matrix $\mathbf{D}$ which measures the Euclidean distance between site $i$

and site $j$ in the protein tertiary structure. $\mathbb{I}_{i=j}(i,j)$ is an indicator function which is equal to 1 if site $i$ and site $j$ are the same site and equal to 0 otherwise. The covariance function contains two free parameters, $l$ and $\sigma$. $l$ is the characteristic length which determines the strength of the spatial correlation of substitution rates. If it is small, the spatial correlation is weak and only nearby sites have similar log substitution rates. Instead, if it is large, the spatial correlation is strong and distant sites can have similar log substitution rates. $\sigma$ is the signal standard deviation which measures the marginal variation of log substitution rates at a single site. Small $\sigma$ implies that the variation of log substitution rates is small. $J$ is a fixed "jitter" term which introduces a small amount of noise to the diagonal elements in $\mathbf{\Sigma}(\mathbf{D}, l, \sigma)$. The "jitter" term ensures that the Cholesky decomposition, a critical numerical algorithm in the MCMC simulations, is numerically stable and improves the mixing of the MCMC simulations (Neal, 1997). The "jitter" term is usually a small positive number (e.g. $J = 0.1$), so it does not significantly change the behavior of the covariance function (Neal, 1997). Clearly Equation 3.5 implies that the covariance of log substitution rates are decreasing with increasing Euclidean distance between two amino acid sites, which is compatible with our intuition that nearby sites tend to have similar substitution rates due to similar functions.

In addition to the Matérn 1.5 covariance function, GP4Rate has two alternative covariance functions for users to choose. One is the Matérn 2.5 covariance function,

$$\Sigma_{ij} = \sigma^2(1 + \frac{\sqrt{5}d_{ij}}{l} + \frac{5d_{ij}^2}{3l^2}) \exp(-\frac{\sqrt{5}d_{ij}}{l}) + \mathbb{I}_{i=j}(i,j)J^2. \tag{3.6}$$

The other is the widely used squared-exponential covariance function,

$$\Sigma_{ij} = \sigma^2 \exp(-\frac{d_{ij}^2}{2l}) + \mathbb{I}_{i=j}(i,j)J^2. \tag{3.7}$$

The three covariance functions are all special cases of the general Matérn covariance function (Rasmussen and Williams, 2005). The major difference between them is that the three covariance functions describe different levels of smoothness in the spatial distribution of site-specific log substitution rates (Rasmussen and Williams, 2005). In the squared-exponential covariance function, the site-specific log substitution rates are smoothly distributed in the protein tertiary structure. Therefore, it is most suitable to model proteins with relatively large functional regions. In contrast, the Matérn 1.5 covariance function is the least smooth one and is suitable to model proteins with small functional patches. In this paper, we used the Matérn 1.5 covariance function in all analyses to allow for proteins that may have relatively small functional patches and could have nearby sites with very different substitution rates.

The hyperparameters in the covariance functions, i.e. $l$ and $\sigma$, follow a prior distribution $P(l, \sigma)$. We assume that the characteristic length, $l$, and the signal standard deviation, $\sigma$, are independent and follow exponential distributions. Therefore, the prior distribution is defined by the following probability density function,

$$P(l, \sigma) = m_l^{-1} m_\sigma^{-1} \exp(-\frac{l}{m_l}) \exp(-\frac{\sigma}{m_\sigma}). \tag{3.8}$$

We choose $m_l$ and $m_\sigma$ to be large so that the prior distribution has relatively weak information.

### 3.5.3 Approximation of the Phylogenetic Likelihood Function

To fully define the unnormalized posterior distribution (Equation 3.4), the likelihood $\mathcal{L}(\Phi_i; \mathbf{X}_i, \mathcal{T})$ must be specified. We follow the standard phylogenetic model first described by Felsenstein (Felsenstein, 1981). We assume that the substitution model in the phylogenetic likelihood function is fixed to the JTT model (Jones *et al.*, 1992; Kosiol and Goldman, 2005) while the phylogenetic tree is fixed to the one provided by the users. The likelihood can be calculated by the pruning algorithm and the gaps in the alignment may be treated as missing data (Felsenstein, 1981). However, the calculation of the likelihood function can easily become the most time consuming step in the MCMC sampling, because we need to evaluate the likelihood millions of times. We have applied a simple linear interpolation method to reduce the computational time of the likelihood evaluation (Press *et al.*, 1992). GP4Rate calculates the site-specific log likelihoods at a set of evenly spaced substitution rates and then approximates the site-specific log likelihoods at other rates by interpolation. Note that the linear interpolation is performed based on the site-specific substitution rates while $\mathbf{\Phi}$ is the vector of their log values, so an exponential transformation, i.e. $\exp(\Phi_i)$, must be performed for each site $i$ before the interpolation. By default, GP4Rate calculates and caches the site-specific log likelihoods at 4000 evenly spaced substitution rates, ranging from $10^{-6}$ to 20. In each step of the likelihood calculation, if $\exp(\Phi_i)$ is between $10^{-6}$ and 20, the corresponding site-specific log likelihood is approximated by the following formula,

$$\log(\mathcal{L}_i(\Phi_i; \mathbf{X}_i, \mathcal{T})) = \log(\mathcal{L}_{i0}) + (\log(\mathcal{L}_{i1}) - \log(\mathcal{L}_{i0}))\frac{\exp(\Phi_i) - R_{i0}}{R_{i1} - R_{i0}}. \tag{3.9}$$

On the right hand side, $R_{i0}$ and $R_{i1}$ are the two cached substitution rates which are closest to $\exp(\Phi_i)$, while $\log(L_{i0})$ and $\log(L_{i1})$ are the site-specific log likelihoods of $R_{i0}$ and $R_{i1}$, respectively. In practice, $\exp(\Phi_i)$ is rarely bigger than 20 or smaller than $10^{-6}$. If it is indeed outside this, the log likelihood at the closest boundary is used as the approximate log likelihood.

### 3.5.4 Markov Chain Monte Carlo Sampling

GP4Rate uses MCMC simulations to sample parameters from their posterior distribution. The algorithm follows previous studies by Neal (1997, 1999). As described in the previous sections, the parameters in GP4Rate have two components. The first one is $\mathbf{\Phi}$ and the second one consists of $\sigma$ and $l$. In each iteration, the two components are sequentially updated by the Metropolis algorithm with symmetric proposals (Metropolis *et al.*, 1953; Hastings, 1970).

To update $\mathbf{\Phi}$, GP4Rate uses a proposal distribution suggested by Neal (Neal, 1997),

$$\mathbf{\Phi}' = \mathbf{\Phi} + \epsilon \mathbf{L}\mathbf{z}. \tag{3.10}$$

In the equation, $\mathbf{\Phi}$ is the current vector of site-specific log substitution rates while $\mathbf{\Phi}'$ is the new proposal. $\mathbf{L}$ is the Cholesky decomposition of the covariance matrix $\mathbf{\Sigma}(\mathbf{D}, l, \sigma)$ and $\mathbf{z}$ is a vector of independent standard Gaussian variables. The proposal distribution is tuned by the constant, $\epsilon$. A large $\epsilon$ leads to large changes of $\mathbf{\Phi}$ while small $\epsilon$ leads to small changes. $\epsilon$ is chosen to make the acceptance rate of new proposals close to 0.25.

Instead of updating $\sigma$ and $l$ in the original scale, we transform them to the log scale. The use of a log scale removes the boundaries of the two parameters and makes the

MCMC sampling of $\sigma$ and $l$ independent from the scale of the data (Neal, 1997). The two parameters are updated by a sliding window method with a bivariate Gaussian proposal (Neal, 1999). The Gaussian proposal is tuned so that the acceptance rate of new proposals is close to 0.25.

In practice, the update of $\mathbf{\Phi}$ is much faster than the update of $\sigma$ and $l$, because the update of $\sigma$ and $l$ requires a Cholesky decomposition whose time complexity is $O(N^3)$, in which $N$ is the total number of sites in the alignment. Therefore, it is reasonable to update $\mathbf{\Phi}$ more often than $\sigma$ and $l$ (Neal, 1997). In each iteration $\mathbf{\Phi}$ is updated 50 times while the pair of $\sigma$ and $l$ is updated once. After every 10 iterations, the values of $l$, $\sigma$, and $\exp(\mathbf{\Phi})$ are recorded.

## 3.6 Acknowledgments

We thank Shozo Yokoyama for kindly sharing the RH1 dataset used in an early version of this chapter and Julien Dutheil for kindly providing help on the Bio++ library. We also thank Nicolas Rodrigue, Jeffrey Thorne, and Claus Wilke for their insightful comments on this work.

# Chapter 4

# FuncPatch: A Web Server for the Fast Bayesian Inference of Conserved Functional Patches in Protein 3D Structures

## 4.1 Abstract

A number of statistical phylogenetic methods have been developed to infer conserved functional sites or regions in proteins. Many methods, e.g. Rate4Site, apply the standard phylogenetic models to infer site-specific substitution rates and totally ignore the spatial correlation of substitution rates in protein tertiary structures, which significantly reduces their power to identify conserved functional patches in protein tertiary structures. The 3D sliding window method has been proposed to infer conserved

functional patches in protein tertiary structures, but the window size, which reflects the strength of the spatial correlation, must be predefined and is not inferred from data. We recently developed GP4Rate to solve these problems under the Bayesian framework. Unfortunately, GP4Rate is computationally slow. Here we present an intuitive web server, FuncPatch, to perform a fast approximate Bayesian inference of conserved functional patches in protein tertiary structures. Both simulations and two case studies of the MAPK1 genes and the SMAD genes suggest that FuncPatch is a good approximation to GP4Rate. However, FuncPatch is orders of magnitudes faster than GP4Rate. In addition, simulations demonstrate that FuncPatch is more robust and generally more powerful than Rate4Site and the 3D sliding window method. The functional patches predicted by FuncPatch in the two case studies are supported by experimental evidence, which corroborates the usefulness of FuncPatch. FuncPatch is freely available from http://info.mcmaster.ca/yifei/FuncPatch.

## 4.2 Introduction

Because of the fast development of sequencing techniques, the amount of sequence data is increasing exponentially. The best ways to extract biological insights from massive sequence data have become important questions in biology. Comparisons between homologous sequences from different species are very common strategies to analyze biological sequences. For example, given a set of homologous protein sequences from different species, we can compare these sequences to identify conserved amino acid sites. These conserved amino acid sites are more likely to be functionally important, since mutations at these sites are more likely to be deleterious.

To infer the conservation levels of amino acid sites, we need evolutionary models to describe the substitution process of amino acids in the evolutionary history. The simplest idea is to use the standard statistical phylogenetic models (Felsenstein, 1981) to infer the site-specific substitution rates and the sites with low substitution rates may be considered to be functional. For example, a widely used web server, ConSurf (Glaser *et al.*, 2003; Ashkenazy *et al.*, 2010), uses the site-specific substitution rates estimated by the Rate4Site program (Mayrose *et al.*, 2004) to infer the conservation levels of amino acid sites. Then, the conservation scores are mapped onto the protein tertiary structure to get insights on the possible functions of the highly conserved sites (Glaser *et al.*, 2003). While the standard phylogenetic models are useful tools for inferring conserved sites in proteins, they typically model the substitution rate variation by some discretized distributions, e.g. the discretized Gamma distribution (Yang, 1994), and ignore the spatial correlation of substitution rates in protein tertiary structures. However, it is well known that functional amino acids are clustered together in protein tertiary structures in many proteins and modeling the spatial clustering can improve the prediction of functional sites (Madabushi *et al.*, 2002; Panchenko *et al.*, 2004). The independence assumption of site-specific substitution rates in the standard phylogenetic methods makes it difficult to infer conserved functional patches in protein tertiary structures.

A few methods have been proposed to relax the independence assumption of site-specific substitution rates to predict conserved protein patches in protein tertiary structures (Dean and Golding, 2000; Nimrod *et al.*, 2005; Landgraf *et al.*, 2001; Capra and Singh, 2007; Panchenko *et al.*, 2004). These methods are useful tools for inferring conserved protein patches. However, most of these methods are based on the 3D

sliding window framework (Dean and Golding, 2000; Landgraf *et al.*, 2001; Capra and Singh, 2007; Panchenko *et al.*, 2004) or other heuristic algorithms (Nimrod *et al.*, 2005). The common problem of these methods is that they cannot infer the strength of the spatial correlation of substitution rates which in turn makes the inference of site-specific substitution rates unreliable. For example, the window size in the 3D sliding window method is typically predefined before analyses. However, the strength of spatial correlation may vary in different proteins, which suggests the optimal window size may vary in different datasets (Suzuki, 2004).

Recently, we developed a phylogenetic Gaussian process model, GP4Rate, which combines standard phylogenetics and Gaussian processes to infer conserved functional regions in protein tertiary structures (Huang and Golding, 2014). Using the Gaussian process as the prior distribution of log values of site-specific substitution rates, GP4Rate naturally captures the spatial correlation of substitution rates in the protein tertiary structure. In addition, GP4Rate can infer the strength of the spatial correlation of substitution rates based on the Bayesian principle. Therefore, it overcomes the drawbacks of the 3D sliding window method and other heuristic methods.

However, GP4Rate is based on Markov chain Mote Carlo (MCMC) methods to generate samples from the posterior distribution of parameters. Because MCMC methods are generally very slow, it typically takes a few hours to a few days to analyze a gene family using GP4Rate. In addition, the command-line interface of GP4Rate may not be intuitive to many experimental biologists. Here we report a new algorithm, FuncPatch, which is designed as a fast approximation to GP4Rate. This method is fast enough to be implemented in a web server. Using a simplified likelihood function and a Laplace approximation, FuncPatch is orders of magnitudes

faster than GP4Rate and the analyses of most gene families can be finished in a few minutes. Simulations and two case studies of the MAPK1 genes and the SMAD genes demonstrate that FuncPatch is a very accurate approximation to GP4Rate and FuncPatch outperforms Rate4Site and the 3D sliding window method. The two case studies also suggest that the spatial correlation of substitution rates is present in real data and that the strength of spatial correlation may vary in different protein families.

## 4.3 Models

### 4.3.1 Overview of FuncPatch

We developed FuncPatch, an algorithm for the fast Bayesian inference of conserved patches in protein tertiary structures. FuncPatch is designed as a fast approximation to GP4Rate (Huang and Golding, 2014) which combines phylogenetics and Gaussian processes to infer conserved functional patches in protein tertiary structures. In this section, we describe the basic idea of FuncPatch in simple terms and ignore mathematical details. Thereafter, we describe technical details of the parameterization and implementation of FuncPatch. Readers who are not very familiar with computational statistics may read this section but skip these technical sections. FuncPatch assumes that the users provide an alignment of proteins and a representative protein tertiary structure. Similar to GP4Rate (Huang and Golding, 2014), it combines the information from the protein alignment and the protein tertiary structure to infer site-specific substitution rate at each amino acid site. The estimated substitution rates are used as proxies of functionality: lower substitution rates suggest that the corresponding sites are more likely to be functionally important.

If conserved functional sites form tertiary patches, e.g. protein-protein interaction interfaces, the site-specific substitution rates may be positively correlated over the protein tertiary structure. Thus physically closely located sites are more likely to have similar substitution rates. Modeling the spatial correlation of substitution rates may in turn improve the prediction of conserved functional patches, since the inference of substitution rate at a focal amino acid site can borrow "statistical information" from closely located sites with similar substitution rates (Huang and Golding, 2014). FuncPatch uses the Bayesian statistical principle to infer site-specific substitution rates. To apply the Bayesian principle, we need to specify a prior distribution over site-specific substitution rates and a likelihood function which describes the probability of the observed data given the site-specific substitution rates. In FuncPatch, the protein tertiary structure is used to specify a prior distribution of substitution rates. Similar to GP4Rate, we assume that the prior distribution of the log values of site-specific substitution rates follows a Gaussian distribution guided by the protein tertiary structure. The Gaussian prior distribution has a very useful property that the log substitution rates of two physically closely located sites are strongly correlated while the log substitution rates of two distant sites are weakly correlated. Therefore, the Gaussian prior distribution encourages the site-specific log substitution rates to be smoothly distributed over the protein tertiary structure.

To fully define the Bayesian model, we also need to specify the likelihood function. In our previous work of GP4Rate (Huang and Golding, 2014), we used the standard phylogenetic likelihood function (Felsenstein, 1981). However, the phylogenetic likelihood function is too computationally expensive for a web server. Therefore, FuncPatch uses a simpler likelihood function. Firstly, we use the parsimony method

implemented in the PROTPARS program in PHYLIP (Felsenstein, 1989) to estimate the most parsimonious number of substitutions at each site. Then, this number of substitutions is used as a summary statistic in the likelihood function. The likelihood function, which measures the probability of the most parsimonious number of substitutions at a site given its site-specific substitution rate, is assumed to follow a Poisson distribution. The Poisson likelihood function significantly simplifies the computation and, more importantly, makes it easy to design efficient approximation algorithms to infer the posterior distribution of site-specific substitution rates. The combination of the Poisson likelihood function and the Gaussian process priors has been studied in the area of Bayesian disease mapping (Vanhatalo *et al.*, 2010; Vanhatalo and Vehtari, 2007). We developed a customized C++ program to implement this model and then a user-friendly web server based on BioPerl (Stajich *et al.*, 2002) and Jmol (Willighagen and Howard, 2007) was developed to make FuncPatch easily usable to biologists.

### 4.3.2 Poisson Likelihood Function

We assume that a protein alignment, a PDB file of a representative protein tertiary structure, a query PDB chain name, a query sequence name, and an optional phylogenetic tree are provided by the users. The query sequence name should be the exact name of a sequence in the protein alignment and the query sequence should correspond to the query PDB chain. If the users do not provide a phylogenetic tree, FuncPatch generates a neighbor-joining tree automatically (Saitou and Nei, 1987). The phylogenetic tree used in FuncPatch is denoted by $\mathcal{T}$. Then FuncPatch uses MUSCLE (Edgar, 2004) to align the query sequence with the query PDB chain to generate a guide alignment. A set of informative sites are chosen from the original

alignment based on the guide alignment. An informative site must meet three conditions: 1) it must match an amino acid in the query PDB chain; 2) it must contain at least 3 amino acids; 3) the number of gaps at the site must be less than a half of the number of sequences in the original alignment. The informative sites form a new alignment denoted by $\mathbf{X}$ and the number of sites in the alignment $\mathbf{X}$ is denoted by $N$. Then, FuncPatch uses the PROTPARS program in PHYLIP (Felsenstein, 1989) to estimate, $\mathbf{C}$, a vector of the most parsimonious number of substitutions for each site in $\mathbf{X}$, in which a single element $C_i$ is the most parsimonious number of substitutions at site $i$.

FuncPatch assumes that, given $\lambda_i$, the expected number of substitutions at site $i$, $C_i$ follows a Poisson distribution,

$$P(C_i|\lambda_i) = \frac{\lambda_i^{C_i}}{C_i!}e^{-\lambda_i}. \tag{4.1}$$

### 4.3.3 Gaussian Prior Distribution

Given the alignment $\mathbf{X}$, FuncPatch calculates, $\beta$, the average number of substitutions over all sites,

$$\beta = \frac{\sum_{i=1}^{N} C_i}{N}. \tag{4.2}$$

Based on the 3D coordinates of the $\alpha$ carbons of amino acids in the user provided PDB file, FuncPatch then calculates a distance matrix $\mathbf{D}$ in which an element $D_{ij}$ measures the Euclidean distance between site $i$ and site $j$ in the alignment $\mathbf{X}$. Similar to GP4Rate (Huang and Golding, 2014), we assume that the prior distribution of $\mathbf{\Phi}$, the vector of site-specific log substitution rates, follows a zero-mean Gaussian

distribution,

$$P(\mathbf{\Phi}|\mathbf{D}, l, \sigma) = \frac{1}{(2\pi)^{\frac{N}{2}}|\mathbf{\Sigma}(\mathbf{D}, l, \sigma)|^{\frac{1}{2}}} \exp(-\frac{\mathbf{\Phi}^T\mathbf{\Sigma}(\mathbf{D}, l, \sigma)^{-1}\mathbf{\Phi}}{2}), \tag{4.3}$$

in which $\mathbf{\Sigma}(\mathbf{D}, l, \sigma)$ is the covariance matrix. We assume that the covariance matrix $\mathbf{\Sigma}(\mathbf{D}, l, \sigma)$ is parameterized by the Matérn 1.5 covariance function,

$$\Sigma_{ij} = \sigma^2(1 + \frac{\sqrt{3}D_{ij}}{l}) \exp(-\frac{\sqrt{3}D_{ij}}{l}), \tag{4.4}$$

in which $D_{ij}$ is the Euclidean distance between site $i$ and site $j$, $l$ is the characteristic length scale, and $\sigma$ is the signal standard deviation. $l$ is a positive number which measures the strength of the spatial correlation of the site-specific log substitution rates over the protein tertiary structure. A large $l$ implies that the spatial correlation is strong while a small $l$ implies that the spatial correlation is weak. $\sigma$ is a positive number which measures the marginal variation of site-specific log substitution rates at a single site. It is easy to show that, in Equation 4.4, $\Sigma_{ij}$ decreases with increasing $D_{ij}$, which implies that closely located sites are more likely to have similar substitution rates than distantly located sites. Therefore, the Gaussian prior distribution naturally captures our intuition that site-specific substitution rates are smoothly distributed over the tertiary structure. Similar to GP4Rate, FuncPatch introduces a very small amount of noise ("jitter" term) to the diagonal elements in the covariance matrix to ensure that its Cholesky decomposition is numerically stable.

To connect the Poisson likelihood function with the Gaussian prior distribution, we assume that the relationship between $\lambda_i$, i.e. the expected number of substitutions at site $i$, and $\Phi_i$, i.e. the log substitution rate at site $i$, can be described by the

following equation,

$$\lambda_i = \beta \exp(\Phi_i), \tag{4.5}$$

in which $\beta$ is the average number of substitutions calculated by Equation 4.2. In this parameterization, the site-specific substitution rate, i.e. $\exp(\Phi_i)$, is a scaling factor as the substitution rate parameters in statistical phylogenetic models.

By inserting Equation 4.5 into Equation 4.1 and then combining Equation 4.3 and Equation 4.1, we obtain the posterior distribution of $\Phi$,

$$\underbrace{P(\mathbf{\Phi}|l,\sigma,\mathbf{C},\mathbf{D})}_{\text{Posterior}} \sim \underbrace{P(\mathbf{\Phi}|\mathbf{D},l,\sigma)}_{\text{Prior}} \prod_{i=1}^{N} \underbrace{P(C_i|\beta\exp(\Phi_i))}_{\text{Likelihood}}. \tag{4.6}$$

Note that the right-hand side of Equation 4.6 is proportional to the posterior distribution up to a constant $Z$ which is the marginal likelihood of the observed data given the hyperparameters, i.e. $l$ and $\sigma$. The posterior distribution is log concave, so it has only a single stationary point which is also the global maximum. FuncPatch uses the L-BFGS-B algorithm (Zhu *et al.*, 1997) to find the global maximum of the posterior distribution and then uses a Laplace approximation to calculate the approximate posterior distribution of $\mathbf{\Phi}$ and the approximate marginal likelihood $Z$ (Rasmussen and Williams, 2005). The Laplace approximation uses the location of the global maximum in the posterior distribution and the second order derivatives at the maximum to construct a Gaussian distribution to approximate the posterior distribution of $\mathbf{\Phi}$ (Rasmussen and Williams, 2005). For each site $i$ in $\mathbf{X}$, the Laplace approximation can calculate $E_i$ and $S_i$ based on the approximate posterior distribution, where $E_i$ is the approximate posterior mean of $\Phi_i$ and $S_i$ is the approximate posterior standard deviation of $\Phi_i$. We use $\exp(E_i)$ as the estimated substitution rate

at site $i$ and the interval $(\exp(E_i - 0.6745S_i), \exp(E_i + 0.6745S_i))$ as the approximate 50% credible interval at site $i$, in which 0.6745 corresponds to the 25% quantile of the standard Gaussian distribution.

### 4.3.4 Inference of Hyperparameters and Bayesian Model Comparison

The descriptions in the previous sections assume that the two hyperparameters, i.e. $l$ and $\sigma$, are known. In real data analyses, we need to estimate these hyperparameters from data. FuncPatch performs a grid search to generate a point estimation of parameters. We choose 20 representative $l$, evenly spaced between 1 Å and 39 Å, and 20 representative $\sigma$, evenly spaced between 0.1 and 3.9. Therefore, there are 400 different combinations of hyperparameters based on these representative values. Then, FuncPatch performs a Laplace approximation for each combination of hyperparameters to calculate the approximate marginal likelihood $Z$ (Rasmussen and Williams, 2005). The combination of hyperparameters with the largest marginal likelihood $Z$ is chosen as the point estimation of the hyperparameters. The average of the marginal likelihoods over all combinations of hyperparameters is used as the overall marginal likelihood of the model, which implies that we put a uniform hyperprior over hyperparameters.

To evaluate whether the spatial correlation of substitution rates is significant in a dataset, we developed a test based on the Bayesian model comparison. The model described above is the alternative model (model 1) in the Bayesian model comparison. We also designed a null model (model 0) in which any spatial correlation of substitution rates is absent. In model 0, we assume that the characteristic length

scale $l$ is always equal to 0, which essentially removes the spatial correlation from the Gaussian prior distribution. Twenty representative signal standard deviations $\sigma$ are evenly spaced between 0.1 and 3.9 as model 1. The average marginal likelihood over the 20 combinations of hyperparameters is used as the overall marginal likelihood of model 0. We suggest that 8 may be used as a conservative cutoff for the log Bayes factor (model 1 *vs* model 0). If the estimated log Bayes factor is larger than 8 in a dataset, we consider that the spatial correlation of site-specific substitution rates is significant in the dataset.

## 4.4 Simulations and Case Studies

### 4.4.1 Simulations

We evaluated the performance of FuncPatch and compared it with the performances of GP4Rate (Huang and Golding, 2014), Rate4Site (Mayrose *et al.*, 2004), and a customized 3D sliding window program based on the Bio++ library (Dutheil *et al.*, 2006; Gueguen *et al.*, 2013). In the comparison, Rate4Site is used as a representative of methods which ignore the spatial correlation of site-specific substitution rates in protein tertiary structures. The 3D sliding window program is similar to the methods described in previous studies (Dean and Golding, 2000; Suzuki, 2004; Berglund *et al.*, 2005). In the 3D sliding window program, the JTT substitution model (Jones *et al.*, 1992) is used to describe the substitution process of amino acids. We assume that a window size and a reference phylogenetic tree have been provided by the users. For each site in the protein tertiary structure, the 3D sliding window program firstly collects the set of sites whose Euclidean distances to the focal site is smaller than

the window size to generate a local alignment and then optimizes the scale of the reference phylogenetic tree given the local alignment. The tree scale is considered to be the estimated substitution rate at the focal site. We used two window sizes in the 3D sliding window program. The small window size (7 Å) may be more powerful to capture small conserved patches while the large window size (15 Å) may be more powerful to capture large conserved patches. The use of two window sizes also let us explore the robustness of the 3D sliding window method.

We used 4 simulated datasets (A, B, C, and D) described in our previous study of GP4Rate (Huang and Golding, 2014). The spatial correlation of substitution rates is present in dataset A, B, and C while it is absent in dataset D. We briefly describe these simulated alignments in this work and more details can be found in Huang and Golding (2014). All the simulated alignments were based on a simple phylogenetic tree with 4 species and the total branch length is equal to 1. On average there is only 1 substitution at each site, which makes it difficult to infer site-specific substitution rates accurately. Therefore, the 4 sets of simulated alignments are excellent to benchmarking the statistical powers of different methods in the scenario of weak information. The 4 sets of simulated alignments are different in the representative protein tertiary structures and the "true" site-specific substitution rates used in the step of generating simulated alignments.

- The 20 alignments in dataset A were based on a 2D toy protein structure in which amino acid sites corresponded to points in a 20 by 20 2D grid. The Euclidean distance between each site to its closest site is equal to 5 Å in the 2D grid, which is comparable to the average distance between physically interacting amino acids. The 20 by 20 2D grid was divided into 4 blocks each of which was

a 10 by 10 2D grid. Two blocks were assigned to a lower "true" substitution rate, i.e. 0.2, and were considered to be conserved functional patches while the other two blocks were assigned to a higher "true" substitution rate, i.e. 1.8, and were considered to be functionally less important.

- The 20 alignments in dataset B were based on the same 2D toy protein structure used in dataset A. However, the 20 by 20 grid was divided into 16 blocks each of which was a 5 by 5 2D grid. A half of the blocks were assigned to the lower "true" substitution rate, i.e. 0.2, while the others were assigned to the higher "true" substitution rate, i.e. 1.8. The blocks with the lower substitution rate were considered to be conserved functional patches.

- The 20 alignments in dataset C were based on the tertiary structure of the human Bcl-xL protein (PDB ID:1MAZ; Muchmore *et al.*, 1996) instead of the 2D toy protein structure. In addition, the "true" substitution rates used in the step of generating alignments were downloaded from the ConSurf-DB database (Goldenberg *et al.*, 2009) which automatically collected a large number of homologs of the human Bcl-xL protein and then used Rate4Site to estimate site-specific substitution rates. The 10% of sites with the lowest "true" substitution rates were considered to be functional.

- The 20 alignments in dataset A were permuted randomly to generate the alignments in dataset D. The permutations destroyed the spatial correlation of substitution rates but kept all other features of the alignments. The half of sites with lower "true" substitution rates were still considered as the functional sites in dataset D. This dataset was designed to test the performance of different

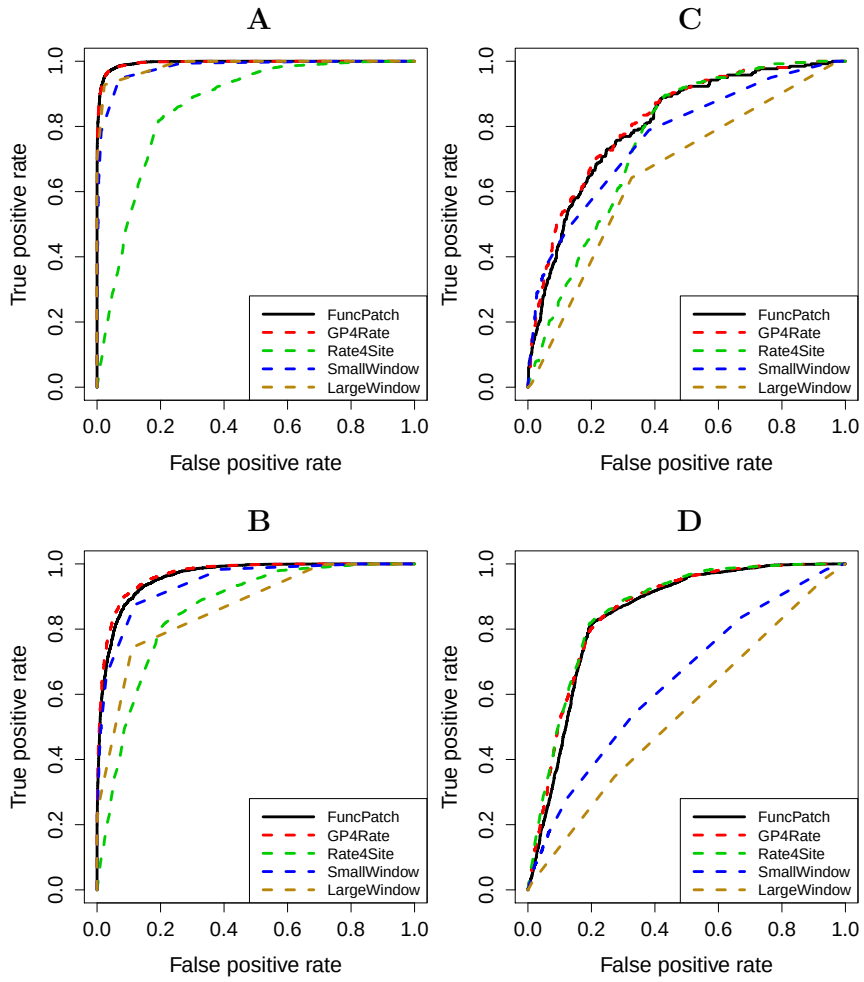


Figure 4.1: The performances of different methods in the simulation study. A: the performances of different methods in the simulated dataset A; B: the performances of different methods in the simulated dataset B; C: the performances of different methods in the simulated dataset C; D: the performances of different methods in the simulated dataset D. Solid black lines: the ROC curves of FuncPatch; dashed red lines: the ROC curves of GP4Rate; dashed green lines: the ROC curves of Rate4Site; dashed blue lines: the ROC curves of the 3D sliding window method with the small window size (7 Å); dashed brown line: the ROC curves of the 3D sliding window method with the large window size (15 Å).

methods when a spatial correlation of substitution rates is absent.

We applied FuncPatch, GP4Rate, Rate4Site, and the customized 3D sliding window program to the 4 sets of simulated alignments. The "true" phylogeny with branch lengths were used as the reference tree in all analyses. For GP4Rate, two independent MCMC chains were implemented for each alignment and the first 30% of samples were discarded as burn-in. We used the ROCR package in R (Sing *et al.*, 2005) to plot the receiver operating characteristic (ROC) curves to evaluate the statistical powers of the 4 programs for identifying conserved functional sites as defined above. As shown in Figure 4.1, FuncPatch and GP4Rate have very similar powers in all the 4 simulated datasets. More importantly, FuncPatch and GP4Rate are always among the most powerful methods. When the spatial correlation of substitution rates is present (Figure 4.1A, 4.1B, and 4.1C), the 3D sliding window method can be more powerful than Rate4Site but FuncPatch and GP4Rate always outperform the 3D sliding window method. In addition, the performance of the 3D sliding window method is sensitive to the choice of the window size. For example, in dataset B and C, using the small window (7 Å) is better than using the large window size (15 Å). In contrast, in dataset A, the difference of the performances between the two window sizes is marginal. When the spatial correlation of substitution rates is absent (Figure 4.1D), the performance of Rate4Site is very similar to the performances of FuncPatch and GP4Rate. In contrast, the performance of the 3D sliding window method is much worse than these three methods in this scenario.

The simulations demonstrate that FuncPatch is a good approximation to GP4Rate and that FuncPatch is always among the most powerful methods regardless of the assumptions of the simulated alignments. In contrast, the performances of Rate4Site and the 3D sliding window method depend on the assumptions of the simulated

alignments and the performance of the 3D sliding window method is sensitive to the window size. In summary, FuncPatch is a robust method and generally outperforms Rate4Site and the 3D sliding window method.

### 4.4.2 Case Study of MAPK1 Genes

To demonstrate the power of FuncPatch in analyses with real data, we applied FuncPatch to two signal transduction related gene families to predict conserved functional patches. As shown in the previous section of simulations, the performance of the 3D sliding window method is sensitive to the choice of the window size and it can lead to misleading results when the spatial correlation of substitution rates is absent. Therefore, we only compared the performances of FuncPatch, GP4Rate, and Rate4Site while the 3D sliding method is not included in the comparison. In this section, we report the results of the MAPK1 (ERK2) gene family which is a central player in signal transduction. The MAPK1 gene family is a member of the mitogen-activated protein kinase (MAPK) superfamily. In the activation of the MAPK/ERK pathway, cell surface receptors are first activated by extra-cellular ligands and then the cell surface receptors activate a series of factors, including MAPKs (Seger and Krebs, 1995). MAPKs thus activate a variety of downstream transcriptional factors.

We downloaded the protein sequences of 17 MAPK1 orthologous genes from the NCBI HomoloGene database (HomoloGene ID: 37670; Sayers *et al.*, 2012). We only included the MAPK1 subfamily in the analysis and excluded other MAPK subfamilies, because the biological functions of the MAPK1 subfamily may be different from the functions of other MAPK subfamilies and the locations of conserved functional patches might be different in different subfamilies. Therefore, the level of sequence

divergence is relatively low in this dataset, which makes it difficult to accurately infer site-specific substitution rates in individual sites. The 17 MAPK1 protein sequences were aligned using MUSCLE with default parameters. The phylogenetic tree was inferred using PhyML with the JTT+$\Gamma$ model (Guindon and Gascuel, 2003). We used the X-ray crystallographic structure of the rat MAPK1 gene (PDB ID: 1ERK; Zhang *et al.*, 1994) as the representative structure.

We applied FuncPatch with default parameters to the MAPK1 dataset. As shown in Table 4.1, the best estimation of characteristic length scale is equal to 21 Å, which is much larger than 0. It suggests that spatial correlation of substitution rates is extended over a very long distance. The statistical significance of the spatial correlation of substitution rates is supported by the Bayesian model comparison. The approximate log Bayes factor (model 1 *vs* model 0) is equal to 148.7 which is significantly larger than the cutoff 8 (Table 4.1). To furthermore demonstrate that the cutoff 8 is valid, we randomly permuted the MAPK1 alignment 1000 times to generate a set of new alignments. The potential spatial correlation of substitution rates has been destroyed in these permuted alignments and we applied FuncPatch to them to generate a null distribution of the log Bayes factors. As shown in Figure 4.2A, only 2.6% of permuted alignments' log Bayes factors are greater than 8, which confirms that the cutoff 8 is conservative. Therefore, the spatial correlation of substitution rates is statistically significant in the MAPK1 dataset via both the Bayes factor and the permutations.

We superimposed the 35 most conserved sites predicted by FuncPatch onto the protein structure of the rat MAPK1 gene (PDB ID: 1ERK). Because the MAPK1 dataset consists of 357 sites, the 35 sites correspond to the 10% of most conserved

Table 4.1: Estimation of parameters and log Bayes factors in the two case studies of the MAPK1 genes and the SMAD genes.

| Dataset | $l$ | $\sigma$ | Log Bayes factor |
|---|---|---|---|
| MAPK1 | 21 | 1.3 | 148.7 |
| SMAD | 9 | 1.1 | 9.19 |

sites in MAPK1. As shown in Figure 4.3A, the 35 most conserved sites form a clearly bounded patch in the protein tertiary structure. The result is not surprising because the estimated characteristic length scale is very large. We further investigated whether this predicted conserved patch is related to MAPK1's activities. Interestingly, previous studies suggest that Asp-147, the second most conserved site predicted by FuncPatch, acts as the catalytic base (Zhang *et al.*, 1994; Canagarajah *et al.*, 1997; Turjanski *et al.*, 2009). Therefore, the predicted conserved patch corresponds to the catalytic site of MAPK1.

We also applied GP4Rate and Rate4Site to the MAPK1 dataset and mapped the 35 most conserved sites onto the protein tertiary structure of the rat MAPK1 gene (PDB ID: 1ERK). As shown in Figure 4.3A, GP4Rate reported essentially the same conserved patch as the one reported by FuncPatch. The consistency between FuncPatch and GP4Rate is not surprising, because the simulation study has already demonstrated that FuncPatch and GP4Rate have similar powers to identify conserved functional patches. However, FuncPatch took only about 1 CPU minute to analyze the MAPK1 dataset while GP4Rate took about 33 CPU hours to analyze the same dataset. Therefore, FuncPatch is orders of magnitudes faster than GP4Rate in the MAPK1 dataset.

In contrast, the most conserved sites predicted by Rate4Site are very different

from the conserved sites predicted by FuncPatch and GP4Rate (Figure 4.3A). Indeed, the most conserved sites predicted by Rate4Site are scattered in the protein tertiary structure and do not form any clearly bounded 3D patch. In addition, Asp-147, i.e. the catalytic base, is not included in the 35 most conserved sites predicted by Rate4Site, even though it is invariant across all sequences in the MAPK1 alignment. The reason that Asp-147 is not included in the set of most conserved sites predicted by Rate4Site is that there are a number of invariant sites in the MAPK1 alignment. Therefore, it is difficult to know which one is more conserved than the others based on the information at individual sites. In contrast, FuncPatch models the spatial correlation of substitution rates and the invariant sites in the core regions of the conserved patches tend to have lower estimated substitution rates than other invariant sites scattered in the protein tertiary structure.

### 4.4.3 Case Study of SMAD Genes

SMAD genes are important factors which mediate the transduction of signals from extracellular ligands to downstream factors (Attisano and Tuen Lee-Hoeflich, 2001). We downloaded the protein sequences of 11 SMAD1 genes (HomoloGene ID: 21196), 9 SMAD5 genes (HomoloGene ID: 4313), and 10 SMAD8 genes (HomoloGene ID: 21198) from the NCBI HomoloGene database (Sayers *et al.*, 2012). All these sequences are receptor-regulated SMAD (R-SMAD) genes regulated by bone morphogenetic proteins (BMPs) (Attisano and Tuen Lee-Hoeflich, 2001). MUSCLE (Edgar, 2004) with default parameters was used to generate an alignment based on the 30 SMAD proteins and then PhyML with the JTT+$\Gamma$ model (Guindon and Gascuel, 2003) was used to generate the reference phylogenetic tree. We did not include other SMAD

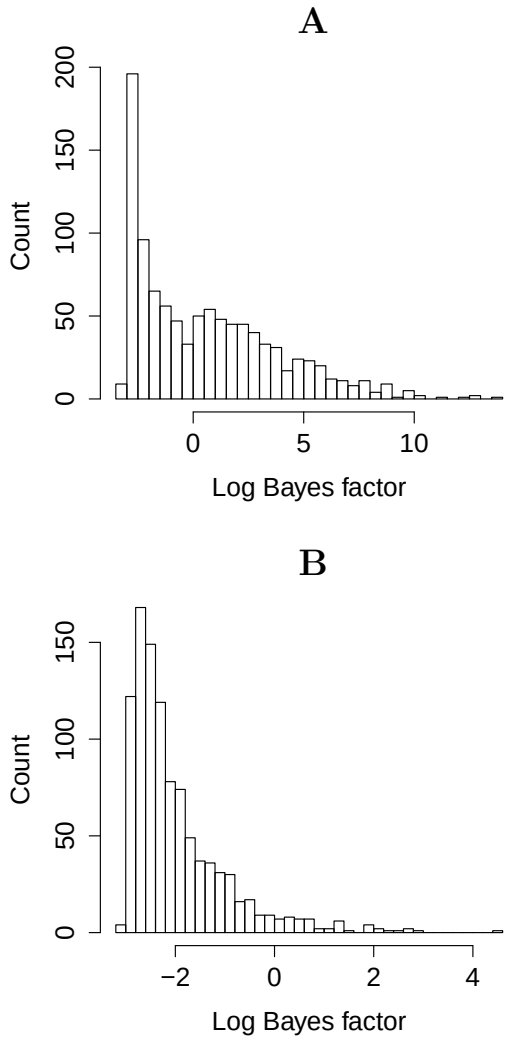


Figure 4.2: The null distributions of approximate log Bayes factors in the two case studies of the MAPK1 genes and the SMAD genes. The The null distributions are generated by applying FuncPatch to the permuted alignments. A: the null distribution of the approximate log Bayes factors in the case study of the MAPK1 genes. B: the null distribution of the approximate log Bayes factors in the case study of the SMAD genes.

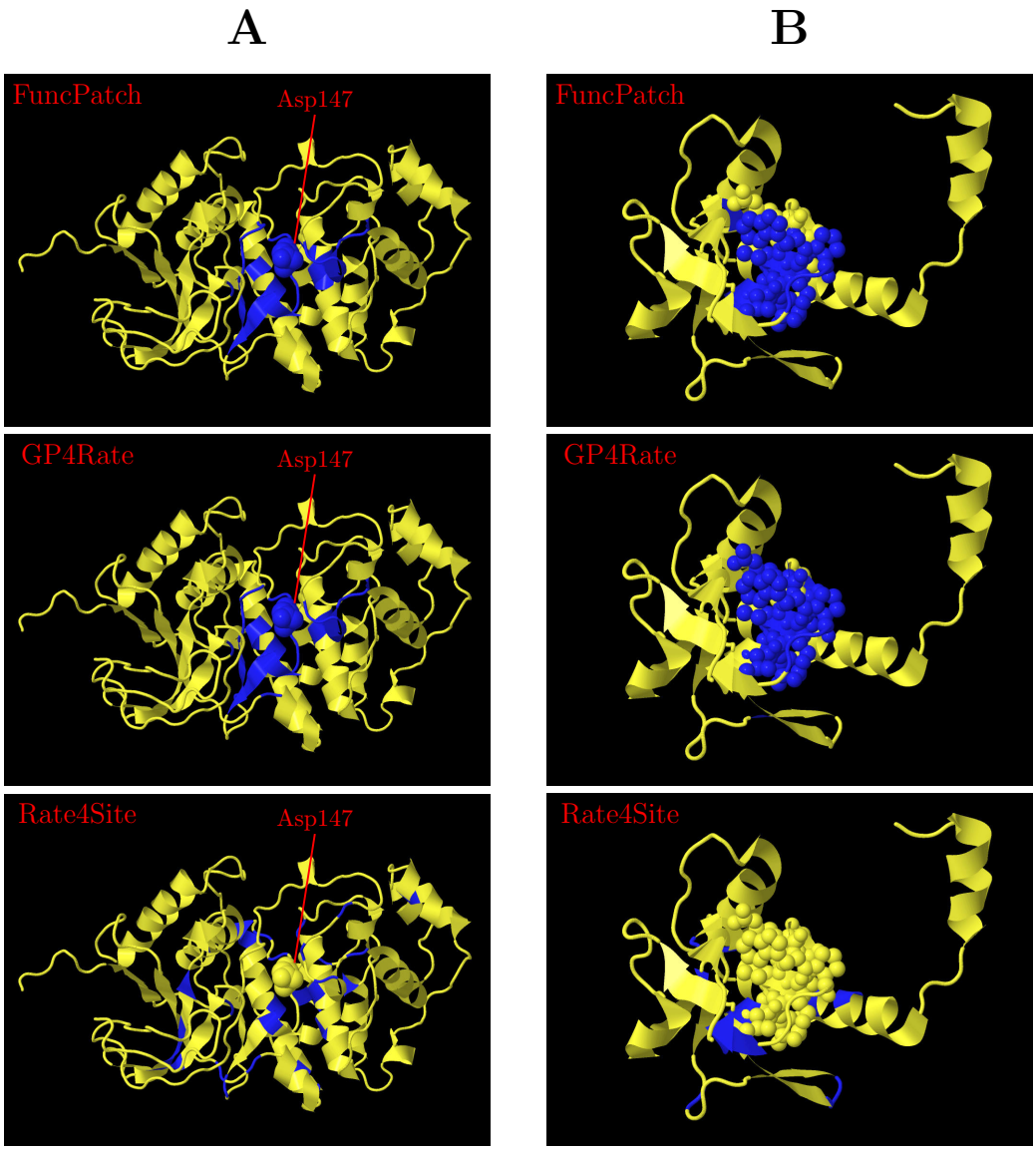


Figure 4.3: The 3D locations of the most conserved sites in the two case studies of the MAPK1 genes and the SMAD genes. A: the 35 most conserved sites predicted by FuncPatch, GP4Rate, and Rate4Site in the MAPK1 dataset (PDB ID: 1ERK). The blue sites are the predicted conserved sites while the yellow sites are not conserved. The space-filled atoms belong to Asp-147, the catalytic base. B: the 12 most conserved sites predicted by FuncPatch, GP4Rate, and Rate4Site in the SMAD dataset (PDB ID: 3KMP). The blue sites are the predicted conserved sites while the yellow sites are not conserved. The space-filled atoms belong to an experimentally verified binding site (Scherer and Graff, 2000). The protein structures are visualized using Jmol (Willighagen and Howard, 2007).

genes in the NCBI HomoloGene database (Sayers *et al.*, 2012), because these SMAD genes are either not R-SMAD genes or not regulated by BMPs and may have different functions from the BMP-regulated R-SMAD genes (Attisano and Tuen Lee-Hoeflich, 2001). The R-SMAD proteins consist of two domains, i.e. MAD homology 1 (MH1) and MAD homology 2 (MH2), connected by a disordered peptide (Attisano and Tuen Lee-Hoeflich, 2001). We used the X-ray crystallographic structure of the MH1 domain in the mouse SMAD1 protein (PDB ID: 3KMP; Baburajendran *et al.*, 2010) as the representative protein tertiary structure.

We applied FuncPatch to the SMAD dataset to infer conserved functional patches in the SMAD MH1 domain. FuncPatch analyzed 124 sites in the SMAD alignment. As shown in Table 4.1, the estimated characteristic length scale is equal to 9 Å in the SMAD dataset. Because the estimated characteristic length scale is larger than 0, a spatial correlation of substitution rates may be present in the SMAD dataset. Indeed, the approximate log Bayes factor reported by FuncPatch is greater than the cutoff 8 (Table 4.1), which suggests that the spatial correlation of substitution rates is statistically significant in the SMAD dataset. To test whether the cutoff 8 is valid in the SMAD dataset, we again generated 1000 permuted alignments based on the SMAD alignment and then applied FuncPatch to these permuted alignments to generate a null distribution of the approximate log Bayes factors. As shown in Figure 4.2, none of the 1000 permuted alignments has a log Bayes factor larger than 8. Therefore, the cutoff 8 is conservative in the SMAD dataset.

We superimposed the 12 most conserved sites predicted by FuncPatch onto the tertiary structure of the MH1 domain (Figure 4.3B), which correspond to the 10% of most conserved sites in the 124 analyzed sites. Obviously the most conserved

sites predicted by FuncPatch are physically close to each other in the protein tertiary structure. In addition, most of these conserved sites overlap with the $\alpha$-helix 4 of the MH1 domain. A previous experimental study has already demonstrated that this region is a binding site which may interact with calmodulin and may contribute to the crosstalk between the calmodulin pathway and the SMAD pathway (Scherer and Graff, 2000). Therefore, the conserved region predicted by FuncPatch is supported by experimental evidence.

We also applied GP4Rate and Rate4Site to the SMAD dataset. As shown in Figure 4.3B, the 12 most conserved sites predicted by GP4Rate largely overlap with the 12 most conserved sites predicted by FuncPatch. Therefore, FuncPatch is a good approximation to GP4Rate. However, FuncPatch took only about 6 CPU seconds to analyze the SMAD dataset while GP4Rate took about 4 CPU hours. Again, FuncPatch is orders of magnitudes faster than GP4Rate. In contrast, the 12 most conserved sites predicted by Rate4Site are scattered in the protein tertiary structure and do not overlap with the sites predicted by FuncPatch (Figure 4.3B). Therefore, FuncPatch identified an experimentally verified conserved region in the SMAD dataset which is overlooked by Rate4Site.

## 4.5 Discussion

Recently, we developed GP4Rate, a phylogenetic Gaussian process model which combines statistical phylogenetics and Gaussian processes to infer conserved functional regions in protein tertiary structures (Huang and Golding, 2014). Our previous study has already shown than GP4Rate is a powerful method to infer conserved functional regions and outperforms Rate4Site, but GP4Rate is a slow MCMC program. In this

work, we present a new statistical method, FuncPatch, which is designed as a fast approximation to GP4Rate. While it takes from hours to days for GP4Rate to analyze a protein family, FuncPatch can finish a similar analysis within a few minutes. An intuitive web based graphical interface of FuncPatch is available and makes it more accessible to experimental biologists. Both simulations and the case studies of the MAPK1 genes and the SMAD genes suggest that FuncPatch is a very accurate approximation to GP4Rate. The simulations also show that FuncPatch is more powerful and robust than Rate4Site and the 3D sliding window method. The conserved patches predicted by FuncPatch in the two case studies are supported by experimental evidence. Therefore, we believe FuncPatch is a useful tool for analyzing protein families and guiding mutagenesis experiments.

Unlike many other alternative methods, e.g. the 3D sliding window method, FuncPatch uses a Gaussian prior distribution which naturally captures the spatial correlation of substitution rates in protein tertiary structures. Therefore, it can infer the strength of the spatial correlation of substitution rates. In addition, a Bayesian comparison method has been implemented in FuncPatch to test whether the spatial correlation of substitution rates is significant in a dataset. The two case studies of the MAPK1 genes and the SMAD genes suggest that the strength of the spatial correlation may vary in different protein families and our preliminary analyses on a few other protein families also suggest that the spatial correlation of substitution rates may be insignificant in some protein families (data not shown). We believe that the ability of inferring the strength and significance of the spatial correlation of substitution rates is a significant advantage of FuncPatch over the 3D sliding window method.

Our case studies in this work focus on two empirical datasets in which the divergence levels of sequences are relatively low. However, it is by no means the case that FuncPatch cannot be used to analyze datasets in which the sequence divergence level is high. Indeed, it is always helpful to include more sequences in the analyses, if it is believed that the sequences share the same conserved patches. However, functional divergence may happen after gene duplication (Gu, 1999; Knudsen and Miyamoto, 2001; Huang and Golding, 2012), which makes it risky to include remote homologs with different functions in the analyses. FuncPatch may alleviate the problem, because the remote homologs may be removed from data without significantly reducing the statistical power of the analyses. Therefore, FuncPatch is particularly useful for analyzing small gene families or gene families that have undergone recent gene duplications.

## 4.6 Acknowledgements

We thank Ben Evans for insightful comments on this work.

# Chapter 5

# Conclusion

In this thesis, we report new statistical phylogenetic models for predicting functionally important regions in protein tertiary/primary structures. These models are based on the intuitive assumptions that functionally important protein regions may be under strong purifying selection in the whole gene family or under different selection pressures in different subfamilies. A large number of classic models have been developed to infer functionally important sites in proteins based on these assumptions. However, most of these classic methods assume that evolutionary patterns are independent and identically distributed (i.i.d.) over sites and focus on the inference of functionally important sites instead of regions. As far as we know, the underrepresentation of rigorous statistical models for inferring functionally important protein regions is mainly due to the difficulty of modeling the spatial correlation of evolutionary patterns in protein tertiary/primary structures. In this thesis, Gaussian processes and hidden Markov models have been used as priors to model the spatial correlation of evolutionary patterns in protein tertiary structures and protein primary structures,

respectively, and then these priors are combined with Felsenstein's phylogenetic likelihood function (Felsenstein, 1981) to infer functionally important protein regions.

To demonstrate that these new models are powerful and robust, we performed systematic simulation studies to evaluate the performs of these new models and the classic models based on the i.i.d. assumption. The simulation studies suggest that, in general, these new models which explicitly model the spatial correlation of evolutionary patterns in protein tertiary/primary structures significantly outperform the classic models based on the i.i.d. assumption when the spatial correlation of evolutionary patterns is present. Furthermore, when the spatial correlation of evolutionary patterns is absent, these new models still have similar powers as the classic models. Therefore, the simulation studies suggest these new models are potentially more powerful than the classic models and are very robust.

A number of case studies have also been performed to compare the results from the new models with those from the classic models. The case studies clearly show that the new models which explicitly model the spatial correlation of evolutionary patterns in protein tertiary/primary structures can lead to results which are very different from the results of the classic models. More importantly, the functionally important regions predicted by these new models are supported by experimental evidence. Therefore, the new statistical models developed in this thesis can provide new insights on the functionally important protein regions which cannot be detected by the classic models.

The new methodologies described in this thesis can also be used to study other important questions in phylogenetics and molecular evolution. The phylogenetic Gaussian process framework is particularly interesting, since it is one of a few rigorous

statistical methods which can capture the spatial correlation of evolutionary patterns in protein tertiary structures. For example, the framework of phylogenetic Gaussian process models may be used to infer protein 3D patches under positive selection. To develop such a model, we may firstly use a Gaussian process to define a prior distribution of $dN/dS$ ratios over a protein tertiary structure. Then, the prior distribution may be combined with a codon model (Goldman and Yang, 1994; Yang *et al.*, 2000) to infer protein 3D patches under positive selection. It is interesting to develop such a phylogenetic Gaussian process model and then compare it with a recently developed codon model in which the Ising model is used to to capture the spatial correlation of $dN/dS$ ratios in protein tertiary structures (Watabe and Kishino, 2013).

In summary, in this thesis we describe a number of new statistical phylogenetic models to infer functionally important regions in proteins. The usefulness and robustness of these models have been demonstrated by simulations and case studies. In addition, the new methodologies developed in these models, e.g. the phylogenetic Gaussian process framework, may open new avenues to develop biologically realistic models for studying important questions in phylogenetics and molecular evolution.

# Appendix A

# Supplementary Material for Chapter 2

## A.1 Proof of the Stationary Distribution

As shown in the main text, the transition probabilities between two states can be parameterized by the following three equations:

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

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

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

In this section, we prove that if the Markov model of $k^2$ states are parameterized by Equation (A.1), (A.2), and (A.3) , its stationary distribution follows

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

It's trivial to show that $\pi(r_{ij}) \geq 0$ for every $r_{ij}$ and $\sum_{0\leq i,j<k} \pi(r_{ij}) = 1$. Therefore, we just need to prove that

$$\pi(r_{i'j'}) = \sum_{0\leq i,j<k} \pi(r_{ij}) P(r_{i'j'}|r_{ij}). \tag{A.5}$$

Due to the symmetry of the one-step transition matrix, we prove the above theory in two cases. In the first case $r_{i'j'} \in M_0$, and in the second case $r_{i'j'} \in M_1 \bigcup M_2$.

If $r_{i'j'} \in M_0$,

$$\begin{aligned}
\sum_{0\le i,j<k} \pi(r_{ij})P(r_{i'j'}|r_{ij}) = & \sum_{0\le i=j<k} \pi(r_{ij})P(r_{i'j'}|r_{ij}) \\
& + \sum_{0\le i\neq j<k} \pi(r_{ij})P(r_{i'j'}|r_{ij}) \\
= & \{\frac{p_1}{(2p_0+p_1)k}\cdot(1-2p_0)\cdot(\lambda_0+\frac{1-\lambda_0}{k}) \\
& + (k-1)\cdot\frac{p_1}{(2p_0+p_1)k}\cdot(1-2p_0)\cdot\frac{1-\lambda_0}{k}\} \\
& + k(k-1)\cdot\frac{2p_0}{(2p_0+p_1)(k-1)k}\cdot\frac{p_1}{k} \\
= & \frac{p_1}{(2p_0+p_1)k} \\
= & \pi(r_{i'j'}).
\end{aligned} \tag{A.6}$$

Therefore, the stationary probabilities of states in $M_0$ follow Equation (A.4).

Similarly, if $r_{i'j'} \in M_1 \cup M_2$,

$$\begin{aligned}
\sum_{0\le i,j<k} \pi(r_{ij})P(r_{i'j'}|r_{ij}) = & \sum_{0\le i=j<k} \pi(r_{ij})P(r_{i'j'}|r_{ij}) \\
& + \sum_{0\le i\neq j<k} \pi(r_{ij})P(r_{i'j'}|r_{ij}) \\
= & \; k \cdot \frac{p_1}{(2p_0+p_1)k} \cdot \frac{2p_0}{k(k-1)} \\
& + \{\frac{2p_0}{(2p_0+p_1)(k-1)k} \cdot (1-p_1) \cdot (\lambda_1 + \frac{2(1-\lambda_1)}{k(k-1)}) \\
& + (\frac{k(k-1)}{2} - 1) \cdot \frac{2p_0}{(2p_0+p_1)(k-1)k} \cdot (1-p_1) \cdot \frac{2(1-\lambda_1)}{k(k-1)}\} \\
= & \; \frac{2p_0}{(2p_0+p_1)(k-1)k} \\
= & \; \pi(r_{i'j'}).
\end{aligned} \tag{A.7}$$

Therefore, the stationary probabilities of states in $M_1$ and $M_2$ follow Equation (A.4).

In conclusion, the one-step transition probabilities defined in Equation (A.1), (A.2), and (A.3) imply stationary probabilities defined in Equation (A.4).

## A.2 Supplementary Figures

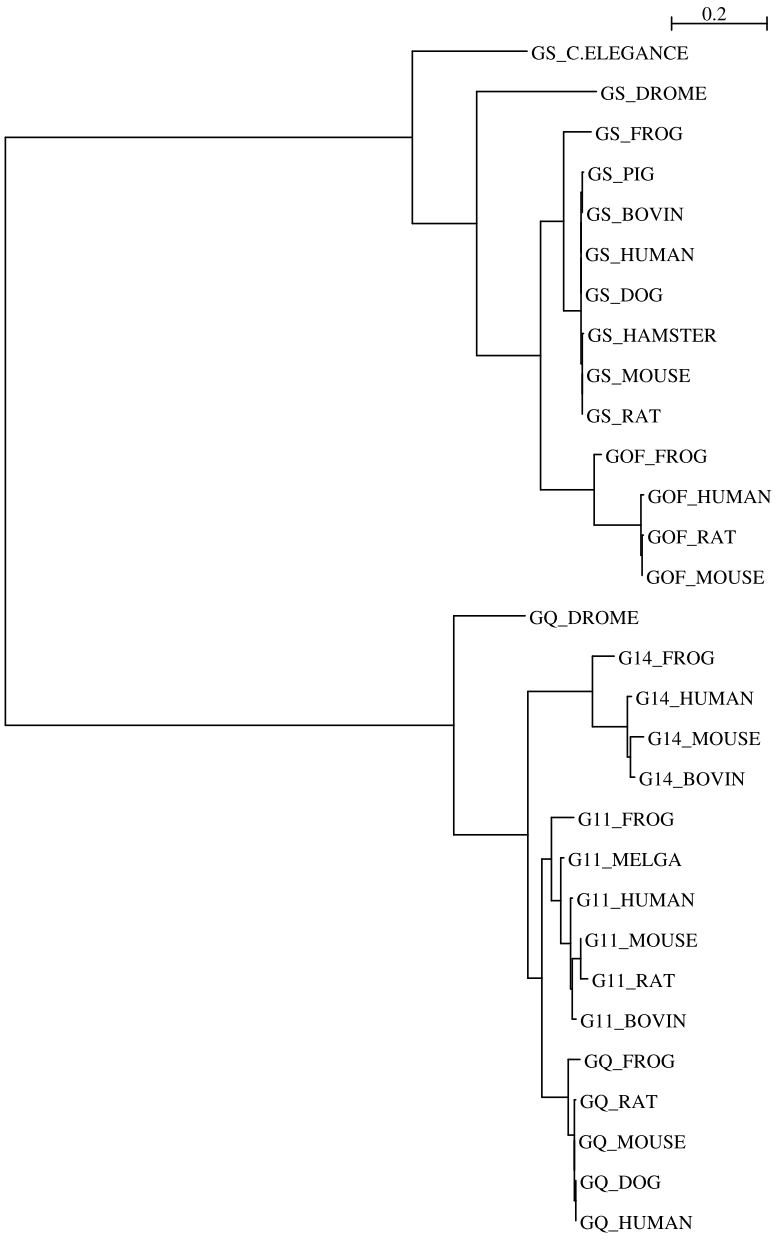


Figure A.1: The phylogenetic tree of G protein $\alpha$ subunits in animals.

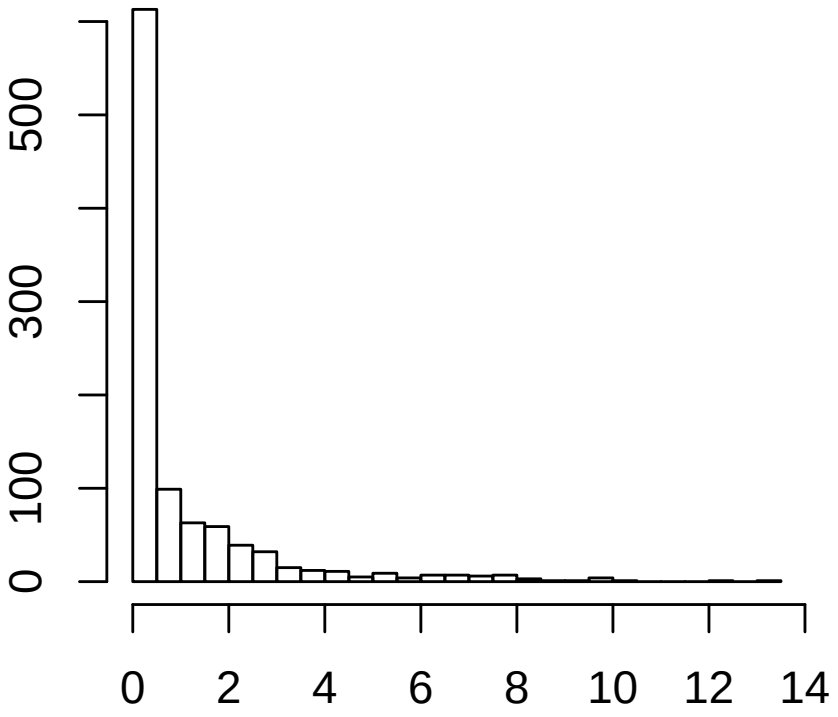


Figure A.2: The distribution of log likelihood ratios in parametric bootstrap. The parameters in the null model described in table 1 in the main text were used to generate 1000 simulated alignments. The X axis represents estimated log likelihood ratios in simulated alignments, while the Y axis represents the number of cases in each bin.

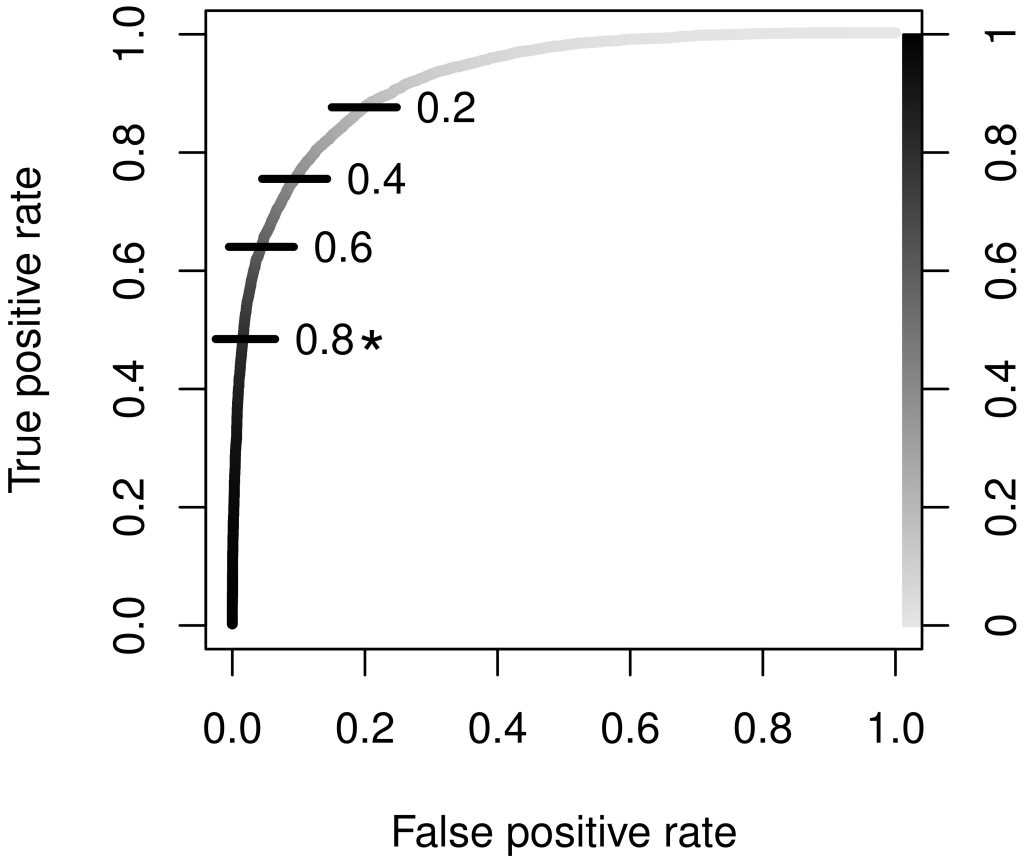


Figure A.3: The performance of HMMDiverge in the set of 50 simulated alignments. The alignments are simulated using parameters in the alternative (full) model in table 1 in the main text. The right-side chart represents the cutoffs used in the identification of sites relevant to typ-I functional divergence. Representative cutoffs are also labeled on the ROC curve directly. $\star$: the cutoff used in the case study of G protein $\alpha$ subunits.

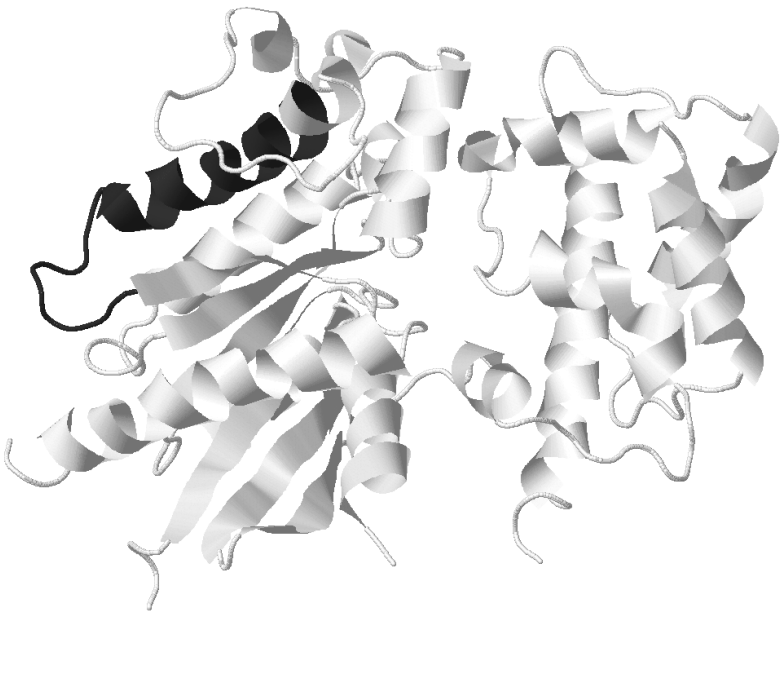


Figure A.4: The spatial distribution of the second region under type-I functional divergence in the 3D protein structure of G protein $\alpha$ subunit. The G protein $\alpha$ subunit consists of two domains, a GTPase domain (left) and a helical domain (right). The dark region is the second candidate region under type-I functional divergence, which overlaps with the $\alpha$-4 helix and the $\alpha$4-$\beta$6 loop in the GTPase domain. The protein structure is visualized using Jalview (Clamp *et al.*, 2004).

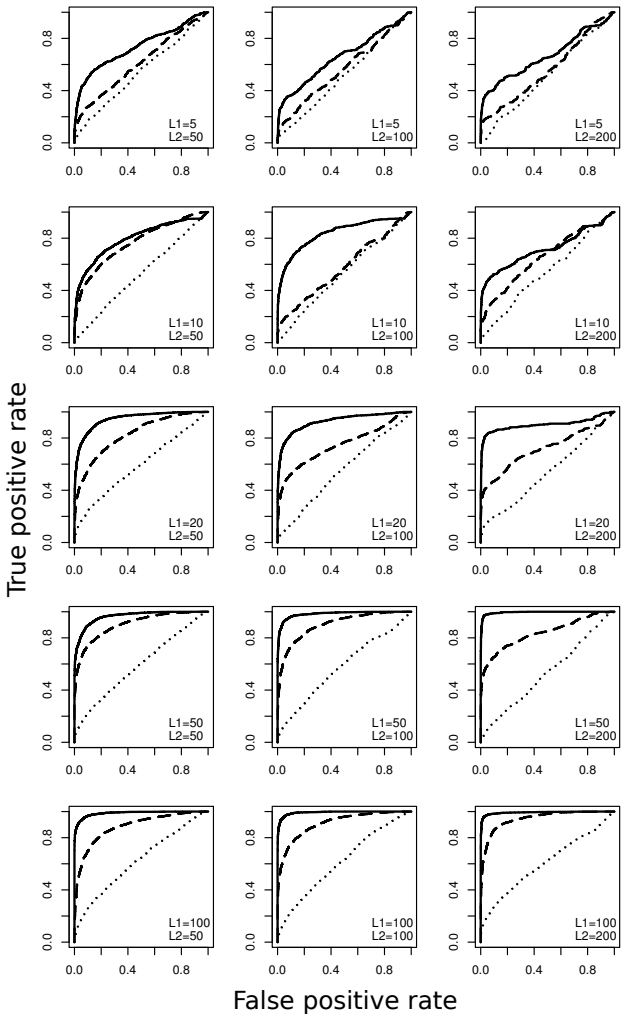


Figure A.5: The performance of HMMDiverge in the first set of additional simulations. X axes represent false positive rates while Y axes represent true positive rates. Each row consists of the ROC curves of multiple simulations having the equal length of divergence relevant regions (L1), which are 5 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 $\alpha$ is 0.2 in all simulations.

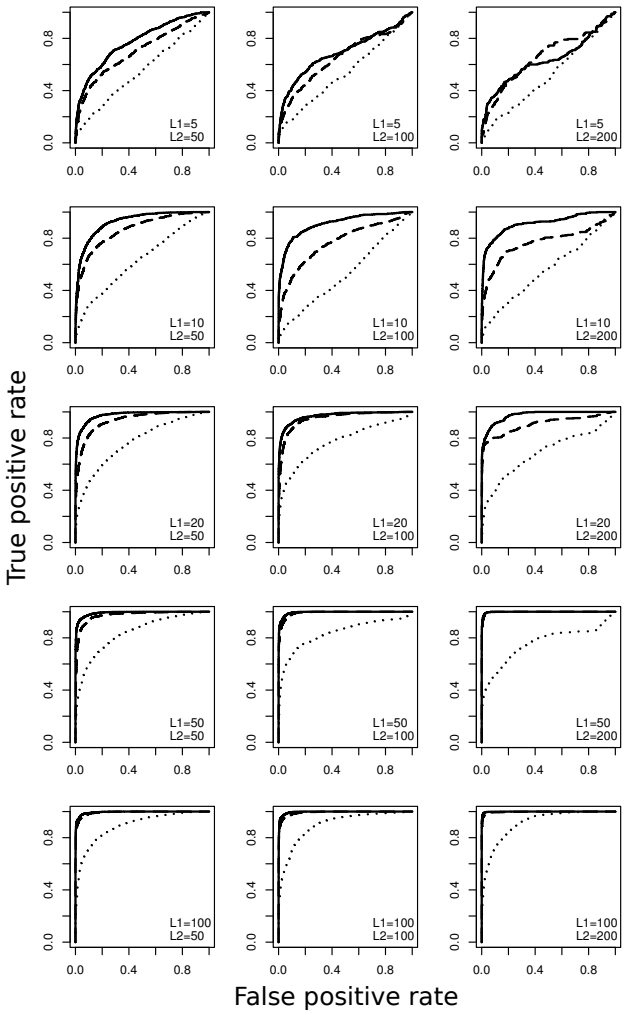


Figure A.6: The performance of HMMDiverge in the second set of additional simulations. X axes represent false positive rates while Y axes represent true positive rates. Each row consists of the ROC curves of multiple simulations having the equal length of divergence relevant regions (L1), which are 5 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 $\alpha$ is 1.0 in all simulations.

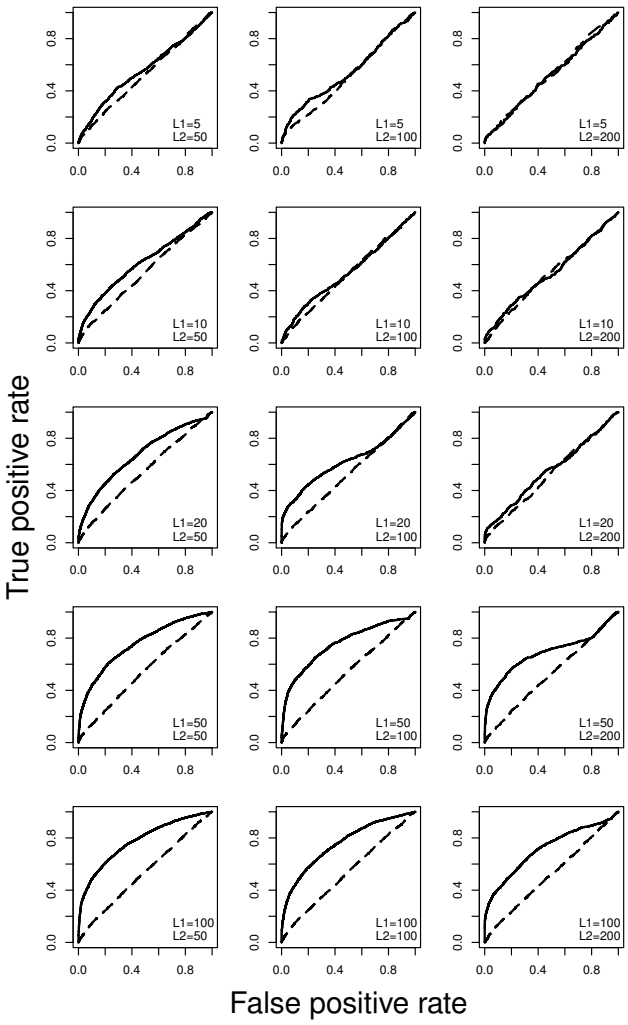


Figure A.7: The comparison of HMMDiverge and DIVERGE2 in the reference simulations. X axes represent false positive rates while Y axes represent true positive rates. Each row consists of the ROC curves of multiple simulations having the equal length of divergence relevant regions (L1), which are 5 aa, 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. The shape parameter $\alpha$ is 0.5 in all simulations. The scale factor is 1.5 in the rapidly evolved subfamily and 0.5 in the slowly evolved subfamily in all simulations. Solid curves: the performances of HMMDiverge. Dashed curves: the performances of DIVERGE2.

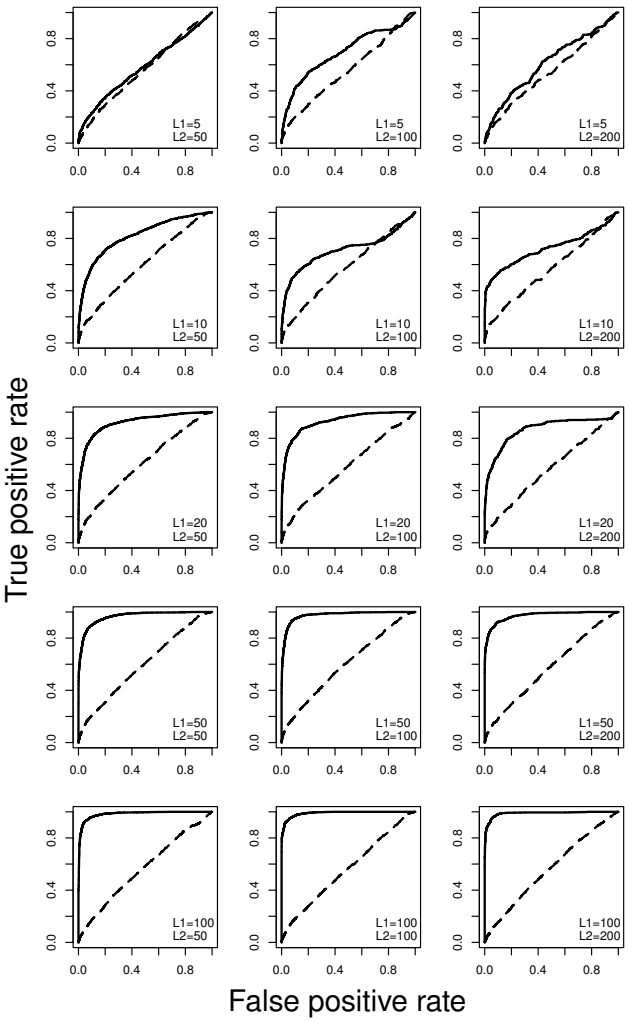


Figure A.8: The comparison of HMMDiverge and DIVERGE2 in the reference simulations. X axes represent false positive rates while Y axes represent true positive rates. Each row consists of the ROC curves of multiple simulations having the equal length of divergence relevant regions (L1), which are 5 aa, 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. The shape parameter $\alpha$ is 0.5 in all simulations. The scale factor is 1.75 in the rapidly evolved subfamily and 0.25 in the slowly evolved subfamily in all simulations. Solid curves: the performances of HMMDiverge. Dashed curves: the performances of DIVERGE2.

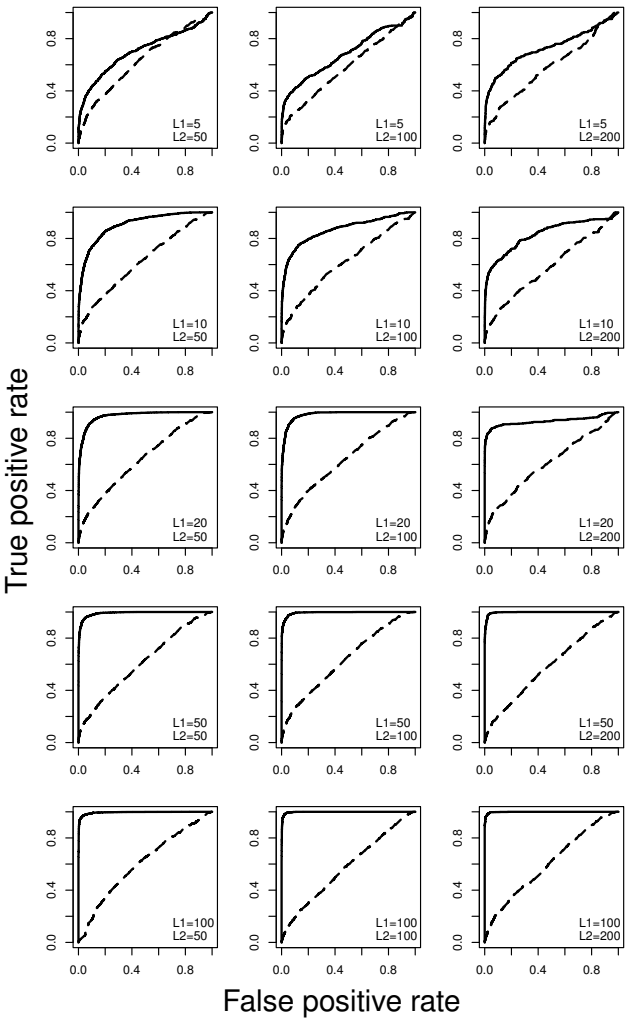


Figure A.9: The comparison of HMMDiverge and DIVERGE2 in the reference simulations. X axes represent false positive rates while Y axes represent true positive rates. Each row consists of the ROC curves of multiple simulations having the equal length of divergence relevant regions (L1), which are 5 aa, 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. The shape parameter $\alpha$ is 0.5 in all simulations. The scale factor is 1.875 in the rapidly evolved subfamily and 0.125 in the slowly evolved subfamily in all simulations. Solid curves: the performances of HMMDiverge. Dashed curves: the performances of DIVERGE2.

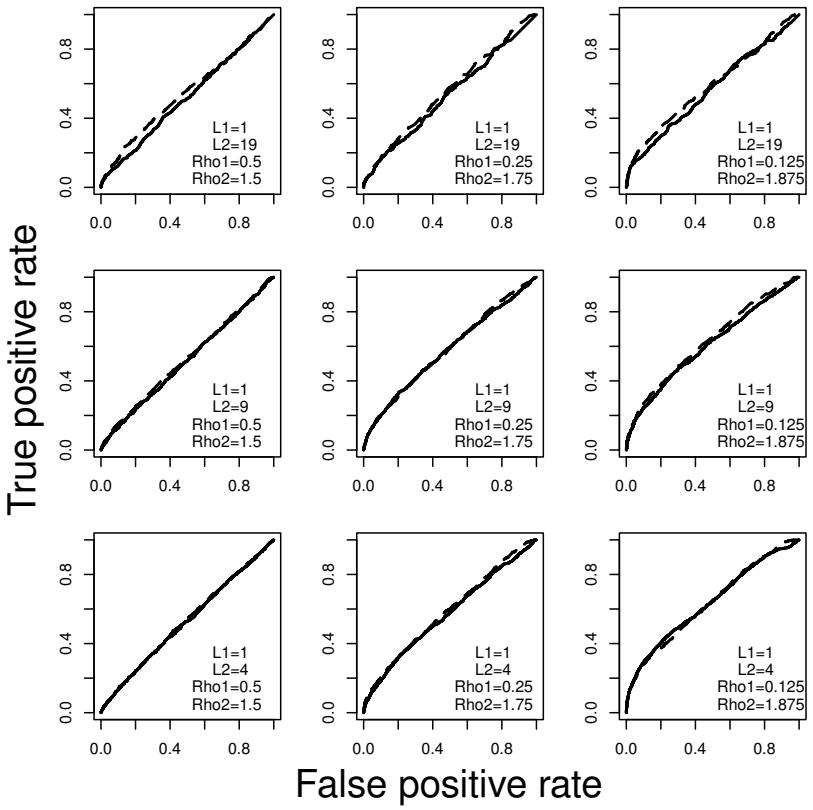


Figure A.10: The comparison of HMMDiverge and DIVERGE2 in the third set of additional simulations. X axes represent false positive rates while Y axes represent true positive rates. The length of 'functional divergence relevant regions', L1, is fixed to 1, which implies individual sites are units of functional divergence. L2: the length of 'functional divergence irrelevant regions'. Rho1: the branch length scale factor in the slowly evolved subfamily. Rho2: the branch length scale factor in the rapidly evolved subfamily. The shape parameter, $\alpha$, is set to 0.5.

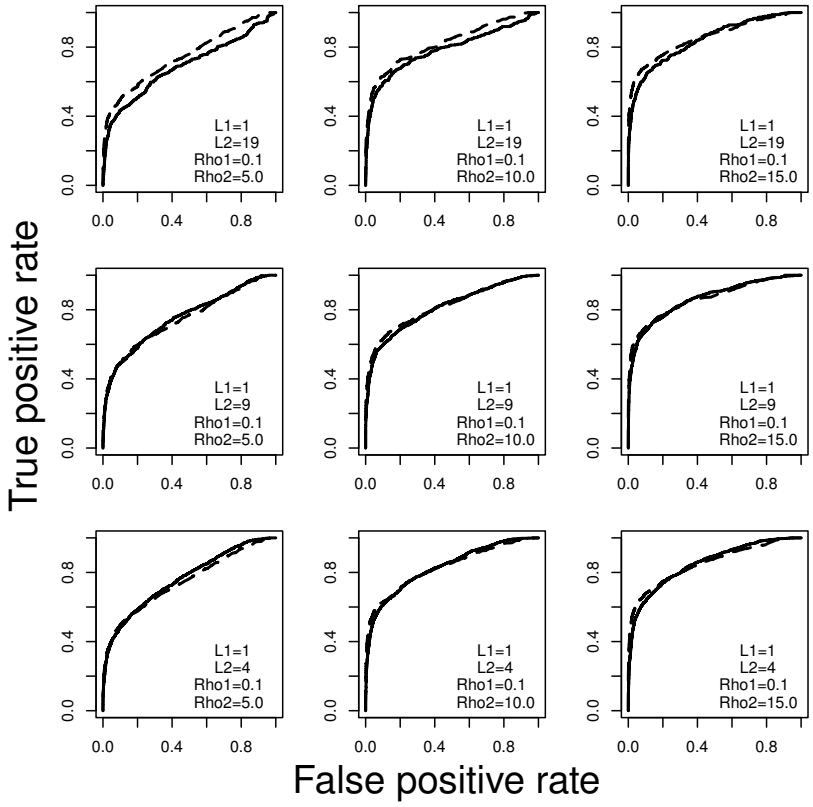


Figure A.11: The comparison of HMMDiverge and DIVERGE2 in the third set of additional simulations. X axes represent false positive rates while Y axes represent true positive rates. The length of 'functional divergence relevant regions', L1, is fixed to 1, which implies individual sites are units of functional divergence. L2: the length of 'functional divergence irrelevant regions'. Rho1: the branch length scale factor in the slowly evolved subfamily. Rho2: the branch length scale factor in the rapidly evolved subfamily. The shape parameter, $\alpha$, is set to 0.5.

# Appendix B

# Supplementary Material for Chapter 3

## B.1 2D Toy Protein Simulations in the Absence of the Spatial Correlation of Site-specific Substitution Rates

Because the strength of spatial correlation of site-specific substitution rates may be very weak in some protein families, we compared the performance of GP4Rate and Rate4Site in simulated alignments in which the spatial correlation of site-specific substitution rates is absent. The simulated alignments were generated by randomly permuting the sites in each alignment in the first spatial configuration of the 2D toy protein simulations. The random permutations destroyed the spatial correlation of site-specific substitution rates but kept the other features of the data. We applied both GP4Rate and Rate4Site to the permuted alignments following the settings described

in the Main Text. Because the spatial correlation of site-specific substitution rates is absent in these permuted alignments, we expected that the characteristic length scales estimated by GP4Rate would be very close to zero. As shown in Figure B.1A, the estimated characteristic length scales are indeed close to zero. The result suggests that GP4Rate can detect the absence of spatial correlation of substitution rates in the permuted alignments.

Because GP4Rate is mainly designed for identifying slowly evolved functional sites in the presence of spatial correlation of substitution rates, it is interesting to test whether it has a similar statistical power as Rate4Site, which explicitly assumes the absence of spatial correlation, in the permuted alignments. Therefore, we plotted the ROC curves to visualize the performance of GP4Rate and Rate4Site. Similar to the 2D toy protein simulations described in the Main Text, we divided the sites into two categories, functional sites and nonfunctional sites, and these two categories were used as true positives and true negatives, respectively, in the ROC curves. The sites that evolved at the lower rate (0.2) were considered to be functional where these that evolved at the higher rate (1.8) were considered to be nonfunctional. As shown in Figure B.2A, GP4Rate and Rate4Site have similar powers as the areas under the ROC curves of GP4Rate and Rate4Site are effectively identical.

As mentioned in the Main Text, ROC curves may not be able to estimate the potential systematic bias of the estimated substitution rates. Therefore, we compared GP4Rate with Rate4Site using the simple loss function proposed in the Main Text. As shown in Figure B.2B, GP4Rate has a lower accuracy than Rate4Site. The higher systematic bias in GP4Rate might be due to the inflexibility of the Gaussian process prior when a spatial correlation is absent. If the spatial correlation of substitution

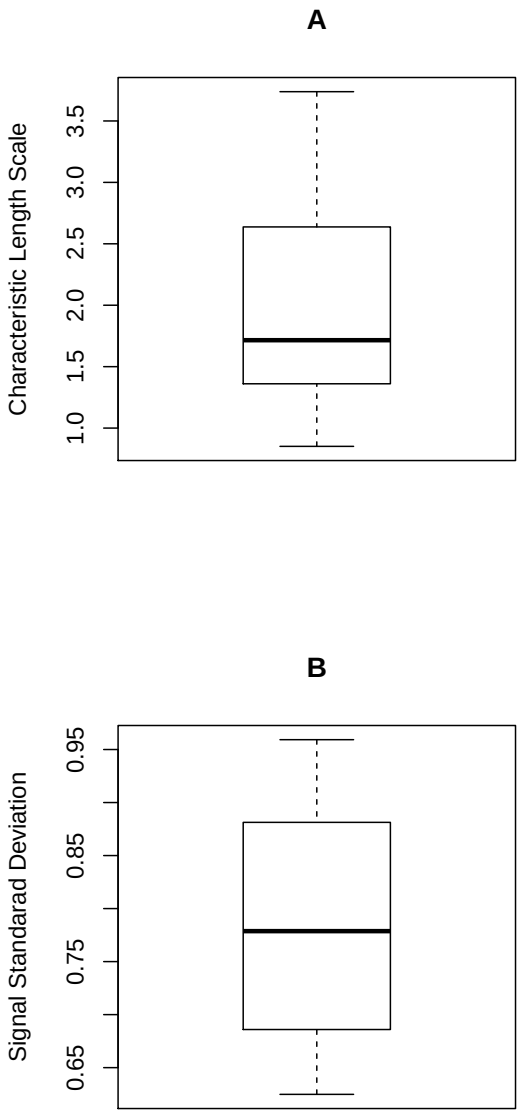


Figure B.1: The hyperparameters estimated by GP4Rate in the 20 permuted alignments. The unit of the characteristic length scale is Å while the signal standard deviation is unitless. (A) the estimated characteristic length scale; (B) the estimated signal standard deviation.

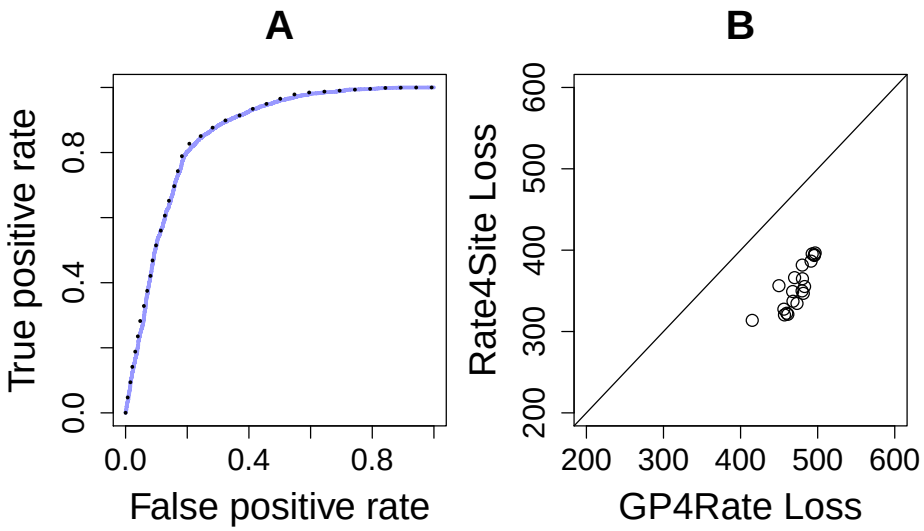


Figure B.2: The quantitative comparison of GP4Rate and Rate4Site in the 20 permuted alignments. (A) the ROC curves of GP4Rate and Rate4Site; (B) the losses of GP4Rate and Rate4Site. In the ROC curves, the solid blue line corresponds to the performance of GP4Rate while the dotted black line corresponds to the performance of Rate4Site. In the plot of losses, each point corresponds to a permuted alignment.

rates is absent, the estimated characteristic length scale will be very close to zero. In this scenario, the site-specific substitution rates are effectively independent and identically distributed (i.i.d) and the Gaussian process prior degenerates to a simple isotropic multivariate Gaussian distribution. Recalling that we assumed that the log values of site-specific substitution rates follow the Gaussian process prior, it means that the site-specific substitution rates effectively follow i.i.d. log-normal distributions. In contrast, Rate4Site assumes that site-specific substitution rates follow i.i.d. discrete Gamma distributions. It is well-known that Gamma distribution is very flexible and can model a variety of distributions with different shapes. In contrast, the log-normal distribution is not as flexible as the Gamma distribution. Nevertheless, in the practice of identifying functional sites, the absolute substitution rates are rarely interesting to researchers, since it is the relative substitution rates that tell us which sites may be functionally important. Because the ROC curves are equivalent between GP4Rate and Rate4Site, GP4Rate should have the same power as Rate4Site for identifying conserved functional sites if the spatial correlation of substitution rates is absent.

## B.2 Bayesian Model Comparision in the Case Study of B7-1 Genes

As mentioned in the main text, it is impractical to compare GP4Rate with Rate4Site directly, since GP4Rate is based on the Bayesian principle while Rate4Site is based on the maximum likelihood principle. Therefore, we developed a Bayesian version of Rate4Site. Because we assumed that both the topology and branch lengths of the phylogenetic tree were fixed in analyses, the only free parameter in Rate4Site is

the shape parameter of the discrete Gamma distribution. In the Bayesian Rate4Site, we assumed that the Gamma shape parameter follows a uniform distribution ranging from 0.05 to 5. The lower boundary was set to 0.05, because very small Gamma shape parameters, which suggest very large variations of site-specific substitution rates, are very unlikely to fit real data well and the discrete Gamma distribution is numerically instable when the Gamma shape parameter is very close to 0. The upper boundary, 5, corresponds to the scenario in which the variation of substitution rates is very small. Because there is only one free parameter in the Bayesian Rate4Site, we numerically integrated it out to calculate the log marginal likelihood of the Bayesian Rate4Site. More specifically, in the numerical integration we divided the range of the Gamma shape parameter into small bins whose sizes are all equal to 0.01. The marginal likelihood may be calculated by the following formula,

$$\mathcal{ML} = \frac{\sum_{i=1}^{K} \mathcal{L}_i^{\text{Mid}}}{K}. \tag{B.1}$$

In the equation, $K$ is the total number of bins in the numerical integration while $\mathcal{L}_i^{\text{Mid}}$ is the phylogenetic likelihood when the Gamma shape parameter is equal to the middle-point of the $i$-th bin. The site-specific substitution rates were also calculated using the same numerical integration algorithm.

To test whether Rate4Site and its Bayesian version lead to similar estimations of the site-specific substitution rates, we applied both the two programs to the B7-1 dataset described in the main text. As shown in Figure B.3, the correlation of estimated site-specific substitution rates is very strong ($\rho > 0.999$). Therefore, the two programs generated essentially the same result and we may use the estimated log marginal likelihood of the Bayesian Rate4Site to measure how good the original

Rate4Site fits the B7-1 dataset.

To calculate the log marginal likelihood of GP4Rate, we applied the steppingstone sampling (SS) algorithm Xie *et al.* (2011). It has been shown that the SS algorithm is a very accurate algorithm to calculate the log marginal likelihood of phylogenetic models Xie *et al.* (2011). The SS algorithm calculates the log marginal likelihood by performing a series of MCMC simulations based on a family of distributions,

$$P(\mathbf{\Phi}, l, \sigma|\mathbf{X}, \mathbf{D}, \mathcal{T}, \beta) \propto P(l, \sigma)P(\mathbf{\Phi}|\mathbf{D}, l, \sigma)\{\prod_{i=1}^{N} \mathcal{L}_i(\Phi_i; \mathbf{X}_i, \mathcal{T})\}^{\beta}. \tag{B.2}$$

The extra parameter $\beta$ reflects the "temperature" of the system. If $\beta = 0$, we essentially sample from the prior distribution. If $\beta = 1$, we essentially sample from the posterior distribution. We choose 21 $\beta$ values which correspond to the quantiles of the Beta$(0.3, 1)$ distribution as suggested by the previous study Xie *et al.* (2011). Then, 20 simulations were performed based on the chosen $\beta$ values, each of which ran $10^6$ iterations. The first 30% of samples were discarded as burn-in. Finally, the log marginal likelihood was calculated based on the 20 simulations Xie *et al.* (2011).

The estimated log marginal likelihood of GP4Rate is equal to $-1705.1$ while the estimated log marginal likelihood of the Bayesian Rate4Site is equal to $-1710.9$. Recall that the Bayes factor is defined as the ratio of the marginal likelihoods of the two alternative models. The Bayes factor of GP4Rate compared with the Bayesian Rate4Site is equal to

$$\mathcal{BF} = e^{-1705.1+1710.9} = 330.3, \tag{B.3}$$

which is significantly greater than 1. Therefore, GP4Rate fits the B7-1 dataset much better than the Bayesian version of Rate4Site.

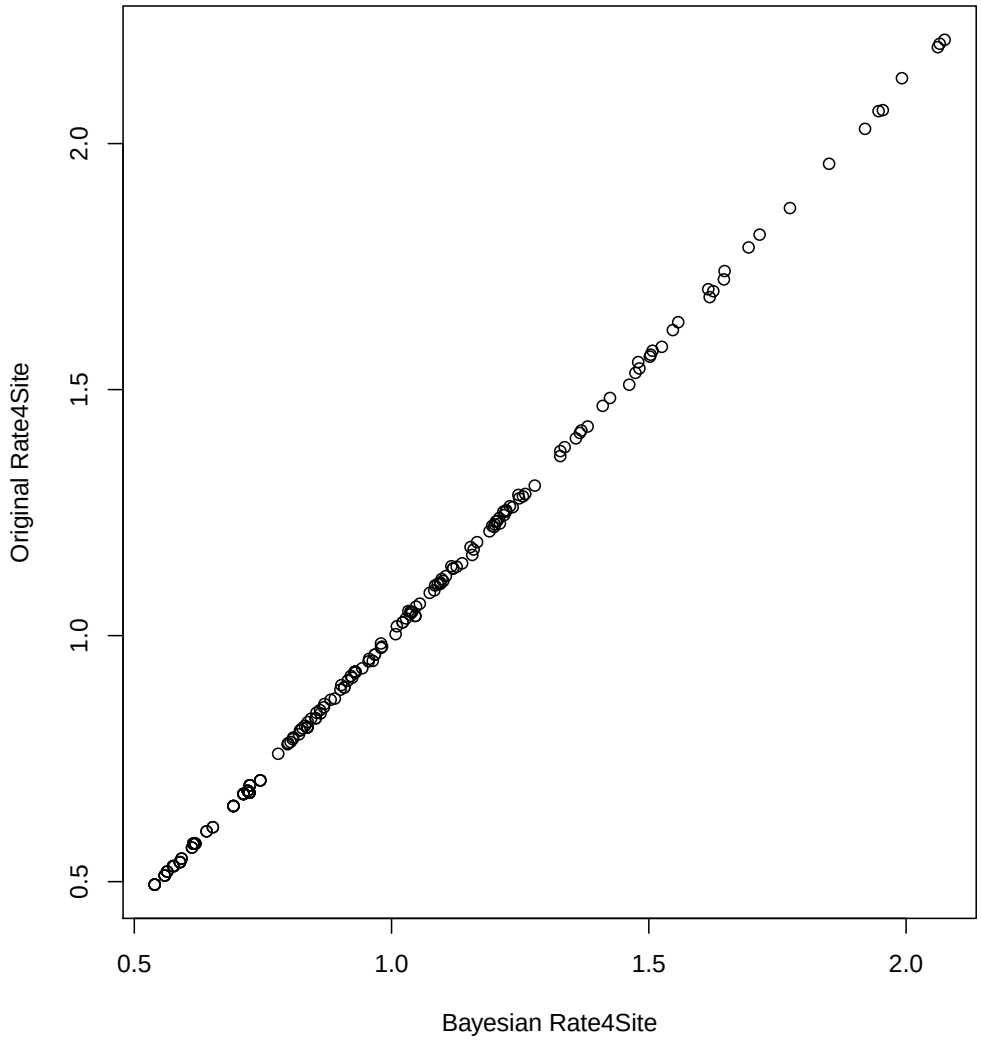


Figure B.3: The site-specific substitution rates estimated by Rate4Site and its Bayesian version in the case study of B7-1 genes. The x-axis corresponds to the site-specific substitution rates estimated by the Bayesian Rate4Site while the y-axis corresponds to the site-specific substitution rates estimated by the original Rate4Site. The Spearman correlation coefficient of the estimated substitution rates is greater than 0.999.

## B.3 List of Conserved Sites Predicted by GP4Rate and Rate4Site in the Case Study of B7-1 Genes

Table B.1: List of the top 20 most conserved sites predicted by GP4Rate and Rate4Site in the case study of B7-1 genes.

| Conserved sites (GP4Rate) | | Conserved sites (Rate4Site) | |
|---|---|---|---|
| site | rate | site | rate |
| 157 | 0.404 | 75 | 0.494 |
| 164 | 0.414 | 131 | 0.494 |
| 158 | 0.418 | 156 | 0.494 |
| 156 | 0.423 | 167 | 0.494 |
| 163 | 0.426 | 168 | 0.494 |
| 154 | 0.452 | 57 | 0.512 |
| 133 | 0.454 | 153 | 0.512 |
| 165 | 0.458 | 154 | 0.512 |
| 159 | 0.463 | 175 | 0.512 |
| 132 | 0.463 | 55 | 0.521 |
| 162 | 0.466 | 144 | 0.521 |
| 166 | 0.470 | 173 | 0.521 |
| 131 | 0.471 | 106 | 0.531 |
| 167 | 0.472 | 30 | 0.532 |
| 153 | 0.473 | 113 | 0.532 |
| 155 | 0.473 | 8 | 0.540 |
| 134 | 0.476 | 39 | 0.540 |
| 168 | 0.497 | 100 | 0.540 |
| 161 | 0.498 | 191 | 0.540 |
| 160 | 0.510 | 96 | 0.547 |

# Bibliography

Abhiman, S. and Sonnhammer, E. L. (2005a). Large-scale prediction of function shift in protein families with a focus on enzymatic function. *Proteins*, **60**(4), 758–768.

Abhiman, S. and Sonnhammer, E. L. L. (2005b). FunShift: a database of function shift analysis on protein subfamilies. *Nucleic Acids Research*, **33**(suppl 1), D197–D200.

Arnau, V., Gallach, M., Lucas, J. I., and Marin, I. (2006). Uvpar: fast detection of functional shifts in duplicate genes. *BMC Bioinformatics*, **7**(1), 174.

Ashkenazy, H., Erez, E., Martz, E., Pupko, T., and Ben-Tal, N. (2010). ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Research*, **38**(suppl 2), W529–W533.

Attisano, L. and Tuen Lee-Hoeflich, S. (2001). The Smads. *Genome Biology*, **2**(8), reviews3010.

Baburajendran, N., Palasingam, P., Narasimhan, K., Sun, W., Prabhakar, S., Jauch, R., and Kolatkar, P. R. (2010). Structure of Smad1 MH1/DNA complex reveals distinctive rearrangements of BMP and TGF-$\beta$ effectors. *Nucleic Acids Research*, **38**(10), 3477–3488.

Bae, H., Anderson, K., Flood, L. A., Skiba, N. P., Hamm, H. E., and Graber, S. G. (1997). Molecular determinants of selectivity in 5-hydroxytryptamine1b receptor-g protein interactions. *Journal of Biological Chemistry*, **272**(51), 32071–32077.

Bae, H., Cabrera-Vera, T. M., Depree, K. M., Graber, S. G., and Hamm, H. E. (1999). Two amino acids within the $\alpha$4 helix of Gi1 mediate coupling with 5-Hydroxytryptamine1B receptors. *Journal of Biological Chemistry*, **274**(21), 14963–14971.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall/CRC, 1 edition.

Berglund, A.-C., Wallner, B., Elofsson, A., and Liberles, D. A. (2005). Tertiary windowing to detect positive diversifying selection. *Journal of Molecular Evolution*, **60**, 499–504.

Bielawski, J. P. and Yang, Z. (2003). Maximum likelihood methods for detecting adaptive evolution after gene duplication. *Journal of Structural and Functional Genomics*, **3**(1), 201–212.

Bishop, C. M. (2007). *Pattern Recognition and Machine Learning*. Springer, 1st ed. 2006. corr. 2nd printing 2011 edition.

Blouin, C., Boucher, Y., and Roger, A. J. (2003). Inferring functional constraints and divergence in protein families using 3d mapping of phylogenetic information. *Nucleic Acids Research*, **31**(2), 790–797.

Cabrera-Vera, T. M., Vanhauwe, J., Thomas, T. O., Medkova, M., Preininger, A.,

Mazzoni, M. R., and Hamm, H. E. (2003). Insights into G protein structure, function, and regulation. *Endocrine Reviews*, **24**(6), 765–781.

Callahan, B., Neher, R. A., Bachtrog, D., Andolfatto, P., and Shraiman, B. I. (2011). Correlated evolution of nearby residues in drosophilid proteins. *PLoS Genetics*, **7**(2), e1001315.

Canagarajah, B. J., Khokhlatchev, A., Cobb, M. H., and Goldsmith, E. J. (1997). Activation mechanism of the MAP kinase ERK2 by dual phosphorylation. *Cell*, **90**(5), 859–869.

Capra, J. A. and Singh, M. (2007). Predicting functionally important residues from sequence conservation. *Bioinformatics*, **23**(15), 1875–1882.

Clamp, M., Cuff, J., Searle, S. M., and Barton, G. J. (2004). The jalview java alignment editor. *Bioinformatics*, **20**(3), 426–427.

Collins, M., Ling, V., and Carreno, B. (2005). The B7 family of immune-regulatory ligands. *Genome Biology*, **6**(6), 223.

Conant, G. C. and Stadler, P. F. (2009). Solvent exposure imparts similar selective pressures across a range of yeast proteins. *Molecular Biology and Evolution*, **26**(5), 1155–1161.

De Maio, N., Holmes, I., Schltterer, C., and Kosiol, C. (2013). Estimating empirical codon hidden markov models. *Molecular Biology and Evolution*, **30**(3), 725–736.

Dean, A. M. and Golding, G. B. (2000). Enzyme evolution explained (sort of). *Pacific Symposium on Biocomputing*, **2000**, 6–17.

Dorman, K. (2007). Identifying dramatic selection shifts in phylogenetic trees. *BMC Evolutionary Biology*, **7**(Suppl 1), S10.

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998). *Biological Sequence Analysis*. Cambridge University Press, eleventh edition.

Dutheil, J., Gaillard, S., Bazin, E., Glemin, S., Ranwez, V., Galtier, N., and Belkhir, K. (2006). Bio++: a set of C++ libraries for sequence analysis, phylogenetics, molecular evolution and population genetics. *BMC Bioinformatics*, **7**(1), 188.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, **32**(5), 1792–1797.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, **17**(6), 368–376.

Felsenstein, J. (1989). PHYLIP – phylogeny inference package (version 3.2). *Cladistics*, **5**, 164–166.

Felsenstein, J. and Churchill, G. A. (1996). A hidden Markov model approach to variation among sites in rate of evolution. *Molecular Biology and Evolution*, **13**(1), 93–104.

Gao, X., Vander Velden, K., Voytas, D., and Gu, X. (2005). SplitTester : software to identify domains responsible for functional divergence in protein family. *BMC Bioinformatics*, **6**(1), 137.

Glaser, F., Pupko, T., Paz, I., Bell, R. E., Bechor-Shental, D., Martz, E., and Ben-Tal, N. (2003). ConSurf: Identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics*, **19**(1), 163–164.

Goldenberg, O., Erez, E., Nimrod, G., and Ben-Tal, N. (2009). The ConSurf-DB: pre-calculated evolutionary conservation profiles of protein structures. *Nucleic Acids Research*, **37**(suppl 1), D323–D327.

Golding, G. B. and Dean, A. M. (1998). The structural basis of molecular adaptation. *Molecular Biology and Evolution*, **15**(4), 355–369.

Goldman, N. and Yang, Z. (1994). A codon-based model of nucleotide substitution for protein-coding dna sequences. *Molecular Biology and Evolution*, **11**(5), 725–736.

Gribaldo, S., Casane, D., Lopez, P., and Philippe, H. (2003). Functional divergence prediction from evolutionary analysis: A case study of vertebrate hemoglobin. *Molecular Biology and Evolution*, **20**(11), 1754–1759.

Gu, X. (1999). Statistical methods for testing functional divergence after gene duplication. *Molecular Biology and Evolution*, **16**(12), 1664–1674.

Gu, X. (2001a). Maximum-likelihood approach for gene family evolution under functional divergence. *Molecular Biology and Evolution*, **18**(4), 453–464.

Gu, X. (2001b). A site-specific measure for rate difference after gene duplication or speciation. *Molecular Biology and Evolution*, **18**(12), 2327–2330.

Gu, X. (2006). A simple statistical method for estimating Type-II (cluster-specific) functional divergence of protein sequences. *Molecular Biology and Evolution*, **23**(10), 1937–1945.

Gueguen, L., Gaillard, S., Boussau, B., Gouy, M., Groussin, M., Rochette, N. C., Bigot, T., Fournier, D., Pouyet, F., Cahais, V., Bernard, A., Scornavacca, C.,

Nabholz, B., Haudry, A., Dachary, L., Galtier, N., Belkhir, K., and Dutheil, J. Y. (2013). Bio++: Efficient extensible libraries and tools for computational molecular evolution. *Molecular Biology and Evolution*, **30**(8), 1745–1750.

Guindon, S. and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, **52**(5), 696–704.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**(1), 97–109.

Huang, Y.-F. and Golding, G. B. (2012). Inferring sequence regions under functional divergence in duplicate genes. *Bioinformatics*, **28**(2), 176–183.

Huang, Y.-F. and Golding, G. B. (2014). Phylogenetic Gaussian process model for the inference of functionally important regions in protein tertiary structures. *PLoS Computational Biology*, **10**(1), e1003429.

Huelsenbeck, J. P. and Suchard, M. A. (2007). A nonparametric method for accommodating and testing across-site rate variation. *Systematic Biology*, **56**(6), 975–987.

Ikemizu, S., Gilbert, R. J., Fennelly, J. A., Collins, A. V., Harlos, K., Jones, E., Stuart, D. I., and Davis, S. J. (2000). Structure and dimerization of a soluble form of B7-1. *Immunity*, **12**, 51 – 60.

Innis, C., Anand, A., and Sowdhamini, R. (2004). Prediction of functional sites in proteins using conserved functional group analysis. *Journal of Molecular Biology*, **337**(4), 1053–1068.

Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Computer Applications in the Biosciences*, **8**(3), 275–282.

Kaziro, Y., Itoh, H., Kozasa, T., Nakafuku, M., and Satoh, T. (1991). Structure and function of signal-transducing GTP-binding proteins. *Annual Review of Biochemistry*, **60**(1), 349–400.

Knudsen, B. and Miyamoto, M. M. (2001). A likelihood ratio test for evolutionary rate shifts and functional divergence among proteins. *Proceedings of the National Academy of Sciences of the United States of America*, **98**(25), 14512–14517.

Knudsen, B., Miyamoto, M. M., Laipis, P. J., and Silverman, D. N. (2003). Using evolutionary rates to investigate protein functional divergence and conservation: A case study of the carbonic anhydrases. *Genetics*, **164**(4), 1261–1269.

Kosiol, C. and Goldman, N. (2005). Different versions of the Dayhoff rate matrix. *Molecular Biology and Evolution*, **22**(2), 193–199.

Kumar, S., Dudley, J. T., Filipski, A., and Liu, L. (2011). Phylomedicine: an evolutionary telescope to explore and diagnose the universe of disease mutations. *Trends in Genetics*, **27**(9), 377–386.

Lambright, D. G., Noel, J., Hamm, H., and Sigler, P. (1994). Structural determinants for activation of the $\alpha$-subunit of a heterotrimeric g protein. *Nature*, **369**(6482), 621–628.

Landgraf, R., Xenarios, I., and Eisenberg, D. (2001). Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. *Journal of Molecular Biology*, **307**(5), 1487–1502.

Lartillot, N. and Philippe, H. (2006). Computing Bayes factors using thermodynamic integration. *Systematic Biology*, **55**(2), 195–207.

Lee, C. H., Katz, A., and Simon, M. I. (1995). Multiple regions of g alpha 16 contribute to the specificity of activation by the c5a receptor. *Molecular Pharmacology*, **47**(2), 218–223.

Liang, H., Zhou, W., and Landweber, L. F. (2006). SWAKK: a web server for detecting positive selection in proteins using a sliding window substitution rate analysis. *Nucleic Acids Research*, **34**, W382–W384.

Lichtarge, O., Bourne, H. R., and Cohen, F. E. (1996). An evolutionary trace method defines binding surfaces common to protein families. *Journal of Molecular Biology*, **257**(2), 342–358.

Madabushi, S., Yao, H., Marsh, M., Kristensen, D. M., Philippi, A., Sowa, M. E., and Lichtarge, O. (2002). Structural clusters of evolutionary trace residues are statistically significant and common in proteins. *Journal of Molecular Biology*, **316**(1), 139–154.

Marin, I., Fares, M. A., Gonzalez-Candelas, F., Barrio, E., and Moya, A. (2001). Detecting changes in the functional constraints of paralogous genes. *Journal of Molecular Evolution*, **52**(1), 17–28.

Mayrose, I., Graur, D., Ben-Tal, N., and Pupko, T. (2004). Comparison of site-specific rate-inference methods for protein sequences: Empirical bayesian methods are superior. *Molecular Biology and Evolution*, **21**(9), 1781–1791.

Mayrose, I., Doron-Faigenboim, A., Bacharach, E., and Pupko, T. (2007). Towards realistic codon models: among site variability and dependency of synonymous and non-synonymous rates. *Bioinformatics*, **23**(13), i319–i327.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, **21**(6), 1087–1092.

Meyer, A. G. and Wilke, C. O. (2013). Integrating sequence variation and protein structure to identify sites under selection. *Molecular Biology and Evolution*, **30**(1), 36–44.

Meyer, A. G., Dawson, E. T., and Wilke, C. O. (2013). Cross-species comparison of site-specific evolutionary-rate variation in influenza haemagglutinin. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **368**(1614).

Muchmore, S. W., Sattler, M., Liang, H., Meadows, R. P., Harlan, J. E., Yoon, H. S., Nettesheim, D., Chang, B. S., Thompson, C. B., Wong, S.-L., Ng, S.-C., and Fesik, S. W. (1996). X-ray and NMR structure of human Bcl-xL, an inhibitor of programmed cell death. *Nature*, **381**(6580), 335–341.

Nam, J., Kaufmann, K., Theiben, G., and Nei, M. (2005). A simple method for predicting the functional differentiation of duplicate genes and its application to mikc-type mads-box genes. *Nucleic Acids Research*, **33**(2), e12.

Neal, R. (1997). Monte Carlo implementation of Gaussian process models for Bayesian regression and classification. Technical report, University of Toronto.

Neal, R. (1999). Regression and classification using Gaussian process priors. *Bayesian Statistics*, **6**, 475–501.

Neer, E. J. (1995). Heterotrimeric c proteins: Organizers of transmembrane signals. *Cell*, **80**(2), 249 – 257.

Neuwald, A. (2010). Bayesian classification of residues associated with protein functional divergence: Arf and arf-like gtpases. *Biology Direct*, **5**(1), 66.

Nimrod, G., Glaser, F., Steinberg, D., Ben-Tal, N., and Pupko, T. (2005). *In silico* identification of functional regions in proteins. *Bioinformatics*, **21**(suppl 1), i328–i337.

Panchenko, A. R., Kondrashov, F., and Bryant, S. (2004). Prediction of functional sites by analysis of sequence and structure conservation. *Protein Science*, **13**(4), 884–892.

Peach, R. J., Bajorath, J., Naemura, J., Leytze, G., Greene, J., Aruffo, A., and Linsley, P. S. (1995). Both extracellular immunoglobin-like domains of cd80 contain residues critical for binding t cell surface receptors CTLA-4 and CD28. *Journal of Biological Chemistry*, **270**(36), 21181–21187.

Press, W., Teukolsky, S., Vetterling, W., and Flannery, B. (1992). *Numerical Recipes in C*. Cambridge University Press, 2nd edition.

Pupko, T. and Galtier, N. (2002). A covarion-based method for detecting molecular

adaptation: application to the evolution of primate mitochondrial genomes. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, **269**(1498), 1313–1316.

Rasmussen, C. E. and Williams, C. K. I. (2005). *Gaussian Processes for Machine Learning*. The MIT Press, 1 edition.

Ridout, K. E., Dixon, C. J., and Filatov, D. A. (2010). Positive selection differs between protein secondary structure elements in *Drosophila*. *Genome Biology and Evolution*, **2**, 166–179.

Robinson, D. M., Jones, D. T., Kishino, H., Goldman, N., and Thorne, J. L. (2003). Protein evolution with dependence among codons due to tertiary structure. *Molecular Biology and Evolution*, **20**(10), 1692–1704.

Rodrigue, N., Lartillot, N., Bryant, D., and Philippe, H. (2005). Site interdependence attributed to tertiary structure in amino acid sequence evolution. *Gene*, **347**(2), 207 – 217. ¡ce:title¿Structural Approaches to Sequence Evolution: Molecules, Networks, Populations Part 2¡/ce:title¿.

Rodrigue, N., Philippe, H., and Lartillot, N. (2006). Assessing site-interdependent phylogenetic models of sequence evolution. *Molecular Biology and Evolution*, **23**(9), 1762–1775.

Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, **4**(4), 406–425.

Sayers, E. W., Barrett, T., Benson, D. A., Bolton, E., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., DiCuccio, M., Federhen, S., Feolo, M., Fingerman,

I. M., Geer, L. Y., Helmberg, W., Kapustin, Y., Krasnov, S., Landsman, D., Lipman, D. J., Lu, Z., Madden, T. L., Madej, T., Maglott, D. R., Marchler-Bauer, A., Miller, V., Karsch-Mizrachi, I., Ostell, J., Panchenko, A., Phan, L., Pruitt, K. D., Schuler, G. D., Sequeira, E., Sherry, S. T., Shumway, M., Sirotkin, K., Slotta, D., Souvorov, A., Starchenko, G., Tatusova, T. A., Wagner, L., Wang, Y., Wilbur, W. J., Yaschenko, E., and Ye, J. (2012). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, **40**(D1), D13–D25.

Scherer, A. and Graff, J. M. (2000). Calmodulin differentially modulates Smad1 and Smad2 signaling. *Journal of Biological Chemistry*, **275**(52), 41430–41438.

Schmid, K. and Yang, Z. (2008). The trouble with sliding windows and the selective pressure in BRCA1. *PLoS ONE*, **3**(11), e3746.

Seger, R. and Krebs, E. G. (1995). The MAPK signaling cascade. *The FASEB Journal*, **9**(9), 726–35.

Siepel, A. and Haussler, D. (2004). Combining phylogenetic and hidden markov models in biosequence analysis. *Journal of Computational Biology*, **11**(2-3), 413–428.

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

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., Weinstock, G. M., Wilson,

R. K., Gibbs, R. A., Kent, W. J., Miller, W., and Haussler, D. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, **15**(8), 1034–1050.

Siepel, A., Pollard, K., and Haussler, D. (2006). New methods for detecting lineage-specific selection. *Lecture Notes in Computer Science*, **3909**, 190–205.

Simon, A. L., Stone, E. A., and Sidow, A. (2002). Inference of functional regions in proteins by quantification of evolutionary constraints. *Proceedings of the National Academy of Sciences*, **99**(5), 2912–2917.

Simon, M., Strathmann, M., and Gautam, N. (1991). Diversity of g proteins in signal transduction. *Science*, **252**(5007), 802–808.

Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005). ROCR: visualizing classifier performance in R. *Bioinformatics*, **21**(20), 3940–3941.

Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., Fuellen, G., Gilbert, J. G., Korf, I., Lapp, H., Lehvaslaiho, H., Matsalla, C., Mungall, C. J., Osborne, B. I., Pocock, M. R., Schattner, P., Senger, M., Stein, L. D., Stupka, E., Wilkinson, M. D., and Birney, E. (2002). The Bioperl toolkit: Perl modules for the life sciences. *Genome Research*, **12**(10), 1611–1618.

Stamper, C. C., Zhang, Y., Tobin, J. F., Erbe, D. V., Ikemizu, S., Davis, S. J., Stahl, M. L., Seehra, J., Somers, W. S., and Mosyak, L. (2001). Crystal structure of the B7-1/CTLA-4 complex that inhibits human immune responses. *Nature*, **410**(6828), 608–611.

Susko, E., Inagaki, Y., Field, C., Holder, M. E., and Roger, A. J. (2002). Testing for differences in rates-across-sites distributions in phylogenetic subtrees. *Molecular Biology and Evolution*, **19**(9), 1514–1523.

Suzuki, Y. (2004). Three-dimensional window analysis for detecting positive selection at structural regions of proteins. *Molecular Biology and Evolution*, **21**(12), 2352–2359.

Turjanski, A. G., Hummer, G., and Gutkind, J. S. (2009). How mitogen-activated protein kinases recognize and phosphorylate their targets: A QM/MM study. *Journal of the American Chemical Society*, **131**(17), 6141–6148.

Tusche, C., Steinbruck, L., and McHardy, A. C. (2012). Detecting patches of protein sites of influenza a viruses under positive selection. *Molecular Biology and Evolution*, **29**(8), 2063–2071.

Vanhatalo, J. and Vehtari, A. (2007). Sparse log Gaussian processes via MCMC for spatial epidemiology. *Journal of Machine Learning Research - Proceedings Track*, **1**, 73–89.

Vanhatalo, J., Pietilainen, V., and Vehtari, A. (2010). Approximate inference for disease mapping with sparse gaussian processes. *Statistics in Medicine*, **29**(15), 1580–1607.

Watabe, T. and Kishino, H. (2013). Spatial distribution of selection pressure on a protein based on the hierarchical bayesian model. *Molecular Biology and Evolution*.

Willighagen, E. and Howard, M. (2007). Fast and scriptable molecular graphics in web browsers without Java3D. *Nature Preceedings*.

Xie, W., Lewis, P. O., Fan, Y., Kuo, L., and Chen, M.-H. (2011). Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Systematic Biology*, **60**(2), 150–160.

Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution*, **39**(3), 306–314.

Yang, Z. (1995). A space-time process model for the evolution of DNA sequences. *Genetics*, **139**(2), 993–1005.

Yang, Z. (1998). Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *MOLECULAR BIOLOGY AND EVOLUTION*, **15**(5), 568–573.

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

Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, **24**(8), 1586–1591.

Yang, Z. and Nielsen, R. (2002). Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Molecular Biology and Evolution*, **19**(6), 908–917.

Yang, Z., Nielsen, R., Goldman, N., and Pedersen, A. (2000). Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, **155**(1), 431–449.

Yokoyama, S., Tada, T., Zhang, H., and Britt, L. (2008). Elucidation of phenotypic adaptations: Molecular analyses of dim-light vision proteins in vertebrates. *Proceedings of the National Academy of Sciences*, **105**(36), 13480–13485.

Zhang, F., Strand, A., Robbins, D., Cobb, M. H., and Goldsmith, E. J. (1994). Atomic structure of the map kinase ERK2 at 2.3 a resolution. *Nature*, **367**(6465), 704–711.

Zhang, Z. and Townsend, J. P. (2009). Maximum-likelihood model averaging to profile clustering of site types across discrete linear sequences. *PLoS Computational Biology*, **5**(6), e1000421.

Zheng, Y., Xu, D., and Gu, X. (2007). Functional divergence after gene duplication and sequence-structure relationship: a case study of g-protein alpha subunits. *Journal of Experimental Zoology Part B*, **308**(1), 85–96.

Zhu, C., Byrd, R., Lu, P., and Nocedal, J. (1997). Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software*, **23**(4), 550–560.

# Glossary

**Bayesian Inference** Bayesian inference is a statistical framework to estimate parameters given an observed dataset. Unlike maximum likelihood inference, Bayesian inference combines the information from both a likelihood function and a prior distribution to generate a posterior distribution which is used to estimate parameters.

**Gaussian Process** A Gaussian process is a stochastic process whose marginal distributions are Gaussian (normal) distributions. Gaussian processes are widely used in machine learning and geostatistics.

**JTT Substitution Model** The JTT substitution model is a protein substitution model firstly described by Jones and colleagues (Jones *et al.*, 1992).

**Laplace Approximation** A Laplace approximation uses a Gaussian (normal) distribution to approximate a complicated probability distribution. The Gaussian distribution is typically constructed by a second-order Taylor expansion at the global maximum of the distribution.

**Likelihood Function** A likelihood function is a function of parameter vectors. It is equal to the probability (or probability density) of the observed data given a

parameter vector.

**Markov Chain Monte Carlo** Markov chain Monte Carlo (MCMC) methods are a collection of algorithms to generate dependent samples from probability distributions using customized Markov chains.

**Markov Model** A Markov model is a stochastic process which assumes that the probability of each state at the current time point is only dependent on the state at the immediately previous time point. Markov models are widely used in phylogenetics to model the evolutionary processes of biological sequences.

**Maximum Likelihood Inference** Maximum likelihood inference is a statistical framework to infer parameters by maximizing the likelihood function with regard to unknown parameters.

**Phylogenetic Hidden Markov Model** A phylogenetic hidden Markov model (phylo-HMM) combines a phylogenetic model and a hidden Markov model to model the spatial distribution of evolutionary patterns along biological sequences.

**Pruning Algorithm** The pruning algorithm is a dynamic programming algorithm to calculate the likelihood function of a phylogenetic model. In machine learning literature, it is also known as the sum-product or the belief-propagation algorithm.

**Receiver Operating Characteristic Curve** A receiver operating characteristic (ROC) curve is a plot to visualize the performance of a binary classifier. The power of a binary classifier can be quantitatively measured by the area under its ROC curve.

**Relative Solvent Accessibility** The relative solvant accessibility (RAS) is the ratio of the measured solvent accessibility of a residue X in a PDB structure and its solvent accessibility in a typical tripeptide, i.e. Gly-X-Gly.

**Type-I Functional Divergence** After gene duplication, some protein sites/regions may evolve at different substitution rates in the two duplicate gene copies, which is referred to as type-I functional divergence.

**Type-II Functional Divergence** In duplicate genes, the substitution rates at some protein sites/regions may increase at the early stage of duplication but decrease at the late stage of duplication, which is referred to as type-II functional divergence.