BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 2 2006, pages 164–171
doi:10.1093/bioinformatics/bti766

Sequence analysis

Prediction of functional specificity determinants from protein sequences using log-likelihood ratios

Jimin Pei[2], Wei Cai[2], Lisa N. Kinch[1] and Nick V. Grishin[1,2,*]

[1]Howard Hughes Medical Institute and [2]Department of Biochemistry, University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, TX 75390-9050, USA

Received on August 9, 2005; revised on and accepted on November 3, 2005
Advance Access publication November 8, 2005
Associate Editor: Dmitrij Frishman

ABSTRACT

Motivation: A number of methods have been developed to predict functional specificity determinants in protein families based on sequence information. Most of these methods rely on pre-defined functional subgroups. Manual subgroup definition is difficult because of the limited number of experimentally characterized subfamilies with differing specificity, while automatic subgroup partitioning using computational tools is a non-trivial task and does not always yield ideal results.

Results: We propose a new approach SPEL (specificity positions by evolutionary likelihood) to detect positions that are likely to be functional specificity determinants. SPEL, which does not require subgroup definition, takes a multiple sequence alignment of a protein family as the only input, and assigns a *P*-value to every position in the alignment. Positions with low *P*-values are likely to be important for functional specificity. An evolutionary tree is reconstructed during the calculation, and *P*-value estimation is based on a random model that involves evolutionary simulations. Evolutionary log-likelihood is chosen as a measure of amino acid distribution at a position. To illustrate the performance of the method, we carried out a detailed analysis of two protein families (LacI/PurR and G protein α subunit), and compared our method with two existing methods (evolutionary trace and mutual information based). All three methods were also compared on a set of protein families with known ligand-bound structures.

Availability: SPEL is freely available for non-commercial use. Its pre-compiled versions for several platforms and alignments used in this work are available at ftp://iole.swmed.edu/pub/SPEL/

Contact: grishin@chop.swmed.edu

Supplementary information: Supplementary materials are available at ftp:/iole.swmed.edu/pub/SPEL/

INTRODUCTION

Patterns of sequence conservation and variability within multiple sequence alignments (MSAs) of protein families reflect various structural and functional constraints required for biological function. Highly conserved positions are likely to play important general roles in a protein family (Aloy *et al*., 2001; Innis *et al*., 2004; Panchenko *et al*., 2004; Pei and Grishin, 2001; Pupko *et al*., 2002). Positions that are conserved within subsets of closely related proteins in a given family, but are variable between the subsets, are likely to be involved in more specific functions in different subgroups (Mirny and Shakhnovich, 1999), such as conferring substrate specificity of enzymes. Here, we are interested in detecting these functional specificity determinants.

A number of computational methods have been developed to identify functionally important sites within protein families (reviewed by Jones and Thornton, 2004). Information used by these methods is from sequence and phylogenetic patterns (Lichtarge *et al*., 1996b), residue physical properties (Elcock, 2001) and three-dimensional (3D) structural patterns (Shulman-Peleg *et al*., 2004). MSAs have been used to deduce functional specificity sites based on hierarchical analysis of residue conservation patterns (Livingstone and Barton, 1993). Evolutionary information has been frequently employed (Armon *et al*., 2001; Bielawski and Yang, 2004; La *et al*., 2005; Soyer and Goldstein, 2004). The Evolutionary Trace (ET) method (Lichtarge *et al*., 1996b) was designed to predict both invariant sites and class-specific sites based on different levels of 'evolutionary time cutoff' in a phylogenetic tree. Although no scores or statistical significance were assigned to the original version of ET predictions, their sensitivity could be gauged by manual adjustment of a cutoff value that defined subgroup partitioning. A later development of the ET method has combined evolutionary and entropic information to rank positions and gives estimates of statistical significance (Mihalek *et al*., 2004). Principle component analysis of sequence space (Casari *et al*., 1995) and relative entropy (Hannenhalli and Russell, 2000) have also been used to detect sites of specificity for functional subtypes. Another computational method developed by Mirny and Gelfand predicts specificity-determining residues using a similar hierarchical analysis of residue conservation (Mirny and Gelfand, 2002). This method assigns a mutual information-based score (MI) to positions and provides statistical significance estimates, but it requires an input alignment with pre-defined subgroups. MI and relative entropy were also used in a further development of methods in detecting specificity positions (Kalinina *et al*., 2004a,b).

Most existing methods for finding functional specificity determinants rely on pre-defined functional subgroups. Manual subgroup definition is often limited by the small number of experimentally characterized subfamilies with differing specificities, while automatic subgroup partitioning using computational tools is a non-trivial task that does not always give ideal results (Wicker *et al*., 2001). Subgroups can be defined by ortholog-paralog relationships (Mirny and Gelfand, 2002). However, it is also not easy to differentiate orthologs and paralogs for many protein families. A phylogenetic tree frequently does not reveal well-defined

*To whom correspondence should be addressed.

subfamilies or orthologous groups. Moreover, different positions with specificity can display different residue association with tree structure. These limitations call for an effective approach that does not require partitioning of the target set of homologous sequences into subgroups. Here, we present such an approach, SPEL (specificity positions by evolutionary likelihood), to predict specificity positions in a given MSA that aims to improve upon existing methods and requires only an MSA as an input. To every position in the MSA, SPEL assigns a statistical significance score to reflect its predicted importance for functional specificity. To compute the score, a phylogenetic tree is automatically reconstructed from the MSA, and *P*-values of an evolutionary likelihood-based score for positions are estimated from a random model that eliminates any functional specificity signal.

We tested the performance of SPEL on two well-studied protein families (LacI/PurR and G protein α subunit). Our results were compared with those of two other methods (MI and ET). In addition, we carried out a larger-scale test of these three methods (SPEL, MI and ET) on two sets of proteins with bound ligands, which have been used in a previous study of functional site prediction (Yao *et al*., 2003). Closeness to the ligands was used as the criterion to judge the performance of these methods. SPEL performed similarly to MI in this large-scale test. We also studied the robustness of these methods with regard to alignment quality and concluded that all methods are reasonably robust, but more accurate alignments yield better predictions.

MATERIALS AND METHODS

Description of the SPEL method

Our method consists of three steps. First, a phylogenetic tree is inferred from a given MSA. Second, an evolutionary likelihood-based score [log-likelihood ratio (LLR)] is calculated for each position in MSA using the tree and a shuffling procedure. Third, using a random model to describe a position expected to be unconstrained by functional specificity, the *P*-value of the likelihood-based score for the observed amino acid distribution at any position is estimated. The flowchart of SPEL process is shown in Figure 1.

Phylogenetic tree reconstruction A phylogenetic tree is built from the input MSA using our implementation of a distance-based method (Cai *et al*., 2004). The evolutionary distances between sequences are estimated by the maximum likelihood approach in which substitution rate variation among positions is taken into account. In the amino acid substitution model used to estimate evolutionary distances, the amino acid exchangeability parameters are taken from the WAG matrix (Whelan and Goldman, 2001). The tree is reconstructed from the distance matrix using Weighbor (Bruno *et al*., 2000), which is an improvement of a neighbor-joining method (Saitou and Nei, 1987) by accounting for inaccuracies of large distances.

Log-likelihood ratio calculation For a position *i*, we calculate the log-likelihood of the amino acids conditioned on the tree ($\log L_i$). The likelihood function is the probability of observing a set of leaf node amino acids given an evolutionary tree and a specific amino acid substitution model. The general assumptions include independence of the positions and a time-reversible, homogeneous and stationary Markov process. We allow the evolutionary rate among positions to vary and estimate position-specific rate factors. The general model and the method for log-likelihood calculation have been thoroughly described in our previous paper on reconstruction of ancestral sequences (Cai *et al*., 2004).

To calculate the LLR, we shuffle the amino acids in the position and map the shuffled set of amino acids onto the same tree. We calculate the log-likelihood for the shuffled set of amino acids conditioned on the original tree

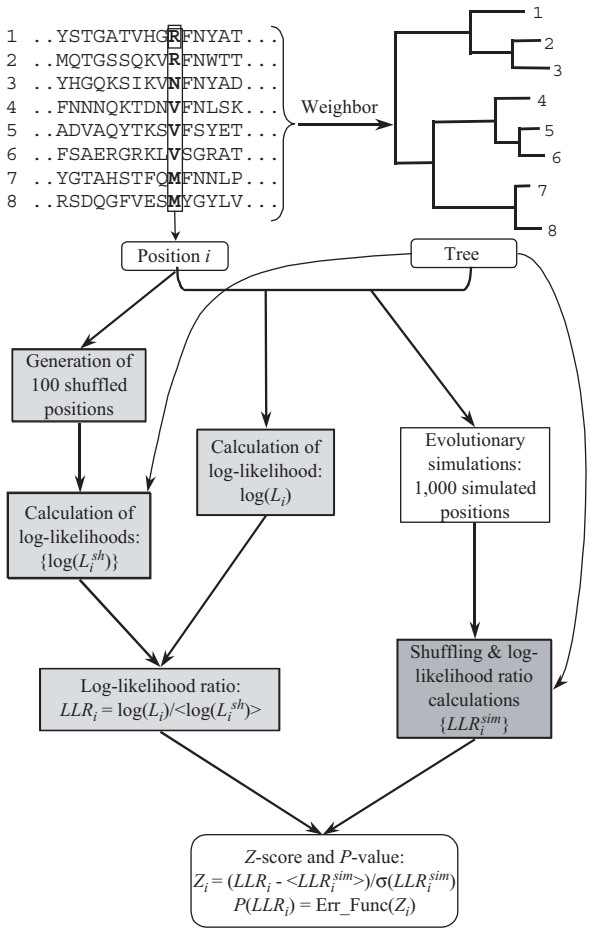

Fig. 1. Flowchart showing the process of SPEL. For every simulated position, the process of shuffling and LLR calculations (darker gray box on the right) has the same steps as shown in the lighter gray boxes on the left. The braces ({ }) are used for a sample of values and the angular brackets (< >) are used for the average over the sample. $\sigma(\mathrm{LLR}_i^{\mathrm{sim}})$ is the variance of the likelihoods of simulated positions for position *i*. The Err_Func is in the formula of Equation (3) in the Materials and methods section.

($\log L_i^{\mathrm{sh}}$). The process of shuffling and log-likelihood calculation is repeated 100 times. The LLR of the original amino acids at position *i* is calculated using the following formula:

$$\mathrm{LLR}_i = \log L_i / <\log L_i^{\mathrm{sh}}>, \quad (1)$$

where $<\log L_i^{\mathrm{sh}}>$ is the average log-likelihood of 100 amino acid sets resulted from independent shuffling at position *i*.

Random model: protein evolutionary simulation and shuffling To make any objective scoring function comparable between positions with different amino acid compositions and to estimate statistical significance of specificity predictions, an appropriate random model is required. The best random generator of a position should model that position as close as possible to match the observed data (sequence composition and phylogenetic patterns), but should exclude the influence of functional selection on that position. The difference between the observed data and randomly generated data would thus be limited to functional constraints that are present in the observed data, but are excluded in the random data. A reasonable random model should take into account amino acid usage at a modeled position and ideally, also the evolutionary tree.

To estimate the statistical significance (*P*-value) of LLR for the observed amino acid distribution, evolutionary simulations are used to get an ensemble of random LLR as described below.

For position i in the alignment, we simulate the evolutionary processes starting from the root of the tree to generate a set of simulated amino acids in the leaf nodes. To do that, we use a least-squares modification of the mid-point rooting procedure to define the root for the tree (Wolf *et al*., 1999). Starting from the root of the tree, we calculate the probability of observing any amino acid at its two child nodes, and randomly select an amino acid for each child node based on these probabilities. The process of probability calculation and that of random amino acid selection are repeated for the rest of the internal nodes until we generate a set of simulated amino acids for all the leaf nodes. This simulation process follows a general amino acids substitution model as defined by the WAG matrix (Whelan and Goldman, 2001), and thus does not introduce any functional specificity information. The resulting set of simulated amino acids should reflect the tree structure without any functional specificity constraints. We then calculate the LLR for the simulated amino acid set ($\mathrm{LLR}_i^{\mathrm{sim}}$) using the same shuffling procedure of getting LLR_i for the original set of amino acids. Such an evolutionary simulation process and the calculation of LLR are repeated 1000 times to get a sample of $\mathrm{LLR}_i^{\mathrm{sim}}$. *Z*-score ($Z_i$) and *P*-value of the LLR_i [$P(\mathrm{LLR}_i)$] for the original set of amino acids are calculated using the mean and the standard deviation of the 1000 simulated $\mathrm{LLR}^{\mathrm{sim}}$ values, with the assumption of a normal distribution of $\mathrm{LLR}^{\mathrm{sim}}$ values.

$$Z_i = (\mathrm{LLR}_i - \langle \mathrm{LLR}_i^{\mathrm{sim}} \rangle)/\sigma(\mathrm{LLR}_i^{\mathrm{sim}}) \quad (2)$$

$$P(\mathrm{LLR}_i) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Z_i} \exp\left(-\frac{z^2}{2}\right) \mathrm{d}z, \quad (3)$$

where $\langle \mathrm{LLR}_i^{\mathrm{sim}} \rangle$ and $\sigma(\mathrm{LLR}_i^{\mathrm{sim}})$ are the mean and standard deviation of simulated LLRs at position i, respectively.

Tests to compare prediction methods

We analyzed in detail the prediction results of SPEL, MI and ET on two well-studied protein families: LacI/PurR (Mirny and Gelfand, 2002) and G protein α subunit (G_α subunit) (Lichtarge *et al*., 1996a; Sowa *et al*., 2000). The LacI/PurR family alignment and subgroup definition were obtained from the authors of the MI method. To predict specificity determining residues for the family of G_α subunits, we generated an MSA of all BLAST-detected members using PCMA (Pei *et al*., 2003) with manual adjustments based on secondary structure and patterns of hydrophobicity. Subgroup definition for MI was defined manually according to the classical subtypes ($G_{\alpha s}$, $G_{\alpha q}$, $G_{\alpha i}$, $G_{\alpha t}$, $G_{\alpha o}$, G_{12}, etc.), sequence similarities and species distribution. The MI program was obtained from Mirny and Gelfand. All ET tests were performed using the web server at http://www-cryst.bioc.cam.ac.uk/~jiye/evoltrace/evoltrace.html, without inputting any structural information. Only class-specific predictions were counted for ET (invariant positions were not considered as predictions).

For a larger-scale test of the methods, we selected two sets of protein families with known structures that have been used previously to test the performance of ET (Yao *et al*., 2003). The smaller set (SGI) comprising 20 proteins were from structural genomics initiatives, while the larger 'protein-ligand' set has 37 protein–ligand complexes (structure IDS available from http://imgen.bcm.tmc.edu/molgen/labs/lichtarge/trace_of_the_week/current.html). Owing to lack of specificity information for every position of the structures, we applied the same criterion as in the previous ET test (Yao *et al*., 2003), namely, residues within 5 Å of any ligand are defined as true positives. An MSA was automatically generated for each protein structure by searching homologs from current non-redundant protein sequence database. We ran PSI-BLAST (Altschul *et al*., 1997) starting from the sequence with known structure for five iterations with an *E*-value cutoff 0.0001. The resulting alignment was processed to remove highly similar sequences (>95% identity), highly divergent sequences (<15% identity to the query) and sequence fragments (sequence length less than half of the query length). To get a representative set of sequences and the subgroup information for the MI method, we clustered the resulting sequences using single-linkage clustering program BLASTCLUST (ftp://ftp.ncbi.nih.gov/blast/documents/blastclust.html) at 45% sequence identity with length coverage on both neighbors set to 0.8. It has been shown that functional specificity changes are rare for sequences with more than 40% sequence identity (Todd *et al*., 2001; Wilson *et al*., 2000). Other clustering strategies have been explored (single-linkage or mean linkage clustering) and the results can be found in supplementary materials. We selected up to 3, 6 or 10 sequences from the top 4 or 8 largest clusters. Each cluster was used as a subgroup for the MI method. For SPEL and MI, the top 10 positions with the lowest *P*-values were selected as the final predictions. We counted the number of correct predictions (residues within 5 Å of any ligand) in the top 10 predictions. We also calculated the receiver operating characteristic (ROC) scores for each family. For the top n false positives, $\mathrm{ROC}_n = (1/nT)\Sigma_{i=1}^{n} t_i$, where t_i is the number of true positives that were ranked ahead of the i-th false positive in the list, and T is the total number of true positives in the dataset (Schaffer *et al*., 2001; Theodoridis, 1999). ROC_{10} (number of false positives up to 10) and ROC_0.1 (number of false positives up to 0.1 fraction of total possible false positives) were reported. Since ET also selects highly conserved positions (not contributing to specificity) as predictions, we ignored all highly conserved positions (having one dominant amino acid with frequency larger than 0.75) when counting top 10 predictions and calculating ROC scores for SPEL, MI and ET. Highly gapped positions (gap fraction larger than 0.5) were also excluded. Amino acid frequency and gap fraction calculations were performed using the program AL2CO (Pei and Grishin, 2001). For the ET method, cutting at different similarity levels usually produces discrete numbers of predictions (not necessarily incrementing by one, e.g. if one level cut produces 5 predicted positions, the next level cut might produce 13 predicted positions). If at one level, exactly 10 predicted class-specific residues were produced by ET, we just used these 10 residues as ET predictions and selected these residues within 5 Å to ligands as correct ET predictions. Otherwise, the level producing residue number just above 10 and the level producing residue number just below 10 were selected. The expected number of correct predictions for an expected top 10 predictions was a weighted average of the number of correct predictions of these two levels. As an illustration, if at one level 7 class-specific residues are produced and 3 among them are correct predictions, and at the next level 15 residues are produced and 8 among them are correct, the expected number of correct predictions assuming the prediction number is 10 will be $3 + (10-7)/(15-7) * (8-3) = 4.875$. The ROC scores for ET predictions were also calculated with the same considerations. The Wilcoxon signed rank test (Wilcoxon, 1947) was applied to compare the performance of SPEL, MI and ET.

RESULTS

A random model based on protein evolutionary simulation and shuffling

A random position can be generated by either shuffling the original position or by the evolutionary simulation using the tree (Mirny and Gelfand, 2002). However, tested on two artificially generated alignments with different properties, such two random models ('shuffling position' based and 'simulation' based) with log-likelihood itself as the scoring function showed some intrinsic problems, such as including positions without strong association of specificity or assigning unreasonable *P*-values (see supplementary materials for details). In SPEL, the random model is a combination of protein evolutionary simulation and shuffling. It was shown to give desirable results on the two artificial alignments (supplementary materials) in terms of giving reasonable *P*-values and being able to rank specificity positions correctly.

Studies on LacI/PurR family and G protein α subunit

We made detailed analysis of the performance of SPEL, MI and ET on two well-studied protein families that have been used for previous prediction of functional specificity: LacI/PurR (Mirny and Gelfand, 2002) and G protein α subunit (Lichtarge *et al*., 1996a; Sowa *et al*., 2000).

LacI/PurR family

The LacI/PurR family represents a large family of bacterial transcription factors that are regulated by small molecules, such as sugars and nucleotides (Sauer, 1996). In addition to available experimental and structural information, the LacI/PurR family orthology has been resolved using positional information from bacterial genomes (Mirny and Gelfand, 2002). This resolution allows a comparison of SPEL performance with that of two established methods: MI, which requires defined groups, and ET. The top ten functional positions with specificity predicted by SPEL and MI (Table 1) include six identical positions (green CPK, Fig. 2A), of which three residues in contact with DNA (T15, T16 and K55) are recognized as specificity determinants by mutagenesis experiments (Glasfeld *et al*., 1999; Lehming *et al*., 1990; Sartorius *et al*., 1991). The other two common residues (D146 and D160) are near the ligand-binding pocket. Of the positions found confidently by SPEL that were not among the top MI predictions (pink CPK in Fig. 2A); two (K5 and H20) bind DNA, one (C123) lines the ligand-binding pocket, and one resides near the dimer interface (C85). Positions predicted by MI and not by SPEL line the ligand-binding pocket (F221, W98, I249 and M122) or contribute to the dimer interface (K114).

While the results of SPEL and MI remain consistent, the ET method predicts a different set of class-specific positions (Table 1). Because ET does not provide confidence scores, the data were partitioned to provide as close to ten predictions as possible (level 10, 12 subtrees with more than one sequence, 9 class-specific predictions). Positions identified by ET at this level display a high degree of conservation without being invariant across the entire sequence set, and therefore do not overlap with top predictions by either MI (none common) or SPEL (only C85 is common) methods. This discrepancy is caused by the definition of class-specific predictions of ET. At a specific similarity level, ET partitions the tree into subtrees and selects any position that is invariant within every subtree as a prediction for this level. Class-specific predictions are defined as those positions with different amino acid types for at least two subtrees. This definition will include positions that have the same amino acid type for most of the subtrees (e.g. 10 out of 12) and different amino acid type(s) for the rest of the few subtrees (e.g. 2 out of 12). Of the 9 class-specific predictions given by ET, 5 positions are highly conserved (with one dominant amino acid having frequency larger than 0.75). These highly conserved positions tend to score low in SPEL, since shuffling a conserved position produces similar amino acid patterns on a tree compared with original pattern.

G protein α subunit

Heterotrimeric G proteins mediate cellular signaling by coupling ligand activated G protein-coupled receptors (GPCRs) to various intracellular effectors (Neer, 1995). The G protein α subunit (G_α) controls these signaling events through a regulated cycle of GTPase activity. G_α subunits of higher eukaryotes fall within four main subtypes based on primary sequence similarity and other criteria. Each of the classes performs different biological functions through specific interactions with effectors (e.g. cyclic GMP phosophodiesterase (PDE)) and regulators [e.g. Regulator of G protein signaling (RGS) domains or GoLoco motif], and the molecular basis of such specificities is an active area of research.

Table 2 shows 10 top-scoring positions predicted by SPEL, MI and ET. Figure 2B depicts the top ten predicted specificity positions

Table 1. Top specificity determinants for LacI/PurR protein family predicted by SPEL, MI method and ET method

Rank	SPEL Position	*P*-value	MI Position	*P*-value	ET Position
1	C85	5.21e−04	**T15**	3.0e−11	Y45
2	**D146**	7.15e−04	**D160**	2.0e−11	R52
3	**T15**	9.56e−04	**D146**	4.0e−09	I62
4	K5	4.23e−03	**K55**	6.0e−09	E82
5	**D160**	4.44e−03	W98	1.0e−08	C85
6	C123	5.90e−03	**T16**	2.0e−08	Y90
7	H20	6.59e−03	K114	2.0e−07	G199
8	W147	7.33e−03	F221	6.0e−07	D248
9	**K55**	7.70e−03	I249	4.0e−06	A251
10	**T16**	1.39e−02	M122	4.0e−06	

Dataset used here is LacI/PurR family of bacterial transcription factors. Positions 15, 16, 55, 146 and 160 (shown in bold) are predicted by both SPEL and MI methods. Rank does not apply to ET predictions.

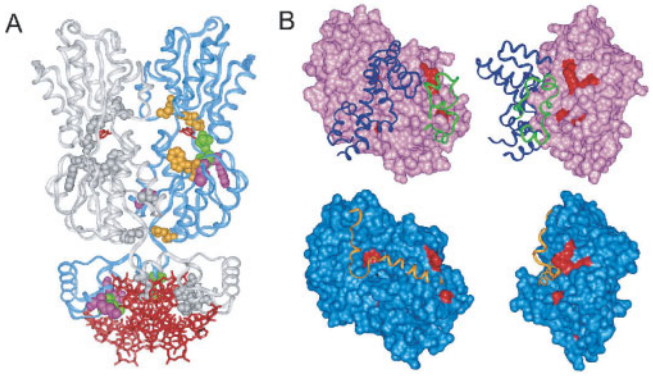

Fig. 2. Structural representation of prediction results. (**A**) Mapping the top 10 predicted functional specificity determinants by SPEL and MI to the protein structure of PurR (PDB ID: 1wet). Two chains of the PurR dimer are illustrated as ribbons (blue and gray). The ligands (small molecule guanine and DNA) are shown as red sticks. The specificity determinants mapped to one chain of the dimer is colored as follows: predicted by SPEL only (pink CPK), predicted by MI only (orange CPK), or predicted by both (Green CPK). The specificity determinants mapped to the other chain are in gray CPK representation. (**B**) Connolly surface renderings of G_α subunits. The upper two structures show the model of G_α (pink surface) bound to RGS (blue ribbon) and PDE (green ribbon) created from PDB ID 1fqj structure. The lower two structures show G_α (light blue surface) bound to GoLoco motif (orange ribbon) created from PDB ID 1kjy structure. The right view is generated by rotating protein structure 90° around the *y*-axis from the left view. Predicted specificity determinants are colored red. The figure is generated by Insight II package (Accelrys Software Inc.).

Table 2. Top 10 specificity determinants for G_α subunits predicted by SPEL, MI and ET

Rank	SPEL Position	*P*-value	MI Position	*P*-value	ET Position
1	**C250**	2.40e−02	**N145**	2.41E−14	K31
2	**T336**	2.62e−02	**T336**	2.58E−12	S40
3	**N145**	3.37e−02	**N252**	1.80E−11	K47
4	V30	3.61e−02	**M236**	3.29E−10	I51
5	**M236**	3.84e−02	**K244**	8.06E−10	W127
6	M243	4.29e−02	S173	3.46E−09	D129
7	S147	4.78e−02	E293	1.07E−08	T177
8	**K244**	5.21e−02	**C250**	1.37E−08	V197
9	**V46**	5.25e−02	**V46**	2.32E−08	C210
10	**N252**	5.48e−02	G296	4.29E−08	L245

Residue numbering is according to the G_α subunit structure with PDB ID 1fqj. Positions predicted by both SPEL and MI are in bold. Rank does not apply to ET predictions.

(Fig. 2B, red surface: C250, T336, N145, V30, M236, M243, S147, K244, V46 and N252, numbered according to PDB ID 1fqj) mapped to the surface of G_α (upper half of Fig. 2B, pink surface) complexed with RGS (upper half of Fig. 2B, blue ribbon) and PDE (upper half of Fig. 2B, green ribbon) and the same positions mapped to the surface of G_α (lower half of Fig. 2B, light blue surface) complexed with GoLoco (lower half of Fig. 2B, orange ribbon). The predicted residues form surface pockets near the G_α binding partners. Notably, the side chains of three residues (M236, K244 and N252) lie within 5 Å of the PDE effector while the side chain of residue N145 lies within 5 Å of the GoLoco motif. Although two residues form surfaces near the RGS binding site (V46 and N145), these surfaces remain some distance (8.5–10 Å, respectively) away from the regulator in this structure. Of the remaining residues, two (M243 and C250) continue the PDE effector-binding surface, one (S147) interacts with N145 and GDP, and two (V30 and T336) are mainly buried in the core of the structure. Our analysis suggests that most of the predicted specificity determining residues of the G_α family contribute to the effector-binding site (M236, K244, N252, M243 and C250).

The predictions of SPEL overlap well with the predictions of MI for the G-protein α subunit, with 7 of top 10 predictions being the same for these two methods (Table 2). However, the ET top 10 class-specific predictions are quite different from SPEL or MI. Similar to the result of the LacI/PurR family (Table 1), inspection of ET predictions revealed that most of them are highly conserved yet not invariant positions (data not shown). A comparison of the top 10 class-specific ET predictions made in our study with a previous ET analysis (Lichtarge *et al*., 1996b) revealed 6 common predictions. Five of them are invariant in the alignment used in the previous study (Lichtarge *et al*., 1996b). Since the G_α sequences in current protein database are much more divergent than those available in previous analysis, these five positions have become variant in our G_α alignment.

Robustness of the methods with regard to alignment quality

The alignments we used for predicting functional specificity positions in the above two protein families are high-quality alignments made with consideration of 3D-structures and/or with manual adjustment. However, a fully automated method for predicting functional sites should provide similar results for potentially lower quality alignments produced automatically. To study the robustness of SPEL, MI and ET, we made alignments automatically by ClustalW (Thompson *et al*., 1994) and PCMA (Pei *et al*., 2003). The prediction results from automatically generated alignments were compared with manually curated alignments used above.

Table 3. Number of predictions shared by different alignments for LacI/PurR family

Method	Manual	PCMA	ClustalW
SPEL			
Manual	20	17	12
PCMA	17	20	10
ClustalW	12	10	20
MI			
Manual	20	17	13
PCMA	17	20	12
ClustalW	13	12	20
ET			
Manual	27	24	25
PCMA	24	25	24
ClustalW	25	24	27

'Manual' is the alignment based on structural superposition and/or manual inspection. 'PCMA' is the alignment produced by PCMA. 'ClustalW' is the alignment produced by ClustalW. The diagonal elements in each table specify the number of top-scoring predictions considered in this analysis. The off-diagonal elements are the number of predictions common to two alignments.

In Table 3, we show the results for LacI/PurR family. For each method (SPEL, MI and ET), the top predictions made from three different alignments (manual, PCMA and ClustalW) are compared with each other. The diagonal elements are the number of top predictions selected for each alignment (for SPEL and MI, it is 20; for ET, it is the number closest to 20 at a cut level). The off-diagonal elements are the number of predictions shared by two alignments. The result from PCMA alignment is highly consistent with that of the manual alignment for all 3 methods, with only 3 positions different between top 20 predicted positions for each method. The result from ClustalW alignment also shares at least half of the predictions produced using the manual alignment, although it is less consistent with that of the manual alignment than PCMA alignment. For the G_α subunit, the same conclusions hold (see supplementary materials). Our results suggest that all three methods (SPEL, MI and ET) are robust with regard to the type of MSAs used. A comparison of alignments reveals that PCMA alignment has 92% position pairs and 63% entire columns aligned exactly according to the manual alignment. For ClustalW alignment, these two percentages are 88 and 39%, respectively. The reason for this robustness is probably that functionally important positions are relatively conserved and are aligned with good quality in all these alignments.

A larger-scale test of methods

We applied SPEL, MI and ET to 57 protein domain families that have been used in a previous comprehensive test of the ET method (Yao *et al*., 2003) (see Materials and methods section). We have

Table 4. Prediction results on 57 protein domain families

Method	No. of selected sequences from each cluster								
	3			6			10		
SPEL	2.95	0.102	0.118	2.81	0.099	0.120	2.75	0.091	0.120
	(0.31	0.017	0.015)	(0.30	0.017	0.016)	(0.30	0.015	0.016)
MI	2.51	0.080	0.091	2.82	0.096	0.110	2.82	0.091	0.105
	(0.32	0.016	0.014)	(0.32	0.018	0.016)	(0.34	0.015	0.012)
ET	2.58	0.083	0.105	2.74	0.097	0.121	2.65	0.098	0.122
	(0.27	0.012	0.011)	(0.25	0.013	0.016)	(0.26	0.013	0.015)

In each cell with 6 numbers, the upper three numbers are the means and the lower numbers are the their standard errors (in parentheses) of correct predictions among top 10 predictions, ROC_{10} and ROC_0.1, respectively.

designed a procedure that automatically selects database homologous sequences for the query structure by PSI-BLAST, and automatically selects sequences and defines subgroups (for MI method) by clustering the sequences using BLASTCLUST. Three statistics were used to compare the performance of SPEL, MI and ET: number of correct prediction among top 10 predictions, ROC_{10} and ROC_0.1 (see Materials and methods section). When selecting up to 3 sequences from the largest 8 clusters as the sequence set, SPEL produced on average 2.95 correct predictions among top 10 predictions. This is better than both MI (2.51 correct predictions) and ET (2.55 correct predictions) using the same sequence sets. The difference between SPEL and MI is statistically significant (P-value < 0.04). In this case, SPEL is also significantly better than MI according to the statistics of ROC_{10} and ROC_0.1. When selecting up to 6 or up to 10 sequences from the largest 8 clusters, the average number of correct top 10 predictions from SPEL slightly decreases to 2.81 and 2.75, respectively. However, the decreases are not statistically significant. For MI, selecting up to 6 or 10 sequences slightly but significantly improves the number of correct top 10 predictions to 2.82. For ET, selecting up to 3, 6 or 10 sequences from each group gives similar results (Table 4). Selecting 4 largest clusters instead of 8 largest clusters as groups decreases the prediction results for both SPEL and MI (supplementary materials). We also studied the effect of using different clustering techniques and found that single-linkage clustering performed similarly to mean-linkage clustering when using the program 'grouper' from the SEALS package (Walker and Koonin, 1997) for MI (supplementary materials). BLASTCLUST gave better results than 'grouper', possibly owing to restriction of length coverage on neighbors.

DISCUSSION

We have developed a new method (SPEL) for predicting functional specificity determinants with an input MSA. A phylogenetic tree is constructed from the MSA and a random model based on evolutionary simulation and shuffling is applied to each position to assign a P-value to it. A smaller P-value of a position suggests that it is more likely to be a functional specificity determinant.

SPEL does not require pre-defined sequence subgroups, in contrast to some other approaches (Hannenhalli and Russell, 2000; Mirny and Gelfand, 2002), nor does it require 2D or 3D structural information. Methods based on mutual information (Mirny and Gelfand, 2002) or relative entropy (Hannenhalli and Russell, 2000; Kalinina *et al*., 2004a) require pre-defined subgroups owing to the intrinsic property of their scoring functions. Subgroup definition poses a few limitations for these methods. First, although subgroup information can be obtained from experimental studies or comparative analysis of genomic sequences (e.g. ortholog/paralog identifications or gene structure analysis) (Overbeek *et al*., 1999; Tatusov *et al*., 1997), such information is not available for many protein families. In addition, the accuracy of subgroup partitions cannot be guaranteed based on either manual inspection of sequence similarities or phylogenetic patterns, or based on some automatic methods (Wicker *et al*., 2001). Second, different specificity positions could have different patterns of amino acid association with subgroups. Third, the phylogenetic information inside each subgroup is not fully explored in MI or relative entropy based methods. Evolutionary trace method (Lichtarge *et al*., 1996b) or its variants (Aloy *et al*., 2001; Armon *et al*., 2001; Innis *et al*., 2000) take into account the phylogenetic tree of the sequence set. However, they also require an *ad hoc* hierarchical partitioning of the tree into subtrees for identification of functional residues. The original evolutionary trace method (Lichtarge *et al*., 1996b) defines class-specific residues for those predictions that are not invariant. However, our analysis showed that many such positions are highly conserved. We expect that highly conserved positions are more likely to play a general role than to determine specificity. Besides, it does not provide estimation of the statistical significance for the positions. Although recent improvements on the evolutionary trace method have incorporated the non-random structural clustering or entropic information in ranking the positions (Madabushi *et al*., 2002; Mihalek *et al*., 2004), the limitation of subtree partitioning still exists.

Our method fully takes into account the phylogenetic information of the target alignment. Unlike ET, we use a more realistic neighbor-joining-based method (Bruno *et al*., 2000) to build the phylogenetic tree without assumption of molecular clock. We choose a log-likelihood (Felsenstein, 1996) based score (LLR) to assign statistical significance of a position being different from randomly generated positions without specificity. Our estimation of log-likelihood (Felsenstein, 1996) of amino acid distribution at a position incorporates the following information: (1) the overall tree, (2) estimated evolutionary rate of the position (3) amino acid composition of aligned protein sequences and (4) a general amino acid substitution model that is derived from analyzing a large set of protein families (Cai *et al*., 2004; Whelan and Goldman, 2001). The log-likelihood is larger for conserved positions (low entropy or generally slower evolutionary rates) than for variable positions (Figure S1A, supplementary materials). If two positions have the same amino acid composition, the position with conservation properties correlated better with the branching in a phylogenetic tree (i.e. the position with higher conservation within subtrees and lower conservation between subtrees) tends to have a higher log-likelihood. This last property combined with the strong statistical basis of log-likelihood, makes log-likelihood based scores appropriate for the task of discriminating specificity determining positions.

One novelty in the SPEL method is the random model based on the combination of evolutionary simulation and shuffling. This more complex random model (protein evolutionary simulation and shuffling) has been shown to give reasonable assignments of P-values and correct ranking of expected specificity positions

(supplementary materials). On the contrary, random models based on either shuffling or evolutionary simulation do not perform as expected (supplementary materials). Compared with the *P*-values given by MI, SPEL *P*-values seem to be more reasonable, with the top-ranked positions having *P*-values in the range of 0.01–0.001 for the LacI/PurR family (Table 1). MI gives much lower and seemingly unreasonable *P*-values for the top predictions for the same family (Table 1); and there are about 60 out of 280 positions that have *P*-values below 0.01 (data not shown). Since there are many positions with low *P*-values, the authors in the MI method have to rely on the combination of *P*-value and mutual information value with *ad hoc* cutoffs to judge the predictions (Mirny and Gelfand, 2002). SPEL gives higher *P*-values (between 0.5 and 0.01) to top predictions of G_α subunit family than the LacI/PurR family. This is probably owing to the higher sequence similarity among the G_α subunit members than the LacI/PurR family, causing G_α specificity positions to have weaker amino acid associations with functional subgroups.

The secondary novelty in SPEL is the design of LLR as the scoring function. The formula of LLR normalizes the log-likelihood of the original amino acid distribution against the average LLR of a sample from shuffled position, making different positions comparable. Since the shuffling procedure destroys any association of amino acid type with the phylogenetic tree, the absolute values of log-likelihoods of shuffled positions tend to be higher than that of the original position. Thus the value of *LLR* is usually between 0 and 1. If the amino acid distribution in a position has a stronger association with subgroups, its LLR will be smaller. This property of LLR and the strong statistical basis of log-likelihood make LLR a good scoring function in selecting specificity-determining positions.

Our method of specificity prediction has a few limitations. Like the methods based on mutual information or relative entropy, the following main assumptions apply to our strategy. First, sequences of proteins with the same functional specificity are more similar to each other than to sequences of proteins with different specificity. Second, all or most of the proteins in the alignment use the same positions as key determinants of specificity. Third, proteins with the same specificity use similar amino acid residues at these key positions and proteins with different specificity employ different residues. Other assumptions in our method are related to tree building or evolutionary simulations (e.g. independence between positions, Markovian substitution process, etc. see Cai *et al*., 2004). It seems reasonable that (1) functional differences in specificity between homologs are correlated with the phylogenetic divergence of their sequences; (2) locations of functional sites are similar in homologs and (3) proteins with the same specificity have conserved amino acids in key positions, but these key positions should show differences in amino acid usage for proteins with different specificity. However, exceptions to these assumptions could exist. For example, functional specificity could be contributed by different positions in different subfamilies; or convergent evolution has resulted in the same or similar amino acid types between phylogenetically distant subfamilies. Our method requires only an MSA. We have shown that automatically generated alignments give similar performance to manually curated alignments. This could be owing to the predicted specificity determinants in sequence motifs that are aligned correctly by automatic alignment programs. However, it is still required that the alignment is relatively divergent and should include paralogous proteins with altered specificity. If an alignment contains only orthologous proteins, there would be no positions having strong association between amino acid types and phylogenetic patterns. In this case, we expect that *P*-values will be high for all positions.

Testing and comparing the performance of different methods on finding specificity determinants is a difficult task owing to lack of specificity information for every position in any protein family. Therefore, accurate definitions of true/false positives/negatives of specificity determinants are difficult for any protein family. Several previous articles about specificity prediction methods (Hannenhalli and Russell, 2000; Kalinina *et al*., 2004a; Mirny and Gelfand, 2002) have chosen a few well-studied protein families to study their performance. Similarly, we exemplify the performance of our method on a few well-studied protein families (LacI/PurR and G protein α subunit; see also the GST family in supplementary materials). Our results were compared with those of two known methods (MI and ET). Generally, predictions given by SPEL are similar to those of MI, and are quite different from ET class-specific predictions that could include many highly conserved positions. In addition to the specific examples, we carried out a larger-scale test of these methods on two sets of proteins with bound ligands, which have been used in a previous study of functional site prediction (Yao *et al*., 2003). Protein active site is usually around the ligand-binding site. Residues close to ligands include highly conserved residues having the same function for every member of a protein family as well as variable residues that do not contribute much to function (DeLano, 2002; Ma *et al*., 2001). Closeness to the ligands can by no means guarantee a residue to be specificity determinant. However, it has been frequently observed that specificity determinants make direct contact with ligands. Therefore, without detailed knowledge about specificity determinants of every residue in the testing set, closeness to the ligands was used as the criterion to judge the performance of the three methods.

We have designed a procedure to automatically select homologs and define subgroups (used only in MI) for a given sequence by similarity searches and clustering. Using more sequences from more clusters should be able to improve performance owing to increase in sample size. Tested on the 57 protein domain families with known structures, we showed that using 8 largest groups did give better results for SPEL, MI and ET than using 4 largest groups. However, selecting up to 6 or 10 sequences from the largest 8 clusters in each group gave similar but slightly worse results for SPEL, compared with selecting up to 3 sequences. The small decrease in SPEL performance when selecting more sequences could be caused by shifting the balance of the sample, as some clusters contain a small number of sequences (<3). Uneven representation of subgroups is a potential issue for SPEL since it uses the shuffling procedure. Using an extreme case to illustrate this point, if one subgroup contains 100 nearly identical sequences while the other two subgroups contain only a couple of sequences, we would not expect significant changes in amino acid distribution on the tree after shuffling, even for specificity-determining positions. Additionally taking into account that the extensive simulation and shuffling procedure in SPEL requires a lot of CPU time (about 3 orders of magnitude slower than MI, see supplementary materials), we recommend using a small but balanced sample of sequences for SPEL. To achieve that, using clustering to select potential subgroups and to remove uneven sampling could be an efficient and a necessary step before applying SPEL.

ACKNOWLEDGEMENTS

We are grateful to Leonid A. Mirny for providing the MSA of LacI/PurR family. This work was supported by the NIH grant GM67165 to N.V.G.

Conflict of Interest: none declared.

REFERENCES

Aloy,P. *et al.* (2001) Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. *J. Mol. Biol.*, **311**, 395–408.

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, **25**, 3389–3402.

Armon,A. *et al.* (2001) ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J. Mol. Biol.*, **307**, 447–463.

Bielawski,J.P. and Yang,Z. (2004) A maximum likelihood method for detecting functional divergence at individual codon sites, with application to gene family evolution. *J. Mol. Evol.*, **59**, 121–132.

Bruno,W.J. *et al.* (2000) Weighted neighbor joining: a likelihood-based approach to distance-based phylogeny reconstruction. *Mol Biol. Evol.*, **17**, 189–197.

Cai,W. *et al.* (2004) Reconstruction of ancestral protein sequences and its applications. *BMC Evol. Biol.*, **4**, 33.

Casari,G. *et al.* (1995) A method to predict functional residues in proteins. *Nat. Struct. Biol.*, **2**, 171–178.

Delano,W.L. (2002) Unraveling hot spots in binding interfaces: progress and challenges. *Curr. Opin. Struct. Biol.*, **12**, 14–20.

Elcock,A.H. (2001) Prediction of functionally important residues based solely on the computed energetics of protein structure. *J. Mol. Biol.*, **312**, 885–896.

Felsenstein,J. (1996) Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Methods Enzymol.*, **266**, 418–427.

Glasfeld,A. *et al.* (1999) The role of lysine 55 in determining the specificity of the purine repressor for its operators through minor groove interactions. *J. Mol. Biol.*, **291**, 347–361.

Hannenhalli,S.S. and Russell,R.B. (2000) Analysis and prediction of functional subtypes from protein sequence alignments. *J. Mol. Biol.*, **303**, 61–76.

Innis,C.A. *et al.* (2004) Prediction of functional sites in proteins using conserved functional group analysis. *J. Mol. Biol.*, **337**, 1053–1068.

Innis,C.A. *et al.* (2000) Evolutionary trace analysis of TGF-beta and related growth factors: implications for site-directed mutagenesis. *Protein Eng.*, **13**, 839–847.

Jones,S. and Thornton,J.M. (2004) Searching for functional sites in protein structures. *Curr. Opin. Chem. Biol.*, **8**, 3–7.

Kalinina,O.V. *et al.* (2004a) Automated selection of positions determining functional specificity of proteins by comparative analysis of orthologous groups in protein families. *Protein Sci.*, **13**, 443–456.

Kalinina,O.V. *et al.* (2004b) SDPpred: a tool for prediction of amino acid residues that determine differences in functional specificity of homologous proteins. *Nucleic Acids Res.*, **32**, W424–W428.

La,D. *et al.* (2005) Predicting protein functional sites with phylogenetic motifs. *Proteins*, **58**, 309–320.

Lehming,N. *et al.* (1990) Mutant lac repressors with new specificities hint at rules for protein—DNA recognition. *EMBO J.*, **9**, 615–621.

Lichtarge,O. *et al.* (1996a) Evolutionarily conserved Galphabetagamma binding surfaces support a model of the G protein-receptor complex. *Proc. Natl. Acad. Sci. USA*, **93**, 7507–7511.

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

Livingstone,C.D. and Barton,G.J. (1993) Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. *Comput. Appl. Biosci.*, **9**, 745–756.

Ma,B. *et al.* (2001) Protein functional epitopes: hot spots, dynamics and combinatorial libraries. *Curr. Opin. Struct. Biol.*, **11**, 364–369.

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

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

Mirny,L.A. and Gelfand,M.S. (2002) Using orthologous and paralogous proteins to identify specificity-determining residues in bacterial transcription factors. *J. Mol. Biol.*, **321**, 7–20.

Mirny,L.A. and Shakhnovich,E.I. (1999) Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. *J. Mol. Biol.*, **291**, 177–196.

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

Overbeek,R. *et al.* (1999) The use of gene clusters to infer functional coupling. *Proc. Natl Acad. Sci. USA*, **96**, 2896–2901.

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

Pei,J. and Grishin,N.V. (2001) AL2CO: calculation of positional conservation in a protein sequence alignment. *Bioinformatics*, **17**, 700–712.

Pei,J. *et al.* (2003) PCMA: fast and accurate multiple sequence alignment based on profile consistency. *Bioinformatics*, **19**, 427–428.

Pupko,T. *et al.* (2002) Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics*, **18** (Suppl. 1), S71–S77.

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

Sartorius,J. *et al.* (1991) The roles of residues 5 and 9 of the recognition helix of Lac repressor in lac operator binding. *J. Mol. Biol.*, **218**, 313–321.

Sauer,R.T. (1996) Lac repressor at last. *Structure*, **4**, 219–222.

Schaffer,A.A. *et al.* (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.*, **29**, 2994–3005.

Shulman-Peleg,A. *et al.* (2004) Recognition of functional sites in protein structures. *J. Mol. Biol.*, **339**, 607–633.

Sowa,M.E. *et al.* (2000) A regulator of G protein signaling interaction surface linked to effector specificity. *Proc. Natl Acad. Sci. USA*, **97**, 1483–1488.

Soyer,O.S. and Goldstein,R.A. (2004) Predicting functional sites in proteins: site-specific evolutionary models and their application to neurotransmitter transporters. *J. Mol. Biol.*, **339**, 227–242.

Tatusov,R.L. *et al.* (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.

Theodoridis,S.K. *et al.* (1999) *Pattern Recognition. Academic Press.*

Thompson,J.D. *et al.* (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.

Todd,A.E. *et al.* (2001) Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.*, **307**, 1113–1143.

Walker,D.R. and Koonin,E.V. (1997) SEALS: a system for easy analysis of lots of sequences. *Intell. Syst. Mol. Biol.*, **5**, 333–339.

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.

Wicker,N. *et al.* (2001) Secator: a program for inferring protein subfamilies from phylogenetic trees. *Mol. Biol. Evol.*, **18**, 1435–1441.

Wilcoxon,F. (1947) Probability tables for individual comparisons by ranking methods. *Biometrics*, **3**, 119–122.

Wilson,C.A. *et al.* (2000) Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J. Mol. Biol.*, **297**, 233–249.

Wolf,Y.I. *et al.* (1999) Evolution of aminoacyl-tRNA synthetases—analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events. *Genome Res.*, **9**, 689–710.

Yao,H. *et al.* (2003) An accurate, sensitive, and scalable method to identify functional sites in protein structures. *J. Mol. Biol.*, **326**, 255–261.