BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 19 2011, pages 2655–2663
doi:10.1093/bioinformatics/btr470

Sequence analysis

Advance Access publication August 11, 2011

A phylogenetic mixture model for the identification of functionally divergent protein residues

Daniel Gaston[1,2], Edward Susko[1,3] and Andrew J. Roger[1,2,*]

[1]Centre for Comparative Genomics and Evolutionary Bioinformatics, [2]Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, Canada, B3H 1X5 and [3]Department of Mathematics and Statistics, Dalhousie University, Halifax, Canada, B3H 3J5

Associate Editor: David Posada

ABSTRACT

Motivation: To understand the evolution of molecular function within protein families, it is important to identify those amino acid residues responsible for functional divergence; i.e. those sites in a protein family that affect cofactor, protein or substrate binding preferences; affinity; catalysis; flexibility; or folding. Type I functional divergence (FD) results from changes in conservation (evolutionary rate) at a site between protein subfamilies, whereas type II FD occurs when there has been a shift in preferences for different amino acid chemical properties. A variety of methods have been developed for identifying both site types in protein subfamilies, both from phylogenetic and information-theoretic angles. However, evaluation of the performance of these methods has typically relied upon a handful of reasonably well-characterized biological datasets or analyses of a single biological example. While experimental validation of many truly functionally divergent sites (true positives) can be relatively straightforward, determining that particular sites do not contribute to functional divergence (i.e. false positives and true negatives) is much more difficult, resulting in noisy 'gold standard' examples.

Results: We describe a novel, phylogeny-based functional divergence classifier, FunDi. Unlike previous approaches, FunDi uses a unified mixture model-based approach to detect type I and type II FD. To assess FunDi's overall classification performance relative to other methods, we introduce two methods for simulating functionally divergent datasets. We find that the FunDi method performs better than several other predictors over a wide variety of simulation conditions.

Availability: http://rogerlab.biochem.dal.ca/Software

Contact: andrew.roger@dal.ca

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

Received on April 18, 2011; revised on July 7, 2011; accepted on August 7, 2011

1 INTRODUCTION

Functional divergence in proteins over evolutionary time includes the processes of sub- and neo-functionalization after gene duplication, as well as specialization or loss of functions of proteins in distinct organismal lineages (Henikoff *et al.*, 1997; Li, 1983). Two main patterns of functional divergence (FD) at the amino acid residue level have been described in the literature and were classified by Gu (1999, 2001) as Type I and Type II. In the case of a protein family composed of two subgroups, Type I functional divergence is characterized by greater conservation at a site in one subfamily versus the other subfamily, indicating a difference in evolutionary rate between them due to fewer selective constraints in the more rapidly evolving group. For Type II, divergence sequence conservation at a site is observed in both subfamilies, but with a marked preference for different amino acids, generally with very different physicochemical properties in each group. Accurate prediction of (FD) residues, also known as 'specificity determining sites' in the case where divergence changes the substrate that is bound (Gerlt and Babbitt, 2000), leads to an enhanced understanding of the mechanisms underlying functional diversification.

Three main approaches have been used for the prediction of FD protein residues that, in broad terms, can be classified as primarily phylogenetic, information theoretic or biophysical. Phylogenetic approaches such as evolutionary trace (Lichtarge *et al.*, 1996), DIVERGE (Gu, 1999, 2001; Gu and Vander Velden, 2002) and various likelihood ratio test/rate shift based tests (Knudsen and Miyamoto, 2001; Knudsen *et al.*, 2003; Susko *et al.*, 2002) explicitly take into account a phylogenetic tree that describes the evolutionary relationships among the sequences in the protein family under consideration. In general, phylogenetic methods for FD prediction correlate observed patterns of amino acid substitution at a site in a multiple sequence alignment across subgroups within a phylogenetic tree. Local conservation (i.e. within a subgroup on a phylogenetic tree) relative to other sequences reflect probable functional specificity of that subgroup if the degree of conservation is large relative to the overall divergences of the sequences within that subgroup. This general case can be extended to more rigorous statistical models of functional divergence such as the type I and type II specific prediction methods employed by DIVERGE (Gu, 1999, 2001; Gu and Vander Velden, 2002).

In contrast, information-theoretic approaches do not generally explicitly consider the relationship between sequences, only the known or predicted divisions into functional subgroups and perhaps some weighting based on overall sequence distances as in GroupSim (Capra and Singh, 2008). These approaches contrast information-theoretic measures of variation of site profiles within a subgroup to those observed at a site across the whole multiple sequence alignment. These profiles may most commonly be represented by some information-theoretic measure of variability among residues

*To whom correspondence should be addressed.

at a site such as the Jensen–Shannon Divergence (Lin, 1991), Relative Entropy/Kullback–Leibler Divergence (Kullback and Leibler, 1951), Sequence Harmony (Feenstra *et al*., 2007; Pirovano *et al.*, 2006) and/or simple Shannon Entropy.

Biophysical/structural methods may also include some measures of sequence diversity/information content as above, with a greater focus on the physicochemical properties of structurally conserved residue positions. Active sites modeling and clustering (de Melo-Minardi *et al.*, 2010) compares profiles of structurally aligned and modeled active sites to identify specificity-determining residues. Surface map techniques have also been developed (Pawlowski and Godzik, 2001; Sael *et al.*, 2008) that compare properties such as charge and hydrophobicity of surface proteins of two proteins. Other methods have been used to predict the substrate specificity of unknown family members in cases where annotation transfer from paralogs (Caffrey *et al.*, 2008) and the related task of identifying functional sites (Capra *et al.*, 2009; Sankararaman *et al.*, 2010) may not be adequate. No structural methods were evaluated as part of this study due to the lack of adequate methods for simulating evolutionary divergence in the context of protein structure.

Here, we introduce a new phylogeny-based method, called FunDi, for detecting FD sites across a phylogenetic split in a protein family tree. By explicitly modeling type I and type II functional divergence using a mixture model, FunDi provides a maximum likelihood phylogenetic framework to predict FD sites using specific models of amino acid substitution. As an open framework for functional divergence classification, FunDi is easily extended to accommodate the latest methods/programs for maximum likelihood-based phylogenetic reconstruction and new, more accurate models of amino acid substitution. We also evaluate whether a weighted average of FunDi's score and the Jensen–Shannon Divergence scores of surrounding residues (Capra and Singh, 2008) improves performance.

A number of well-characterized protein datasets have been used for evaluations of the performance of some functional divergence/specificity-determining classifiers (FD classifiers) (Chakrabarti *et al.*, 2007). One limitation of these biological datasets is the difficulty in assigning the labels of 'true negative' or 'false positive' to sites that are not involved in functional divergence. Thorough molecular characterization of every amino acid position in a protein family is practically infeasible; requiring mutagenesis and functional studies not only on a single representative sequence, but also over the biological sequence diversity represented by the protein family. While these biological datasets are unavoidably 'noisy' for testing the efficacy of functional divergence predictors for these reasons, their true positive sites are often well supported with robust experimental validation. We evaluated the performance of FunDi, and several other FD classifiers over 11 of these biological datasets with two phylogenetically distinct subfamilies each (Chakrabarti *et al.*, 2007). In order to provide a more robust estimate of performance on less noisy data, we also introduce two alternative frameworks for simulating functional divergence. In this framework, we examine the impact of taxon sampling and the scale of branch lengths on the predictive performance of functional divergence classifiers, because undersampling of phylogenetic diversity (taxon sampling) and overall sequence divergence are two well-known factors influencing the accuracy and error associated with phylogenetic reconstructions (Susko *et al.*, 2005; Zwickl and Hillis, 2002). In the case of functional divergence, it seems likely that undersampling of meaningful phylogenetic diversity can lead to incorrect observations of substitution patterns and sequence conservation levels (Blouin *et al.*, 2005). By explicitly taking the phylogeny of protein families into account with an appropriate model of functional divergence, we expect improved predictive performance relative to programs that do not use this information.

2 METHODS

2.1 FunDi

We assume that a given multiple sequence alignment is composed of sites that fall into two classes, those contributing to functional divergence and non-divergent sites. To capture the dynamics of the FD class, we construct a two-component phylogenetic mixture model, where non-FD sites evolve across a shared phylogenetic tree (standard evolutionary model/dependent component), whereas FD sites are treated as being evolutionarily 'uncoupled', evolve on independent subtrees (FD component).

Specifically, the dependent component models amino acid residues whose evolutionary constraints remain similar across a single phylogenetic tree. This is captured by a standard substitution model of protein evolution such as JTT (Jones *et al.*, 1992), WAG (Whelan and Goldman, 2001) or LG (Le and Gascuel, 2008) with rates across sites (RAS) modeled using a discrete rate approximation to the gamma distribution.

During functional divergence, this 'standard' model of evolution is violated. Under type I functional divergence, a rate shift has occurred (heterotachy) such that a site can no longer be adequately modeled by the same rate category in different lineages of the tree. A similar argument can be made for type II functional divergence where a site has undergone a shift in the amino acid preferences across a phylogenetic tree. In both cases, the normal assumption of a homogeneous substitution process across lineages no longer holds. In order to capture functional divergence, we introduce an 'independent component' approximation where sites in subtrees are modeled as if they were completely independent observations. In the maximum likelihood (ML) framework, model parameters such as the alpha shape parameter, amino acid frequencies and branch lengths are allowed to be independently optimized in each subgroup. The total likelihood of a site under this simplified approximate FD model will therefore be the product of the site likelihoods for each subgroup. Note that this is equivalent to assuming that the length of the internal branchlength between the subtrees, b, is effectively infinite and approximates the period of rapid evolution that immediately follows the changes in functional constraints at a site associated with FD. For two subgroups, the likelihood of a site x under this independence model is given by:

$$L_x = P(X_1|T_1)P(X_2|T_2) \quad (1)$$

T_1 and T_2 are the phylogenies and associated branch lengths for each of the two subtrees while X_1 and X_2 are the data patterns in the two subgroups at that site. In a mixture model context, the likelihood of a site is given as the weighted sum of the dependent and FD components:

$$L_x = \rho P(X_1, X_2|T, b) + (1-\rho)(P(X_1|T_1)P(X_2|T_2)) \quad (2)$$

where ρ represents the optimized class weight parameter and T refers to the entire phylogenetic tree comprised of T_1 and T_2 linked by an internal branch of length b. The site likelihoods of each component are calculated by standard ML phylogenetic estimation software using a supplied tree and alignment along with subgroup assignments for taxa. The ρ parameter is optimized using a two-step grid search procedure to two decimal places of precision.

While the observed FD site patterns in the two subtrees are not expected to be completely independent (i.e. they retain a shared evolutionary history/trajectory), approximating them as independent offers several advantages. First, it allows for maximum flexibility of ML model choice. Any phylogenetic software tool that outputs site likelihoods can be used as the back-end engine for likelihood calculations. Currently, FunDi can

accept site log-likelihood values from PUZZLE (Schmidt *et al.*, 2002), RAxML version 7.2.6 (Stamatakis, 2006), QmmRAxML (Wang *et al.*, 2008) or FastTree (Price *et al.*, 2009, 2010). This 'Plug-And-Play' utility allows FunDi to rapidly accommodate new, complex models of sequence evolution as they are developed. In addition, by implementing FunDi as a mixture model containing both independent and dependent components, the shared evolutionary history of FD sites is not completely ignored. All sites will be modeled with likelihood contributions from both components. It is the relative contribution of the independent component, measured by the site-wise posterior probability of the FD class, that serves as an estimator of the FD character for a given site.

Here, the performance of FunDi using either the 'base' RAxML v.7.2.6 (called FunDi-RAxML) or QmmRAxML (called FunDi-QmmRAxML) is evaluated. In brief, QmmRAxML is a mixture model of a user-defined number of rate matrix classes (Q_i's) each with an associated weight (w_i) that is optimized by ML. Here for each class i, we define entries of an instantaneous rate matrix $Q_{jk}(i) = R_{jk}\Pi_k(i)$ for all pairwise combinations of amino acids k and j. R_{jk} is the standard amino acid exchangeability of amino acid j for amino acid k from an exchangeability matrix (WAG in this case) and the Π_i's represent nine commonly occurring amino acid frequency profiles estimated by Sjölander *et al.* (1996). The WAG database frequencies form a 10th, catch-all, class. $\Pi_k(i)$ is therefore the frequency of the amino acid k in the frequency profile class i. Under the FD model, this model has the advantage that each subgroup can optimize toward different class (profile) preferences and rates, allowing for functional shifts at particular sites across the split.

FunDi outputs the posterior probability of each site belonging to the FD, class. FunDi can optionally be run with the ConsWin windowing method (FunDi-ConsWin) as described in Capra and Singh (2008), which weights site scores based on the Jensen–Shannon Divergence of surrounding amino acid residues. The Jensen–Shannon divergence score for all sites is calculated using a python script as detailed by Capra and Singh (2007). The average Jensen–Shannon divergence score of a window of surrounding columns in the alignment is then weighted and added to the posterior probability of functional divergence:

$$S(FD)_x = \lambda P(FD)_x + (1-\lambda)JSD_{avg} \quad (3)$$

where $S(FD)_x$ is the functional divergence score at site x, λ is the weight for the posterior probability of functional divergence (P(FD)) at site x, and JSD_{avg} is the average Jensen–Shannon Divergence score of the window. Here we used the recommended optimal values (Capra and Singh, 2007) of 0.7 for λ and a window size of 3 to either side of the column under consideration. This sliding-window scheme has been shown to improve predictive performance both in the GroupSim method and with other classifiers (Capra and Singh, 2008).

2.2 Simulations

We have implemented two simulation strategies for functional divergence in order to evaluate the relative performance of various FD classifiers. Alignments containing both FD and non-divergent (non-FD) sites were simulated over a variety of tree topologies in order to provide a comprehensive analysis of performance.

2.2.1 Strategy I: site-specific amino acid profiles INDELible (Fletcher and Yang, 2009) was used to simulate alignments consisting of both FD and non-divergent sites using the 10-component QmmRAxML mixture model described above. In order to conform to GroupSim assumptions, FD sites were required to be located within windows of non-FD sites in the primary amino acid sequence and all sites were selected, as described below, to have specific Jensen–Shannon divergence score distributions. Distributions for each site type were estimated from biological datasets that have been used in previous studies (Chakrabarti *et al.*, 2007), with five sets of distributions used. Jensen–Shannon divergence scores were calculated for all FD sites, sites located in a three residue windows on either side of an FD site, and all other sites separately. This was done for all the two-family alignments used in Chakrabarti *et al.* (2007). Four sets of the above estimates were used directly while a fifth set of score distributions was set to be of intermediate values compared with the other four. The divergence score distributions for Window and other non-FD sites in this set were equal to one another (Supplementary Table S1). One hundred random trees and corresponding alignments were simulated under each of these five sets.

Phylogenetic trees: Random phylogenetic trees were generated using INDELible with a birth–death (BD) process. Trees were randomly chosen to have between 10 and 50 taxa. Birth, death and mutation rates were randomly selected from a uniform distribution between 0 and 1 while the sampling parameter was constrained to a value of 1 [for details about the BD process see Yang and Rannala (1997)]. For each of the five sets of Jensen–Shannon divergence scores, one thousand individual trees were simulated and from these 100 pairs were randomly selected and joined by a midpoint rooted internal branch of length 1 expected substitution per site (0.5 on either side of root). This represented 500 protein family trees undergoing functional divergence across a central split.

Simulated alignments: Ten thousand non-FD sites were simulated from a random ancestral root sequence under each of the 10 mixture model components. Sites were then sampled from these sets randomly to construct both the non-FD windows around divergent sites and the remainder of non-FD sites in an alignment. To generate FD data, we simulated, from a shared ancestral sequence, over each subtree independently with a zero internal branchlength between the subtrees. Subtrees of non-FD sites were separated by an internal branchlength of 1. For all site types, four discrete Γ site-rate categories were used based on an α-shape parameter of 0.5.

Type I FD sites (i.e. rate-shifted sites) were simulated using the standard WAG model of evolution. A root sequence was sampled from the WAG model frequencies and sequences for each subtree were simulated separately, with the same root sequence to allow for independence of rates. The simulated pool of sites was then filtered to remove all columns where an identical evolutionary rate was randomly assigned by INDELible.

For Type II FD sites, nine pairs of mixture model components were selected from the mixture model such that amino acids with a high frequency in one component of the pair will have a low frequency in the other and vice versa. For each component pair, we simulated 10 000 sites. Root sequences were randomly sampled from each of the two amino acid distributions of the components used. As for Type I sites, an alignment was simulated for each subtree independently with the same ancestral sequence. To simulate the effect of selection for differing physico-chemical properties, each branch in the subtree was allowed to evolve according to the proportional model (i.e. rates of interchange are proportional to the frequency of the target amino acid, similar to the CAT-Poisson model of Lartillot and Philippe, 2004) using pairs of the 10 component amino acid profiles selected as described above. The resulting 90 000 simulated sites were combined into a single pool of type II FD sites. To accentuate the differences between subtrees, type I and type II simulated datasets were then filtered to remove any columns where the most prevalent amino acid in one subtree represented 30% or more of sites in its counterpart. Alignment columns were also sampled from the type I and type II pools to have appropriate Jensen–Shannon divergence score distributions comparable to one of five biological datasets. To accomplish this, all site types (Type I FD, Type II FD, non-FD) were simulated in excess as described above and a subset was sampled so that proportions of Jensen–Shannon divergence scores in a given subset roughly matched those of the biological datasets. The final alignments used were 400 residues in length, 40 of which were FD sites (20 Type I and 20 Type II). Each FD site was given a window of three non-FD sites to either side in the final alignment to conform to the assumptions made by the GroupSim ConsWin (Capra and Singh, 2008) method as described above.

Taxon sampling: To test the impact of taxon sampling on FD prediction, two phylogenetic trees were chosen to represent best and worst case examples

based on the performance of FunDi relative to other classifiers. Taxa were randomly re-sampled from these datasets in groups of 10, 15, 20, 25, 30 and 35 with the only constraint being that a minimum of four taxa were present in each subgroup. Ten replicate samplings were conducted for each number of taxa. A phylogenetic tree was then re-estimated from the data using RAxML version 7.2.6 (Stamatakis, 2006) and predictions of functional divergence made with each of the tested prediction methods.

Branch length scaling: The two tree topologies discussed above were again used as best and worst cases to investigate the impact of branch lengths on predictive performance of FD detection. For each of the two trees, the branches in the subtrees were re-scaled by a factor of 0.5, 1.5, 2, 3, 4, 5 or 10, or the internal branch separating the two subtrees was set to a length of 1.5, 3, 5 or 10. A simulated dataset was generated as described above on this new phylogenetic tree and evaluated using each of the chosen prediction methods.

2.2.2 Strategy II: defined motifs The 'evolutionary motif' method in Indel-Seq-Gen version 2 (Strope *et al.*, 2007; 2009) was also used. In brief, the sequence motifs of FD subsites from select datasets used in prior performance evaluations (Chakrabarti *et al.*, 2007; Chakrabarti and Panchenko, 2009) were compiled for each of the two subgroups in a given family (Supplementary Table S2).

FD sites were constrained to the motifs found in the biological datasets selected as recommended by Strope *et al.* (2009). An ancestral character state for FD sites in both subtrees was randomly selected and evolved according to the differing motifs of the subtrees. As before, window sites were constructed surrounding each FD residue, but in this case were constrained to be 100% conserved in order to provide optimal conditions for the ConsWin windowing method and GroupSim. Non-FD sites for the remainder of the sequence length were simulated with INDELible using the 10-component amino acid profile mixture with WAG exchangeabilities, 4 Γ rate categories and an α shape parameter of 0.5 using the original tree from the protein family. Non-FD sites were simulated under each of these 10 components, 25 sites per component for 250 non-FD positions unconstrained in their conservation level. For each of the selected biological datasets, 10 independent simulations were performed, to yield 70 simulated alignments over 7 different phylogenetic trees. This simulation strategy allows true FD sites to be simulated based on known biologically derived parameters/motifs while removing the serious problem, in biological datasets, of undetected positives from being incorrectly labeled as negatives (false negatives).

2.3 Testing divergence

For all programs, where appropriate, default values and raw scores were used to produce ordered lists with sites labeled 0 or 1 according to whether they were a truly FD (1) versus a non-FD (0) site. When necessary, raw scores of programs were rescaled to be between 0 and 1, with high scores being indicative of functional divergence (Supplementary Materials). To evaluate overall performance receiver operator characteristic (ROC) and precision-recall (PR) curves, as well as the total area under the curve for both curves (AUC-PR and AUC-ROC), values were calculated using AUCCalculator 0.2 (Davis and Goadrich, 2006). The AUC values each yield a single relative performance score for evaluation of the overall classification performance; the greater the AUC, the better the predictor averaged over all thresholds. An AUC value = 1 indicates perfect performance according to the criterion. We also used the 'average ranks' evaluation method that averages the rank of all true positive sites (ordered by decreasing FD score) in a tested dataset or series of datasets (in this case all 500 or 70 datasets for a simulation method). The lower the average rank, the better the performance of the method. AUC values and calculated average ranks were then used to generate boxplots using the R statistical package. All programs were evaluated over the larger 500 alignment and tree set (simulation set 1) and the smaller 70 alignment set with motifs (simulation set 2) as well as a set of 11 biological datasets (Chakrabarti *et al.*, 2007).

3 RESULTS

We investigated and compared the performance of the FunDi methods and several other methods for FD site prediction over a range of tree topologies and sizes, identifying particular tree topologies that may prove problematic for prediction of functional divergence. The impact of taxon sampling (recovery of true molecular diversity), length of the branch separating the subtrees and overall tree length was also investigated in order to build a robust picture of the behavior of the various functional divergence prediction algorithms and their performance over phylogenetically diverse data.

Several programs were selected based on their performance in previous studies (Brandt *et al.*, 2010; Capra and Singh, 2008; Chakrabarti and Panchenko, 2009) as well as their ability to be used in a large-scale testing pipeline. We tested the performance of our own method, FunDi (using both QmmRAxML and RAxML for site log-likelihood calculation), FunDi+ConsWin, SPEER (Chakrabarti *et al.*, 2007), GroupSim (Capra and Singh, 2008), Sequence Harmony (Feenstra *et al.*, 2007) and Multi-RELIEF (Ye *et al.*, 2008). Both Sequence Harmony and Multi-RELIEF were used as implemented in Multi-Harmony (Brandt *et al.*, 2010). These programs represent the top-performing methods as determined by prior studies and include both phylogenetic and information-theoretic approaches. A likelihood ratio test method for FD detection was also evaluated (Knudsen and Miyamoto, 2001; Knudsen *et al.*, 2003) on both the simulated and real biological datasets. However, as it had poor performance in initial tests we did not do a complete set of analyses (see Supplementary Materials).

Eleven two subfamily biological datasets were selected from Chakrabarti *et al.* (2007) and performance was evaluated. The 11 datasets were selected with the requirement that each subfamily had to be phylogenetically distinct and contain a minimum of four taxa per subfamily. The datasets selected feature a broad range in terms of number of taxa and number of FD sites. The performance of most classifiers as measured by AUC-ROC was very similar, with more variation seen in the AUC-PR metrics (Supplementary Figure S1). Using the medians of the AUC-PR distribution to judge the overall performance, GroupSim appeared to have the overall best performance with FunDi-ConsWin and Multi-RELIEF as the next best performers in the AUC-PR plots. Median performance as measured by AUC-ROC is slightly higher on these 11 biological datasets than that observed under either of the two simulation conditions examined below, but not significantly, except in the case of SPEER in case of simulated dataset 2. We also compared the performance of the Real-Value Evolutionary Trace (Mihalek *et al.*, 2004) and Difference-ET methods (Madabushi *et al.*, 2004; Raviscioni *et al.*, 2006) here, but due to technical constraints were unable to perform those evaluations on our larger simulated datasets.

In simulation Set 1 across all 500 datasets, the performances of FunDi using either QmmRAxML or RAxML were highly similar, outperforming all other methods tested as measured by the area under the precision–recall curve (AUC-PR), the area under the ROC curve (AUC-ROC) and the average rank of true positive FD sites (Fig. 1). The program GroupSim (Capra and Singh, 2008) applies a simple windowing method (ConsWin) for adjusting

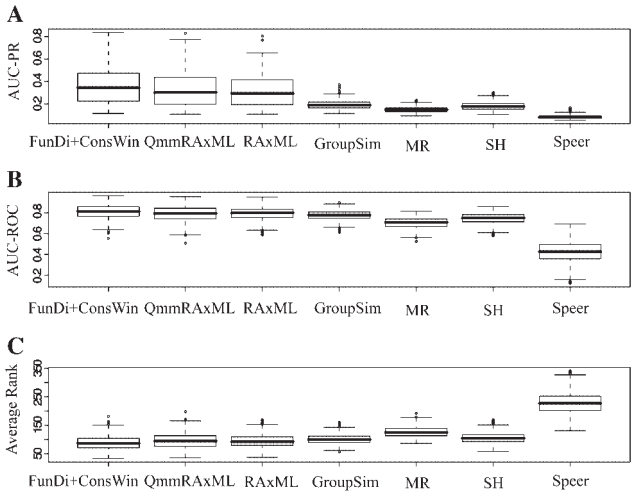

Fig. 1. Boxplots showing performance of several functional divergence classifiers on 500 simulated datasets as measured by the area under the precision–recall curve (**A**) and receiver operating characteristic (**B**) curve. Higher values reflect increased performance with a maximum value of 1.0. Additionally performance was characterized by the average rank of true positive functionally divergent sites (**C**) with sites ordered by the respective FD score of the program tested. All scores transformed (if required) to be between 0 and 1 with high scores reflecting a better functional divergence score. For average rank lower median values show increased performance. The methods evaluated in all cases are FunDi with QmmRAxML, FunDi with QmmRAxML and the ConsWin windowing method, FunDi with RAxML, GroupSim, Multi-Harmony (MR), Sequence-Harmony (SH) and SPEER. The 500 datasets were simulated with varying conditions over randomly generated tree topologies.

scores of functional divergence based on neighboring residues in the primary sequence (described previously). We applied this same method to the posterior probabilities of functional divergence P(FD) generated by FunDi with QmmRAxML to see if it yielded an improvement and to ensure that our simulation settings were appropriate for Group-Sim's prediction strategy. FunDi+ConsWin displayed an increase in predictive performance compared with the non-windowed P(FD) scores alone (Fig. 1). GroupSim was the next best classifier after the three FunDi-based methods across all datasets. Surprisingly given previous studies (Chakrabarti and Panchenko, 2009), SPEER appeared to have the lowest performance for all three scoring metrics. The distribution of AUC-PR values for FunDi was significantly different from the other predictors in all pairwise comparisons by the Wilcoxon signed-rank test with a Bonferroni correction for multiple comparisons ($P < 8.8e\text{-}16$). All other tests were also significantly different from one another with Bonferroni corrected $P \ll 0.05$.

3.1 The impact of phylogenetic tree shape and the number of taxa

We also investigated performance of the classifiers as a function of several common phylogenetic tree shape statistics. The clearest trend indicated that performance was greatly influenced by normalized tree length (i.e. overall sum of branch lengths divided by the number of taxa) as shown in Figure 2. All programs that we examined exhibited some increase in AUC-PR as the normalized tree length

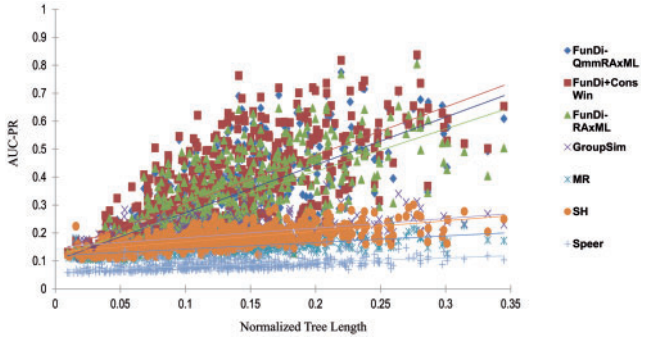

Fig. 2. Performance (area under the precision–recall curve) versus normalized tree length (total summed length of all branches in the phylogenetic tree divided by the number of taxa) for several functional divergence classifiers evaluated over 500 randomly simulated datasets. Linear trend lines for the data points are also shown for each classifier. Larger normalized tree lengths result in increased predictive performance, particularly for the three versions of FunDi tested here.

increased; however, this trend was much stronger in the three variations of FunDi tested. As the normalized tree length increases, the performance gap between FunDi and the other prediction programs increases, with the FunDi-based methods doing much better in general at longer tree lengths. We also identified two tree topologies with identical normalized tree lengths (NTL = 0.18) where the performance of FunDi differs significantly (Fig. 3). For the tree in Figure 3A, FunDi performs poorly (AUC-PR = 0.33), whereas in the tree in Figure 3B, it performs much better (AUC-PR = 0.51). Curiously, the AUC-PR results under the poorly performing tree are nearly indistinguishable from those of Group-Sim (second best performing method overall), while there is a large difference in AUC-PRs under the 'high-performance' tree. These two topologies differ mainly in their tree shapes, with a large discrepancy between the branch lengths in the subtrees. These two tree topologies are best and worst case examples for FunDi and were selected for further analyses on the effect of branch length and taxon sampling on functional divergence prediction. While large discrepancies between branch lengths of subtrees in datasets with functional divergence is not uncommon, in this case a large branch length discrepancy is compounded with relatively short branches throughout the tree when compared with similar trees from 11 biological datasets examined (Supplementary Table S3). This may explain why performance increases so dramatically when branches in the subtrees are made longer (see below).

3.2 The impact of taxon sampling

To test the effects of taxon sampling on functional divergence prediction, random taxon subsets were created from the simulated datasets from trees A and B in Figure 3. Results are shown in Figure 4. For both trees, addition of more data in the form of additional taxonomic coverage improved the performance of FunDi (AUC-PR; AUC-ROC shown in Supplementary Figure S2) relative to other tested methods, although the trend is much more pronounced for tree B. On the other hand, information-theoretic methods appear to be relatively insensitive to taxon sampling, showing only moderate performance increases (or some apparent decreases, e.g. Fig. 4B).

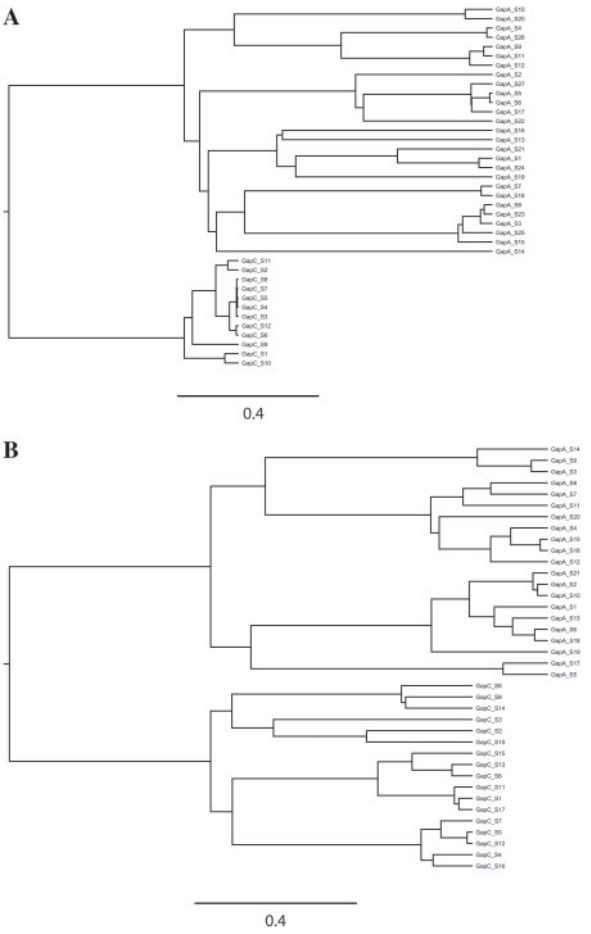

Fig. 3. Trees used for simulations that led to (**A**) poor performance of FunDi (relative to other classifiers) and (**B**) good performance. These trees were selected for further analyses of the impact of taxon sampling and branch length rescaling as best and worst case examples of phylogenetic tree shapes, balance and differences between subtrees. Both tree topologies have an identical normalized tree length of 0.18. Performance of FunDi on tree topology A (as measured by the area under the precision–recall curve) was 0.33, while it was 0.51 for tree topology B.

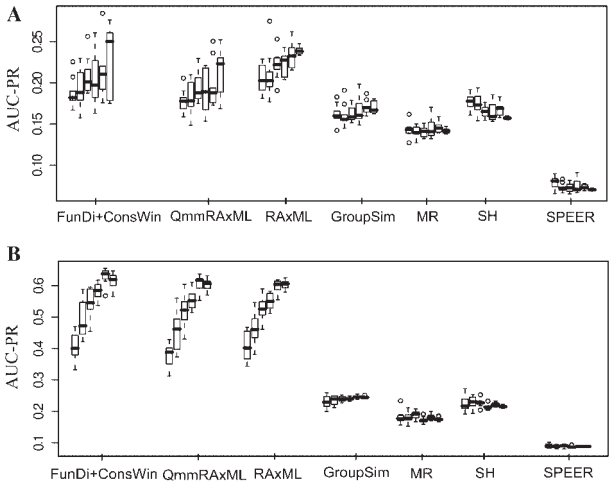

Fig. 4. Boxplots showing the impact of taxon sampling on performance as measured by the area under the precision–recall curve (**A** and **B**) on tree topologies A (A) and B (B) from Figure 3. For each tree 10 subsampled replicates of 10, 15, 20, 25, 30 or 35 taxa were constructed and the performances of each of the listed classifiers assessed.

3.3 The impact of branch length scaling

We also investigated the impact of branch lengths on performance using trees A and B as examples. We present only the results for AUC-PR as they showed the clearest trends. For both trees, either the branch lengths in the subtrees or the internal branch length separating the two subtrees were rescaled and for each case, a dataset was simulated and tested. The branch length effect is dependent on which branch is being rescaled (Fig. 5), as well as the given tree. Increasing the internal branch (Fig. 5A and B) separating subtrees results in a decrease in predictive performance for all classifiers for tree B. For tree A, it is difficult to discern a clear trend as the performances of most classifiers do not change dramatically, although at the longest branch length setting (10) all methods do generally poorer than at shorter lengths. This general effect is expected in terms of the performance of FunDi; as the internal branch between subgroups increases, the whole tree becomes closer and closer to the independence model, decreasing the distinction (and separatability) of the two components of the mixture model.

In the case of varying branch lengths in subtrees (Fig. 5C and D), AUC-PR clearly increases for all classifiers and on either tree as the branch lengths are increased. The trend is most dramatic for all three versions of FunDi relative to the other programs. As each subtree is evolving under its own evolutionary model, longer branches in subtrees provide additional time for substitutions that allow discrimination between the two site types (FD versus non-FD) to appear (if such a substitution did not occur along the internal branch) and provide more information upon which a classification can be made. This may be particularly true for Type I functional divergence as longer branches lead to more scrambling of the amino acid states at that site in the subtree with relaxed selective pressures. In tree A (the 'poorly' performing tree topology), we see rapid increases in performance for the FunDi-based methods as branch lengths increase.

3.4 Prediction of type I versus type II FD sites

We evaluated whether there were differences in the ability of methods to predict Type I versus Type II FD sites. While better performance is observed for Type II sites (Supplementary Figure S3), the difference is not great as median values for AUC-PR or AUC-ROC metrics for Type I or Type II sites fall within the other site type's interquartile range.

3.5 Performance with defined evolutionary motifs

The functional divergence prediction programs were tested using the same performance metrics on the 70 alignments from simulated dataset 2, which used the 'defined evolutionary motifs' simulation strategy. Overall, the same general trends in relative performance are observed. The median AUC-PR was highest for FunDi (RAxML, QmmRAxML and QmmRAxML + ConsWin) with GroupSim, the next best performing prediction method (Fig. 6A). The large range

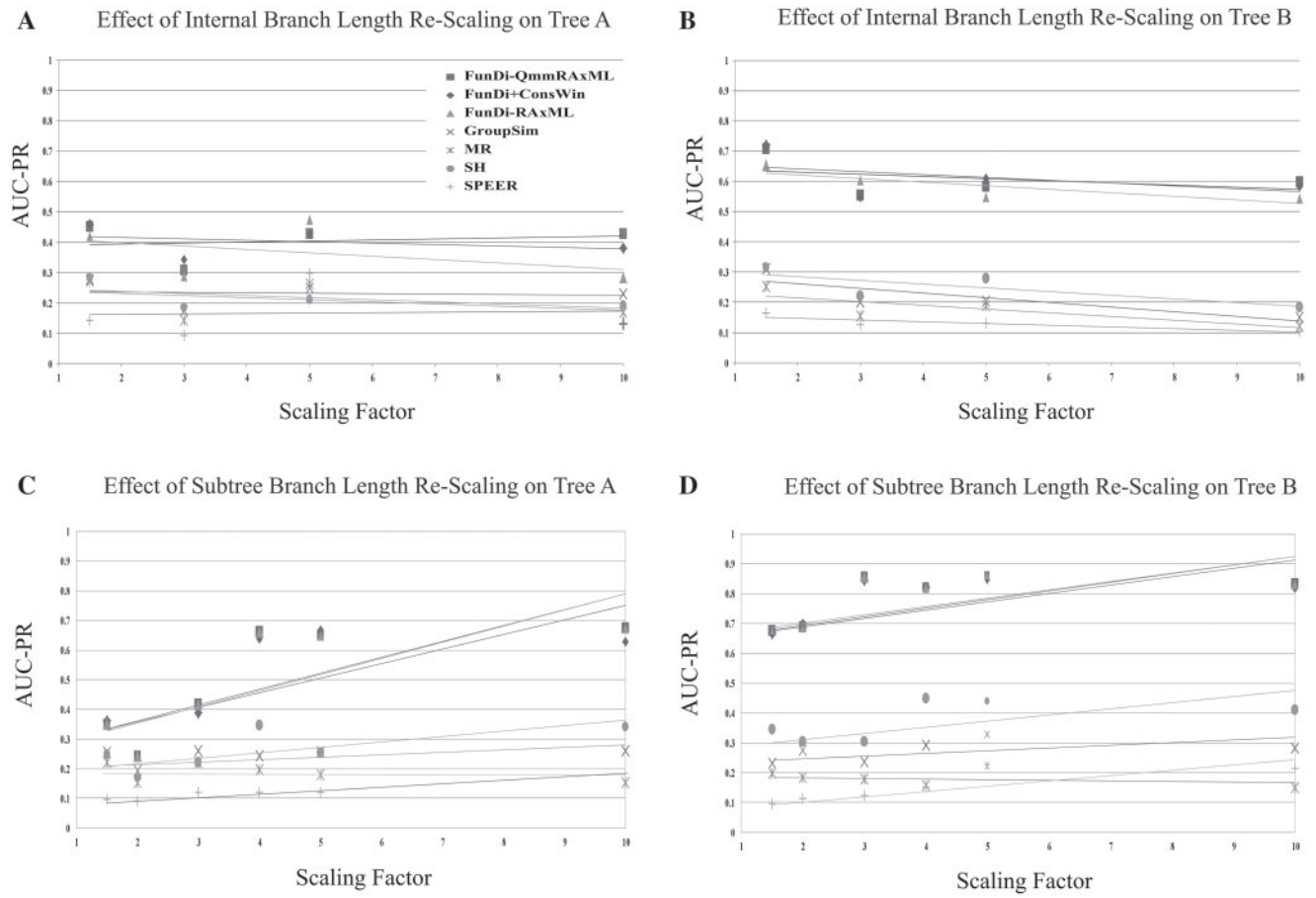

Fig. 5. Impact of branch length rescaling on internal (**A** and **B**) or subtree (**C** and **D**) branches as measured by the area under the precision–recall curve. Rescaling was applied to both tree topologies A and B from Figure 3. Scaling factors are shown on the *x*-axis with AUC-PR scores on the *y*-axis. For each of the two topologies rescaling was applied to the indicated branch length(s) and a random dataset simulated and performance of the tested functional divergence classifiers assessed as described in Section 2.

of performance scores observed with the AUC-PR data can best be explained by the varying performance on individual motifs and phylogenetic tree shapes used (data not shown). When evaluated using AUC-ROC, FunDi + ConsWin was the best performing prediction method, followed by GroupSim, then by FunDi (QmmRAxML or RAxML) without the windowing method (Fig. 6B). Performance of FunDi (AUC-PR values) is significantly different from other predictors as measured by the Wilcoxon signed-rank test ($P < 2.65e\text{-}12$). If non-FD windows that are more highly conserved compared with the majority of non-FD sites do tend to be located around FD sites, a windowing method such as ConsWin is of clear benefit, regardless of the testing methodology, as described previously (Capra and Singh, 2008).

4 DISCUSSION

Although not dramatic, there is an apparent discrepancy between the performance results for the 11 real datasets versus those of the two simulation studies. GroupSim does best overall for the real datasets with FunDi-based methods among the next best performers, whereas in the two kinds of simulations, FunDi-based methods (especially FunDi+ConsWin) typically outperform GroupSim and other methods. The source of the discrepancy is not very clear, but we would suggest that all the performance metrics are inherently less trustable for the 11 real datasets because, for these, only the true positive class of sites is known with certainty. As the relative rank of performance of the various methods is similar over two completely distinct functional divergence simulation settings and multiple performance indicators, we believe that the simulation results are more representative of the true performance properties of the various methods.

While FunDi performs better overall, its predictive power, as measured by precision–recall curves, remains relatively low (although performance as measured by ROC curves appears quite strong). This low predictive power is due to a variety of factors. Even under our simulation conditions, some functional shifts may result in relatively subtle amino acid substitutions, especially in complex situations that are an apparent mix of type I and type II functional divergence types, or what have been termed 'Marginally Conserved' sites. It is these marginally conserved sites that prove to be the most difficult in terms of prediction (Chakrabarti *et al.*, 2007). In addition, the majority of existing approaches essentially search

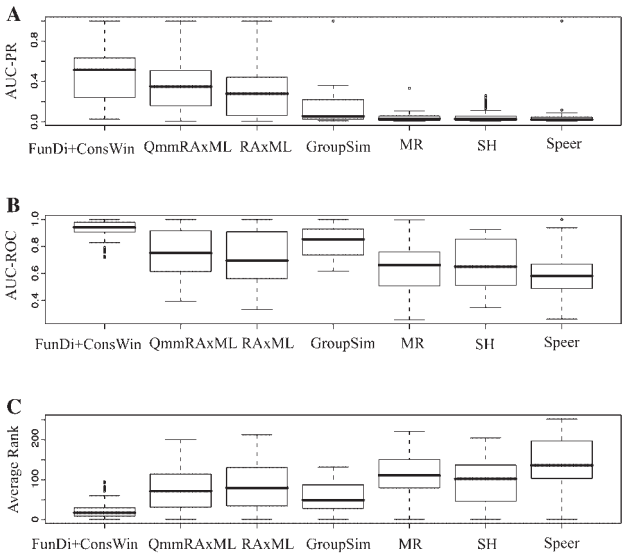

Fig. 6. Boxplots of performance of functional divergence classifiers on 70 datasets simulated with defined motifs using Indel-Seq-Gen-v2 as measured by area under the precision–recall (**A**) and receiver operator characteristic (**B**) curves as well as the average rank (**C**). The 70 simulated datasets are simulated with 10 replicates over each of 7 tree topologies (and with defined motifs) from real biological datasets from Chakrabarti *et al*. (2007 and 2009).

for particular patterns of amino acid usage, patterns which can arise in an evolutionary context due simply to stochastic neutral changes over the underlying phylogeny of the protein family without any functional shift occurring, making adjustments for the underlying phylogeny of great importance. FunDi can also leverage improved models of amino acid evolution, such as the 10 component amino acid profile mixture models implemented in QmmRAxML (Wang *et al.*, 2008) using WAG (used in this study), JTT, LG or user-supplied exchangeabilities. Judicious model selection allows the incorporation of some prior knowledge of the protein's evolutionary history, structure, function and amino acid frequencies.

We have also introduced two new simulation strategies for functional divergence that are useful for benchmarking new prediction programs, and improving existing ones, especially in phylogenetic 'trouble spots' (e.g. worst case tree used in our analyses of the impact of branch lengths and taxon sampling). Unfortunately, we were unable to compare our results with some of the other phylogenetically based methods such as DIVERGE (Gu, 1999; 2001; Gu and Vander Velden, 2002), because the software implementation and run times of the latter precluded analyses of large simulated datasets. Furthermore, the Gu 2001-based predictions could not be obtained for several of the biological datasets examined. Since FunDi is: (i) scalable to the analysis of multiple (potentially thousands) of large protein datasets; (ii) has a single coherent framework for the prediction of type I and type II FD sites; and (iii) can be used with any phylogenetic model of protein evolution implemented in a maximum likelihood framework, it has distinct advantages as compared with other phylogenetic-based functional divergence predictors currently in use.

Our analysis on a large, phylogenetically diverse set of simulated FD datasets shows that taking into account the phylogeny of a protein family is an important part of the prediction of FD sites. While non-phylogenetically aware prediction schemes such as GroupSim can be characterized as insensitive to issues of taxon sampling and phylogenetic tree topologies, they also do not increase in predictive accuracy under appropriate phylogenetic conditions and show generally poorer performance under a wide range of conditions. FunDi, as a phylogenetically aware prediction program, shows marked improvement in the quality of its predictions under increased taxon sampling (recovery of true biological diversity) as well as increased evolutionary time as measured by the normalized tree length and illustrated in our branch length rescaling experiments.

The main problem with using real biological data to evaluate the performance of functional divergence methods is the infeasibility of experimentally testing false and true negative prediction; there are simply too many sites and too many character state combinations to test comprehensively. The two simulation strategies described here therefore provide much cleaner data, with less noise than true biological data and so can be used to evaluate the performance of functional divergence methods over a wide range of possible evolutionary conditions such as tree topologies and taxon sampling. The ability to specify particular sequence motifs for FD residues based on observed biological data, as we have done here in the second set of simulations, may be useful in developing methods that have better performance on difficult-to-classify functionally divergent residues.

ACKNOWLEDGEMENTS

D.G. would like to thank William Fletcher for implementing several suggested changes to INDELible, to B.W. Brandt for providing a script for the Multi-Harmony web server that allowed testing of a large number of datasets and to Olivier Lichtarge and Angela Dawn Wilkins for running real value ET on the 11 biological datasets

Funding: Nova Scotia Health Research Foundation graduate student research award (to D.G.); Natural Sciences and Engineering Research Council of Canada, Discovery Grant (227085-2011 to A.J.R. and E.S.).

Conflict of Interest: none declared.

REFERENCES

Blouin,C. *et al.* (2005) Impact of taxon sampling on the estimation of rates of evolution at sites. *Mol. Biol. Evol.*, **22**, 784–791.

Brandt,B.W. *et al.* (2010) Multi-Harmony: detecting functional specificity from sequence alignment. *Nucleic Acids Res.*, **38**, W35–W40.

Caffrey,D.R. *et al.* (2008) Prediction of specificity-determining residues for small-molecule kinase inhibitors. *BMC Bioinformatics*, **9**, 49.

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

Capra,J.A. and Singh,M. (2008) Characterization and prediction of residues determining protein functional specificity. *Bioinformatics*, **24**, 1473–1480.

Capra,J.A *et al.* (2009) Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. *PLoS Comput. Biol.*, **5**, e1000585.

Chakrabarti,S. and Panchenko,A.R. (2009) Ensemble approach to predict specificity determinants: benchmarking and validation. *BMC Bioinformatics*, **10**, 207.

Chakrabarti,S. *et al.* (2007) Functional specificity lies within the properties and evolutionary changes of amino acids. *J. Mol. Biol.*, **373**, 801–810.

Davis,J. and Goadrich,D. (2006) The relationship between precision-recall and ROC curves. In *23rd International Conference on Machine Learning (ICML), Pittsburgh, PA, USA*. ACM, New York, NY.

de Melo-Minardi,R.C. *et al.* (2010) Identification of subfamily-specific sites based on active sites modeling and clustering. *Bioinformatics*, **26**, 3075–3082.

Feenstra,K.A. *et al.* (2007) Sequence harmony: detecting functional specificity from alignments. *Nucleic Acids Res.*, **35**, W495–W498.

Fletcher,W. and Yang,Z. (2009) INDELible: a flexible simulator of biological sequence evolution. *Mol. Biol. Evol.*, **26**, 1879–1888.

Gerlt,J.A. and Babbitt,P.C. (2000) Can sequence determine function? *Genome Biol.*, **1**, reviews0005.1–0005.10.

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

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

Gu,X. and Vander Velden,K. (2002) DIVERGE: phylogeny-based analysis for functional-structural divergence of a protein family. *Bioinformatics*, **18**, 500–501.

Henikoff,S. *et al*. (1997) Gene families: the taxonomy of protein paralogs and chimeras. *Science*, **278**, 609–614.

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. USA*, **98**, 14512–14517.

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

Kullback,S. and Leibler,R.A. (1951) On information and sufficiency. *Ann. Math. Stat.*, **22**, 79–86.

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

Le,S.Q. and Gascuel,O. (2008) An improved general amino acid replacement matrix. *Mol. Biol. Evol.*, **25**, 1307–1320.

Li,W.H. (1983) Evolution of duplicated genes. In Nei,M. and Koehn,R.K. (eds) *Evolution of Genes and Proteins*. Sinauer Associates, Sunderland, MA, pp. 14–37.

Lichtarge,O. *et al.* (1996) An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol*., **257**, 342–358.

Lin,J. (1991) Divergence measures based on the shannon entropy. *IEEE Trans. Informat. Theory*, **37**, 145–151.

Madabushi,S. *et al*. (2004) Evolutionary trace of G protein-coupled receptors reveals clusters of residues that determine global and class-specific functions. *J. Biol. Chem.*, **279**, 8126–8132.

Mihalek,I. *et al.* (2004) A family of evolution-entropy hybrid methods for ranking protein residues by importance. *J. Mol. Biol.*, **336**, 1265–1282.

Pawlowski,K. and Godzik,A. (2001) Surface map comparison: studying function diversity of homologous proteins. *J. Mol. Biol.*, **309**, 793–806.

Pirovano,W. *et al*. (2006) Sequence comparison by sequence harmony identifies subtype-specific functional sites. *Nucleic Acids Res.*, **34**, 6540–6548.

Price,M.N. *et al.* (2009) FastTree: computing large minimum-evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.*, **26**, 1641–1650.

Price,M.N. *et al.* (2010) FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS One*, **5**, e9490.

Raviscioni,M. *et al*. (2006) Evolutionary identification of a subtype specific functional site in the ligand binding domain of steroid receptors. *Bioinformatics*, **1057**, 1046–1057.

Sael,L. *et al*. (2008) Rapid comparison of properties on protein surface. *Proteins*, **73**, 1–10.

Sankararaman,S. *et al*. (2010) Active site prediction using evolutionary and structural information. *Bioinformatics*, **26**, 617–624.

Schmidt,H.A. *et al*. (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics*, **18**, 502–504.

Sjölander,K. *et al*. (1996) Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. *Comput. Appl. Biosci.*, **12**, 327–345.

Stamatakis,A. (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, **22**, 2688–2690.

Strope,C.L. *et al.* (2007) indel-Seq-Gen: a new protein family simulator incorporating domains, motifs, and indels. *Mol. Biol. Evol.*, **24**, 640–649.

Strope,C.L. *et al*. (2009) Biological sequence simulation for testing complex evolutionary hypotheses: indel-Seq-Gen version 2.0. *Mol. Biol. Evol.*, **26**, 2581–2593.

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

Susko,E. *et al.* (2005) Biases in phylogenetic estimation can be caused by random sequence segments. *J. Mol. Evol.*, **61**, 351–359.

Wang,H.C. *et al.* (2008) A class frequency mixture model that adjusts for site-specific amino acid frequencies and improves inference of protein phylogeny. *BMC Evol. Biol.*, **8**, 331.

Whelan,S. and Goldman,N. (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.*, **18**, 691–699.

Yang,Z. and Rannala,B. (1997) Bayesian phylogenetic inferences using DNA sequences: a Markov chain Monte Carlo method. *Mol. Biol. Evol.*, **14**, 717–724.

Ye,K. *et al.* (2008) Multi-RELIEF: a method to recognize specificity determining residues from multiple sequence alignments using a Machine-Learning approach for feature weighting. *Bioinformatics*, **24**, 18–25.

Zwickl,D.J. and Hillis,D.M. (2002) Increased taxon sampling greatly reduces phylogenetic error. *Syst. Biol.*, **51**, 588–589.