BIOINFORMATICS ORIGINAL PAPER

Vol. 31 no. 4 2015, pages 523–531
doi:10.1093/bioinformatics/btu673

Sequence analysis

Advance Access publication October 15, 2014

FuncPatch: a web server for the fast Bayesian inference of conserved functional patches in protein 3D structures

Yi-Fei Huang[†] and G. Brian Golding[*]

Department of Biology, McMaster University, Hamilton, ON L8S4K1, Canada

Associate Editor: John Hancock

ABSTRACT

Motivation: 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 may reduce their power to identify conserved functional patches in protein tertiary structures when the sequences used in the analysis are highly similar. 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.

Results: Both simulations and four case studies based on empirical data suggest that FuncPatch is a good approximation to GP4Rate. However, FuncPatch is orders of magnitudes faster than GP4Rate. In addition, simulations suggest that FuncPatch is potentially a useful tool complementary to Rate4Site, but the 3D sliding window method is less powerful than FuncPatch and Rate4Site. The functional patches predicted by FuncPatch in the four case studies are supported by experimental evidence, which corroborates the usefulness of FuncPatch.

Availability and implementation: The software FuncPatch is freely available at the web site, http://info.mcmaster.ca/yifei/FuncPatch

Contact: golding@mcmaster.ca

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

Received on January 24, 2014; revised on October 1, 2014; accepted on October 13, 2014

1 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 (Ashkenazy *et al*., 2010; Glaser *et al*., 2003), 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 may make it difficult to infer conserved functional patches in protein tertiary structures. The problem is especially acute when the sequences used in the analysis are very similar to each other, because it is difficult to estimate site-specific substitution rates accurately in this scenario due to the limited information at each amino acid site.

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 (Capra and Singh, 2007; Dean and Golding, 2000; Landgraf *et al*., 2001; Nimrod *et al*., 2005; Panchenko *et al*., 2004;). These methods are useful tools for inferring conserved protein patches, especially when the sequences in the dataset are highly similar. However, most of these methods are based on the 3D sliding window framework (Capra and Singh, 2007; Dean and Golding, 2000; Landgraf *et al*., 2001; 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

[*]To whom correspondence should be addressed.

[†]Present address: Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, BC V6T2B5, Canada

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 that 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 Monte 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. Both simulations and four case studies based on empirical data demonstrate that FuncPatch is a very accurate approximation to GP4Rate. In simulations, FuncPatch is always among the most powerful methods regardless of the assumptions of the simulations. Furthermore, FuncPatch could be more powerful than Rate4Site when a spatial correlation of substitution rates is present and the sequences in the simulated alignments are highly similar. In contrast, the performance of the 3D sliding window method is generally very poor. Therefore, FuncPatch is complementary to Rate4Site and is particularly useful to infer conserved 3D patches in a set of highly similar protein sequences. The four case studies 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.

2 MODELS

2.1 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 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. First, 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 and Vehtari, 2007; Vanhatalo *et al.*, 2010). 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.

2.2 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: (i) it must match an amino acid in the query PDB chain; (ii) at least three sequences in the original alignment must have amino acids instead of gaps at the site; (iii) 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 **X** and the number of sites in the alignment **X** is denoted by N. Then, FuncPatch uses the PROTPARS program in PHYLIP (Felsenstein, 1989) to estimate, **C**, a vector of the most parsimonious number of substitutions for each site in **X**, in which a single element C_i is the most parsimonious number of substitutions at site i.

FuncPatch assumes that C_i, the most parsimonious number of substitutions at site i, follows a Poisson distribution,

$$P(C_i|\lambda_i) = \frac{\lambda_i^{C_i}}{C_i!} e^{-\lambda_i}, \tag{1}$$

where λ_i denotes the expected number of substitutions at site i [which will be described in detail in Equation (5)].

2.3 Gaussian prior distribution

Given the alignment **X**, FuncPatch calculates, β, the average number of substitutions over all sites,

$$\beta = \frac{\sum_{i=1}^{N} C_i}{N}. \tag{2}$$

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

$$P(\Phi|\mathbf{D}, l, \sigma) = \frac{1}{(2\pi)^{\frac{N}{2}}|\Sigma(\mathbf{D}, l, \sigma)|^{\frac{1}{2}}} \exp\left(-\frac{\Phi^T \Sigma(\mathbf{D}, l, \sigma)^{-1}\Phi}{2}\right), \tag{3}$$

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

$$\Sigma_{ij} = \sigma^2\left(1 + \frac{\sqrt{3}D_{ij}}{l}\right)\exp\left(-\frac{\sqrt{3}D_{ij}}{l}\right), \tag{4}$$

where D_{ij} is the Euclidean distance between sites i and j, l is the characteristic length scale and σ 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. σ 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), Σ_{ij} decreases with increasing D_{ij}, which implies that the closely located sites are more likely to have similar substitution rates than the 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 λ_i, i.e. the expected number of substitutions at site i, and Φ_i, i.e. the log substitution rate at site i, can be described by the following equation,

$$\lambda_i = \beta \exp(\Phi_i), \tag{5}$$

where β is the average number of substitutions calculated by Equation (2). In this parameterization, the site-specific substitution rate, i.e. $\exp(\Phi_i)$, is a scaling factor, which is analogous to the substitution rate parameters in statistical phylogenetic models.

By inserting Equation (5) into Equation (1) and then combining Equations (3) and (1), we obtain the posterior distribution of Φ,

$$\underbrace{P(\Phi|l, \sigma, \mathbf{C}, \mathbf{D})}_{\text{Posterior}} \sim \underbrace{P(\Phi|\mathbf{D}, l, \sigma)}_{\text{Prior}} \prod_{i=1}^{N} \underbrace{P(C_i|\beta \exp(\Phi_i))}_{\text{Likelihood}}. \tag{6}$$

Note that the right-hand side of Equation (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 σ. 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 Φ 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 Φ (Rasmussen and Williams, 2005). For each site i in **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 Φ_i and S_i is the approximate posterior standard deviation of Φ_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 75% quantile of the standard Gaussian distribution.

2.4 Inference of hyperparameters and Bayesian model comparison

The descriptions in the previous sections assume that the two hyperparameters, i.e. l and σ, 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 representatives l, evenly spaced between 1 and 39 Å, and 20 representatives σ, 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 σ 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 versus 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.

3 SIMULATIONS AND CASE STUDIES

3.1 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 (Berglund *et al.*, 2005; Dean and Golding, 2000; Suzuki, 2004). 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 first 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.

We generated four sets of simulated alignments (A–D) based on the human Bcl-xL protein to benchmark the performances of FuncPatch, GP4Rate, Rate4Site and the customized 3D sliding window program, respectively. To generate simulated alignments, we need to specify a substitution model, a reference protein structure, a reference phylogenetic tree and a set of reference site-specific substitution rates. In all simulations, the JTT substitution model (Jones *et al.*, 1992) was used to describe the substitution process of amino acids and the protein tertiary structure of the human Bcl-xL protein (PDB ID: 1MAZ; Muchmore *et al.*, 1996) was used as the reference structure. The reference phylogenetic trees and the reference site-specific substitution rates were different in the four simulated datasets, which will be described in detail below.

- The 100 alignments in dataset A were generated using a small reference phylogenetic tree. As shown in Supplementary Figure S1A, the small phylogenetic tree consists of four sequences and the total branch length is equal to 1. Therefore, the expected number of substitutions per site is equal to 1, which is comparable to the expected numbers of substitutions per site in many empirical datasets which only consist of orthologous sequences. We expect that it is relatively difficult to estimate site-specific substitution rates accurately in this scenario due to the relatively small number of substitutions per site. The reference 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. It is expected that the spatial correlation of substitution rates is present in this dataset, since a previous study based on Rate4Site has suggested that the inferred conserved sites are clustered together in the protein tertiary structure of the human Bcl-xL protein (Glaser *et al.*, 2003).
- The 100 alignments in dataset A were permuted randomly to generate the 100 alignments in dataset B. The permutations destroyed the spatial correlation of substitution rates but kept all other features of the alignments. This dataset was designed to test the performances of different methods when the reference phylogenetic tree is small and a spatial correlation of substitution rates is absent.
- The 100 alignments in dataset C were based on a large reference phylogenetic tree (Supplementary Fig. S1B). There are 12 sequences in the phylogenetic tree and the expected number of substitutions per site is equal to 4.2, which is comparable to the expected numbers of substitutions per site in many empirical datasets consisting of paralogous sequences. Therefore, we predict that it is relatively easy to estimate site-specific substitution rates in this scenario, because the generated protein sequences are

highly diverged. Similar to dataset A, the substitution rates downloaded from the ConSurf-DB database (Goldenberg *et al*., 2009) were used to generate dataset C. Therefore, the spatial correlation of site-specific substitution rates is present in this dataset.

- The 100 alignments in dataset C were permuted randomly to generate the 100 alignments in dataset D. This dataset was designed to test the performances of different methods when the reference phylogenetic tree is large while a spatial correlation of substitution rates is absent.

We applied FuncPatch, GP4Rate, Rate4Site and the customized 3D sliding window program to the four sets of simulated alignments. The reference phylogenetic trees (Supplementary Fig. S1) were used as the input trees 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 four programs. To perform the ROC analyses, we need to define a cutoff of percentage which corresponds to the proportion of functional sites in the simulated alignments. We assumed that the cutoff of percentage is equal to 10%, which suggests that the 10% of sites with the lowest reference substitution rates were considered as functional sites. Similar results are reached using different cutoffs, e.g. 30 and 50% (data not shown).

As shown in Figure 1, FuncPatch and GP4Rate generally have similar powers in the four simulated datasets. More importantly, FuncPatch and GP4Rate are always among the most powerful methods. Therefore, FuncPatch is a good approximation to GP4Rate and the two methods are robust against different assumptions. More specifically, when the spatial correlation of substitution rates is present and the reference phylogenetic tree is small (Fig. 1A), FuncPatch and GP4Rate could be more powerful than Rate4Site. When a spatial correlation of substitution rates is absent and the reference phylogenetic tree is small, FuncPatch is only marginally less powerful than GP4Rate and Rate4site (Fig. 1B). When a large phylogenetic tree is used in simulations, the powers of FuncPatch and GP4Rate are very similar to that of Rate4Site regardless of whether the spatial correlation of substitution rates is present or not (Fig. 1C and D). It is not surprising, since when the phylogenetic tree is large, each site in the alignment is informative and it is relatively easy to infer the site-specific substitution rate accurately based on the information at a single site. The 3D sliding window method is almost always among the least powerful methods in the four simulated datasets, which may be due to the difficulty of determining the optimal window size.

In summary, the simulations demonstrate that FuncPatch is a good approximation to GP4Rate and may be a useful tool complementary to Rate4Site. In contrast, the performance of the 3D sliding window method is generally very poor.

3.2 Case study of MAPK1 genes

To demonstrate the power of FuncPatch in analyses with real data, we applied FuncPatch to four signal transduction-related gene families to predict conserved functional patches. As shown

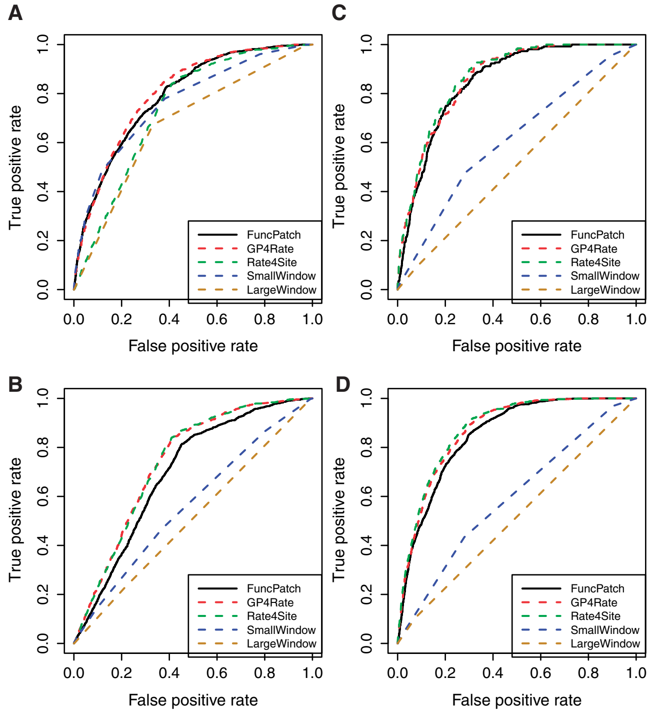

Fig. 1. The performances of different methods in the simulation study. **A** The performances of different methods in the simulated dataset A. **B** The performances of different methods in the simulated dataset B. **C** The performances of different methods in the simulated dataset C. **D** The performances 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)

in the previous section of simulations, the performance of the 3D sliding window method is poor, so we only compared FuncPatch with GP4Rate and Rate4Site. 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). The dataset corresponds to a set of representative MAPK1 sequences collected from vertebrates, invertebrates, plants and fungi. 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

Table 1. Estimation of parameters and log Bayes factors in the case study of the MAPK1 genes and the case study of the SMAD genes

Dataset	l	σ	Log Bayes factor
MAPK1	21	1.3	148.7
SMAD	9	1.1	9.19

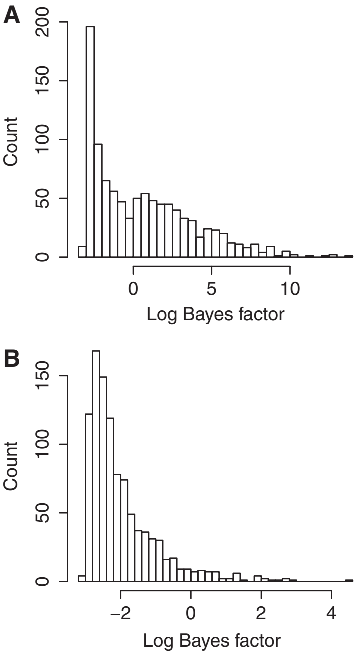

Fig. 2. The null distributions of approximate log Bayes factors in the case study of the MAPK1 genes and the case study of the SMAD genes. The null distributions are generated by applying FuncPatch to the permuted alignments. **A** The null distribution of the approximate log Bayes factors in the case study of the MAPK1 genes. **B** The null distribution of the approximate log Bayes factors in the case study of the SMAD genes

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 (Edgar, 2004) with default parameters. The phylogenetic tree was inferred using PhyML with the JTT + Γ 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 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 versus model 0) is equal to 148.7 which is significantly larger than the cutoff 8 (Table 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 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 sites in MAPK1. As shown in Figure 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 residue (Canagarajah *et al.*, 1997; Turjanski *et al.*, 2009; Zhang *et al.*, 1994). Therefore, the predicted conserved patch corresponds to the catalytic site of MAPK1. The high conservation of neighboring sites of Asp-147 suggests that these sites might be important to mediate the interaction between MAPK1 and its substrate proteins or maintain a suitable local environment for the kinase reaction.

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 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 whereas 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 (Fig. 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 residue, 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.

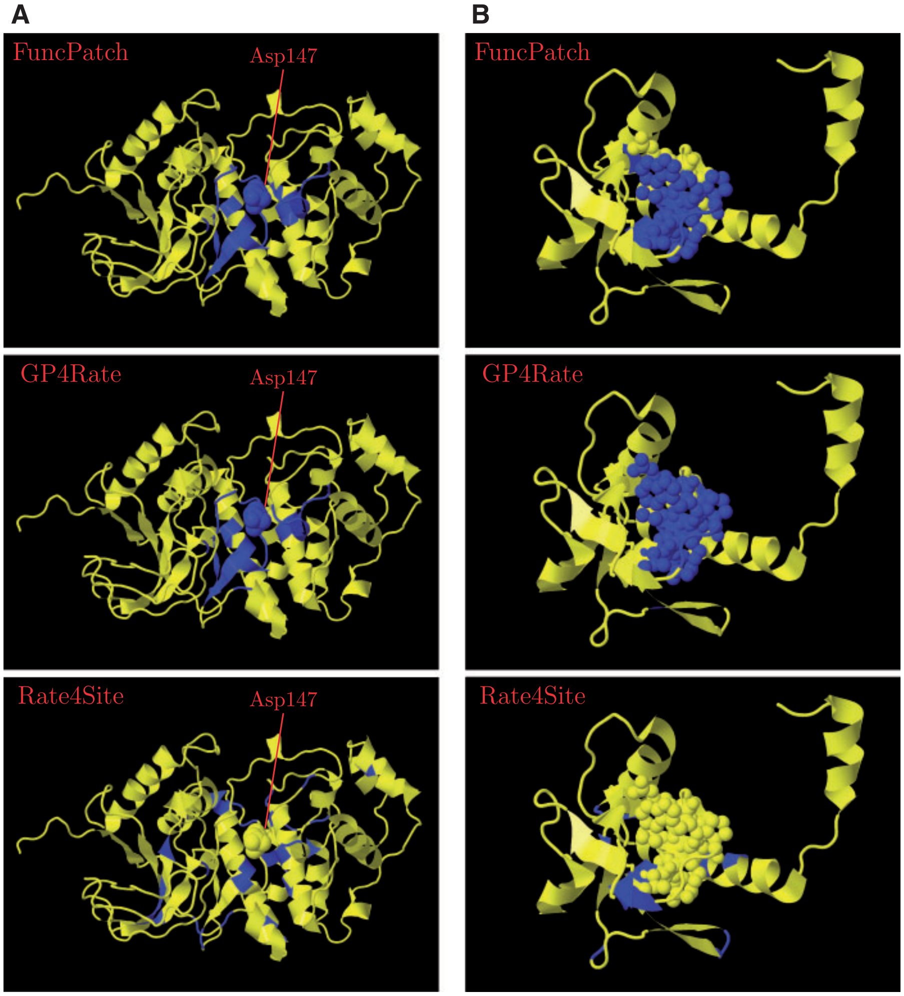

Fig. 3. The 3D locations of the most conserved sites in the case study of the MAPK1 genes and the case study of the SMAD genes. **A** The 35 most conserved sites predicted by FuncPatch, GP4Rate and Rate4Site in the MAPK1 dataset (PDB ID: 1ERK). The blue sites are the predicted conserved sites while the yellow sites are not conserved. The space-filled atoms belong to Asp-147, the catalytic residue. **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)

3.3 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 + Γ model (Guindon and Gascuel, 2003) was used to generate the reference phylogenetic tree. We did not include other SMAD 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, 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 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 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 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 (Fig. 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 α-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 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 whereas 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 (Fig. 3B). Therefore, FuncPatch identified an experimentally verified conserved region in the SMAD dataset which is overlooked by Rate4Site.

3.4 Case studies of MDM2 genes and RAS genes

MDM2 genes and RAS genes are important players in signal transduction networks and many somatic mutations in these genes can be drivers of cancer progression (Weinberg, 2013). Because MDM2 genes and RAS genes are important and well studied, we applied FuncPatch, GP4Rate and Rate4Site to infer conserved functional sites in these genes. We summarize the results in this section and more details are described in the Supplementary Material.

In both of the two case studies, FuncPatch and GP4Rate predicted conserved patches while Rate4Site predicted spatially scattered conserved sites. It is not surprising, since Rate4Site implicitly ignores the possibility of spatial clustering of conserved sites in protein tertiary structures. In addition, the conserved patches predicted by FuncPatch and GP4Rate are essentially identical, which again confirms that FuncPatch is a good approximation to GP4Rate. However, FuncPatch is orders of magnitudes faster than GP4Rate. By comparing the predicted conserved sites with previous experimental studies, we found that many conserved functional sites predicted by FuncPatch are supported by previous experimental studies (Freedman *et al.*, 1997; Vetter *et al.*, 1999). In contrast, only one conserved functional site predicted by Rate4Site is directly supported by existing experimental evidence. These results again highlight that FuncPatch may be a potentially useful tool complementary to Rate4Site.

4 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 that GP4Rate is a powerful method to infer conserved functional regions and is potentially a useful complementary to Rate4Site, but GP4Rate is a slow MCMC program. In this study, 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 four case studies suggest that FuncPatch is an accurate approximation to GP4Rate. The simulations also show that FuncPatch may be a useful complementary to Rate4Site but the 3D sliding window method typically leads to bad results. The conserved patches predicted by FuncPatch in the four 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 model comparison method has been implemented in FuncPatch to test whether the spatial correlation of substitution rates is significant in a dataset. The four case studies of empirical data 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 four 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. As shown in the Supplementary Material, the site-specific substitution rates estimated by FuncPatch are strongly correlated with the rates estimated by Rate4Site in two large gene families each of which consists of highly diverged sequences. Together with the results in simulations, it seems that FuncPatch and Rate4Site typically give similar results when the sequence divergence levels are high. In practice, 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). Therefore, including remote homologs may reduce the power of detecting functional patches unique to an orthologous family. If researchers believe that a family of orthologous genes has distinct biological functions from its remote homologs, it is beneficial to infer conserved patches solely based on the family itself. The traditional methods, e.g. Rate4Site, may not be very powerful in this scenario because of the limited information about the substitution rate at each site. FuncPatch may alleviate the problem, because the Gaussian process prior used in FuncPatch can let an amino acid site to borrow statistical strength from neighboring sites with similar substitution rates. Therefore, FuncPatch is particularly useful for analyzing small gene families or gene subfamilies derived from recent gene duplication events.

ACKNOWLEDGEMENTS

We thank Ben Evans, Terri Porter, Wilson Sung and the anonymous reviewers for insightful comments on this work.

Funding: This work was financially supported by a National Sciences and Engineering Research Council of Canada (NSERC) Discovery grant to GBG [grant number RGPIN140221-10].

Conflict of interest: None declared.

REFERENCES

Ashkenazy,H. *et al.* (2010) ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res.*, **38**, W529–W533.

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

Baburajendran,N. *et al.* (2010) Structure of Smad1 MH1/DNA complex reveals distinctive rearrangements of BMP and TGF-β effectors. *Nucleic Acids Res.*, **38**, 3477–3488.

Berglund,A.-C. *et al.* (2005) Tertiary windowing to detect positive diversifying selection. *J. Mol. Evol.*, **60**, 499–504.

Canagarajah,B.J. *et al.* (1997) Activation mechanism of the MAP kinase ERK2 by dual phosphorylation. *Cell*, **90**, 859–869.

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

Dean,A.M. and Golding,G.B. (2000) Enzyme evolution explained (sort of). *Pac. Symp. Biocomput.*, **2000**, 6–17.

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

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

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

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

Freedman,D.A. *et al.* (1997) A genetic approach to mapping the p53 binding site in the MDM2 protein. *Mol. Med.*, **3**, 248–259.

Glaser,F. *et al.* (2003) ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics*, **19**, 163–164.

Goldenberg,O. *et al.* (2009) The ConSurf-DB: pre-calculated evolutionary conservation profiles of protein structures. *Nucleic Acids Res.*, **37**, D323–D327.

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

Gueguen,L. *et al.* (2013) Bio++: efficient extensible libraries and tools for computational molecular evolution. *Mol. Biol. Evol.*, **30**, 1745–1750.

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

Huang,Y.-F. and Golding,G.B. (2012) Inferring sequence regions under functional divergence in duplicate genes. *Bioinformatics*, **28**, 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 Comput. Biol.*, **10**, e1003429.

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

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

Landgraf,R. *et al.* (2001) Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. *J. Mol. Biol.*, **307**, 1487–1502.

Madabushi,S. *et al.* (2002) Structural clusters of evolutionary trace residues are statistically significant and common in proteins. *J. Mol. Biol.*, **316**, 139–154.

Mayrose,I. *et al.* (2004) Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. *Mol. Biol. Evol.*, **21**, 1781–1791.

Muchmore,S.W. *et al.* (1996) X-ray and NMR structure of human Bcl-xL, an inhibitor of programmed cell death. *Nature*, **381**, 335–341.

Nimrod,G. *et al.* (2005) *In silico* identification of functional regions in proteins. *Bioinformatics*, **21**, i328–i337.

Panchenko,A.R. *et al.* (2004) Prediction of functional sites by analysis of sequence and structure conservation. *Protein Sci.*, **13**, 884–892.

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

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

Sayers,E.W. *et al.* (2012) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **40**, D13–D25.

Scherer,A. and Graff,J.M. (2000) Calmodulin differentially modulates Smad1 and Smad2 signaling. *J. Biol. Chem.*, **275**, 41430–41438.

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

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

Stajich,J.E. *et al.* (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.*, **12**, 1611–1618.

Suzuki,Y. (2004) Three-dimensional window analysis for detecting positive selection at structural regions of proteins. *Mol. Biol. Evol.*, **21**, 2352–2359.

Turjanski,A.G. *et al.* (2009) How mitogen-activated protein kinases recognize and phosphorylate their targets: a QM/MM study. *J. Am. Chem. Soc.*, **131**, 6141–6148.

Vanhatalo,J. and Vehtari,A. (2007) Sparse log Gaussian processes via MCMC for spatial epidemiology. *J. Mach. Learn. Res. – Proc. Track*, **1**, 73–89.

Vanhatalo,J. *et al.* (2010) Approximate inference for disease mapping with sparse Gaussian processes. *Stat. Med.*, **29**, 1580–1607.

Vetter,I. *et al.* (1999) Structural and biochemical analysis of RAS-effector signaling via RaLGDS. *FEBS Lett.*, **451**, 175–180.

Weinberg,R.A. (2013) *The Biology of Cancer*. Garland Publishing, New York.

Willighagen,E. and Howard,M. (2007) Fast and scriptable molecular graphics in web browsers without Java3D. *Nature Precedings*, http://dx.doi.rog/10.1038.npre.2007.50.1.

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

Zhang,F. *et al.* (1994) Atomic structure of the map kinase ERK2 at 2.3 a resolution. *Nature*, **367**, 704–711.

Zhu,C. *et al.* (1997) Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Trans. Math. Softw.*, **23**, 550–560.