doi:10.1016/j.jmb.2007.08.036 J. Mol. Biol. (2007) 373, 801–810

JMB

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Functional Specificity Lies within the Properties and Evolutionary Changes of Amino Acids

Saikat Chakrabarti*, Stephen H. Bryant and Anna R. Panchenko

National Center for Biotechnology Information National Library of Medicine National Institutes of Health Bethesda, MD 20894, USA

Received 1 May 2007; received in revised form 3 July 2007; accepted 16 August 2007
Available online 22 August 2007

The rapid increase in the amount of protein sequence data has created a need for automated identification of sites that determine functional specificity among related subfamilies of proteins. A significant fraction of subfamily specific sites are only marginally conserved, which makes it extremely challenging to detect those amino acid changes that lead to functional diversification. To address this critical problem we developed a method named SPEER (specificity prediction using amino acids' properties, entropy and evolution rate) to distinguish specificity determining sites from others. SPEER encodes the conservation patterns of amino acid types using their physico-chemical properties and the heterogeneity of evolutionary changes between and within the subfamilies. To test the method, we compiled a test set containing 13 protein families with known specificity determining sites. Extensive benchmarking by comparing the performance of SPEER with other specificity site prediction algorithms has shown that it performs better in predicting several categories of subfamily specific sites.

Keywords: functional divergence; subfamily specificity; physico-chemical properties; combined relative entropy; evolutionary rate

Edited by F. E. Cohen

Introduction

According to the neutral theory of molecular evolution the majority of mutations are selectively neutral at the molecular level and do not affect the fitness of the organism.[1] As a consequence many protein sites undergo random amino acid changes, which are apparently not functional and are not conserved in evolution. Other sites are under more stringent evolutionary constraints that are reflected in the more prominent conservation of sequence and structural properties. It has been argued that changes in the conservation or evolutionary rate at a particular site reflect functional divergence after the gene duplication.[2,3] Indeed, after duplication of a gene, one copy evolves under relaxed evolutionary constraints, which allow it to accumulate changes and develop new functions and specificities.[3,4] Such mechanisms of functional diversification have recently been studied in proteins with promiscuous functions,[5–7] and two types of functional divergence have been distinguished.[8] Type I functional divergence is the result of the change in evolutionary rate where the site is conserved for one subfamily and is variable in another. Type II divergence is a consequence of the rate change where purifying selection causes similar levels of conservation of different amino acid types for different protein subfamilies.

Various site-specific conservation scores have been offered to distinguish conserved functionally important sites from the background of neutral changes.[9] Some of them are based on combinatorics and information theory, including different variations of Shannon entropy and frequency scores.[10–14] Others take into account amino acid stereochemical properties[15–19] and amino acid substitution matrices.[20,21] Since there is heterogeneity in evolutionary rates between sites, models, which account for the difference in rates and amino acid substitution probabilities among different sites can be very valuable as well.[22,23] It has also been shown that prediction of functional sites and site-specific rate inference can be improved considerably if phylogenetic trees and evolutionary models

*Corresponding author. E-mail address: chakraba@ncbi.nlm.nih.gov.

Abbreviations used: ED, Euclidean distance; ER, evolution rate; CRE, combined relative entropy; SPEER, specificity prediction using amino acids' properties, entropy and evolution rate; MC, marginally conserved; GHPS, dihydropteroate synthase; ROC, receiver operating characteristic.

0022-2836/$ - see front matter. Published by Elsevier Ltd.

are considered.[24–31] Other methods attempt to identify functional sites based not only on the sequence conservation but also on their location in the 3D structure.[26,32–38]

Several computational methods have been developed which are exclusively designed to predict specificity determinants. Earlier algorithms applied principal component analysis to a vector representation of protein sequences[39] or self-organizing maps to retrieve sequence patterns characteristic of subfamilies.[40] The evolutionary trace method, for example, identified invariant specific residues by partitioning the phylogenetic tree into subgroups of similar sequences and its later versions estimate the statistical significance of the predictions.[27,41] Some more recent methods use multiple sequence alignments and various conservation scores like relative entropy, mutual entropy or "sequence harmony" to predict subfamily specific sites.[42–46] The majority of specificity determining methods require pre-defined grouping into subfamilies while several of them overcome this limitation by simultaneous identification of optimal groups and conserved positions.[47,48] In the first approach the likelihood score is calculated for each position using the phylogenetic tree and a shuffling procedure.[47] The second approach uses a Bayesian-based model for identification of specificity determinants, and in this case the Bayes factors allow one to estimate the uncertainty level of the solution.[48]

It is extremely difficult to detect amino acid changes which lead to functional divergence. It is indeed much easier to distinguish globally conserved sites from the overall background rather than differentiate between the two types of conservation in various subfamilies. The reason is that specificity is determined by subtle changes in residue stereochemistry and the residue conservation score should be tuned to detect these changes. Moreover, in many cases sites responsible for specificity are located on flexible or disordered loops that are difficult to characterize.[5] Finally, experiments on specificity determinants are difficult and compiling a comprehensive dataset for testing these prediction methods is a major task.

Indeed, despite all efforts at predicting subfamily specific sites, accuracy remains very limited and some methods are tuned to predict only type I functional sites while others are biased toward the type II functional sites. In reality it is almost impossible to judge their performance using a few test families, which is the case for most of the studies. In our work we compiled a more comprehensive test set which consisted of 13 protein families with the pre-determined specificity sites. Using this test set we analyzed the site attributes which can distinguish between different subfamilies of the same family alignment. We developed a method named SPEER (specificity prediction using amino acids' properties, entropy and evolution rate) that encodes the specific conservation pattern of amino acid types together with their physico-chemical properties and the evolutionary rates between the subfamilies. We have also undertaken by far the most extensive benchmarking analysis in this field where the SPEER method has been compared to other available specificity site prediction methods. Comparison results suggest better performance of SPEER with respect to other methods. The prediction sensitivity provided by our combinatorial approach is good (close to 70% at 15% error rate) and our findings are encouraging for future investigations.

Results and Discussion

Characterization of subfamily specific sites

Subfamily specific sites (110 sites altogether) collected from 13 families are categorized into three major classes, type I, type II and marginally conserved (MC) sites (Figure 1). As can be seen on this Figure, about half of subfamily specific sites are only marginally conserved which reflects the lack of regularity in conservation pattern and thereby illustrates the difficulties in identifying them through prediction methods. Another half constitute type I and type II sites; these two types of conservation are shown to occur more frequently among subfamily specific rather than non-subfamily specific sites.

We developed a scoring function (SPEER score) that represents a linear combination of Euclidean distances (ED score) based on amino acids' physico-chemical properties, evolution rate (ER) and combined relative entropy (CRE). All three terms account for the variability of sites within the subfamilies in terms of their physico-chemical properties, evolutionary rates and amino acid types. Figure 2 shows the distribution of three components of our combined scoring scheme, ED score, CRE and ER scores together with the combined SPEER score calculated for subfamily specific and all other sites in the alignments. Although not all scores demonstrate good discrimination between subfamily

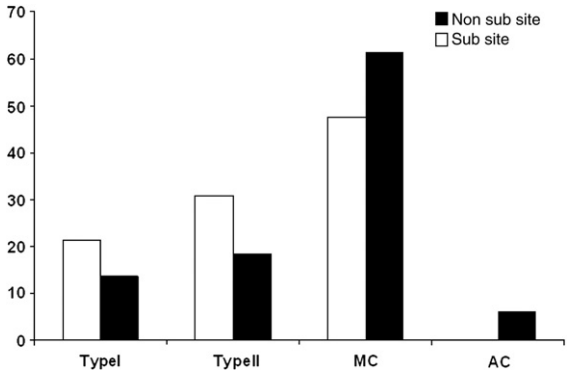

Figure 1. Distribution of different categories of subfamily specific sites. Percentage of typeI, typeII, marginally conserved (MC) and absolutely conserved (AC) sites are shown in known subfamily specific sites (Sub site; white bar) and non-subfamily specific sites (Non sub site; black bar).

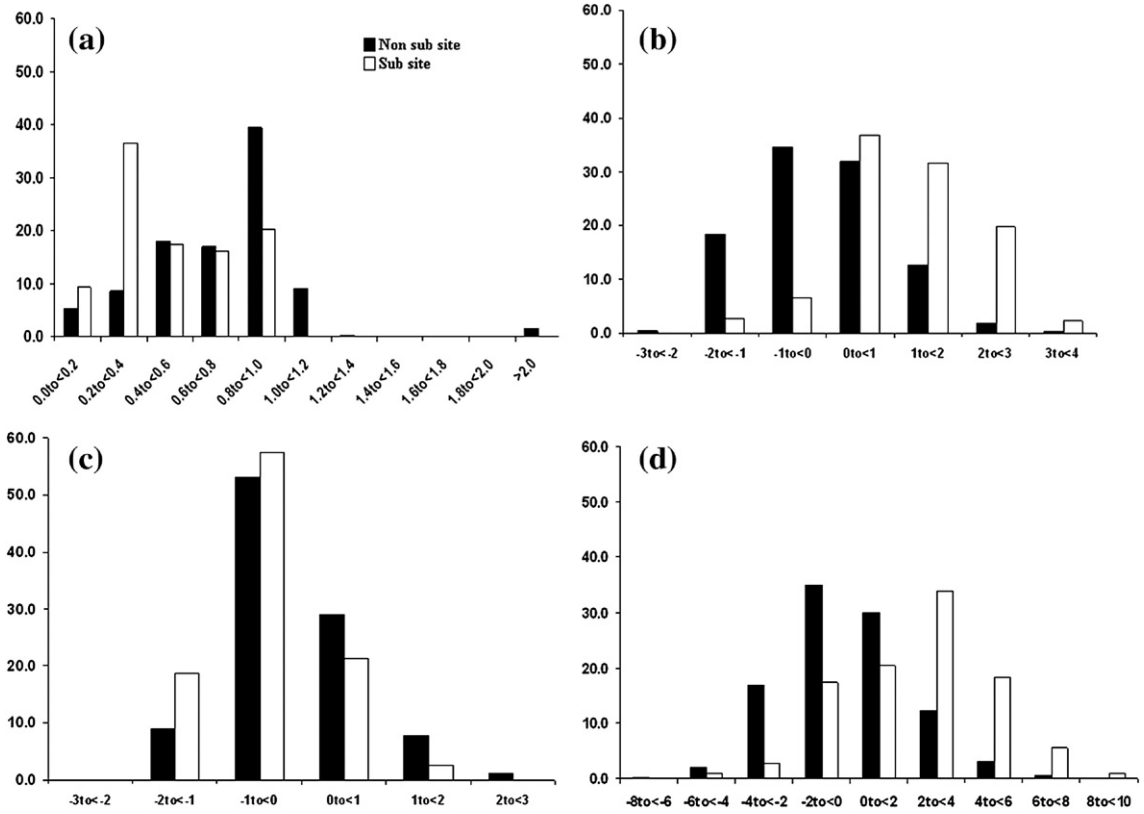

Figure 2. Distribution of three component scores, i.e. (a) ED score, (b) CRE score and (c) ER along with the combined SPEER score (d) are shown for subfamily specific sites (Sub site; white bar) and non-subfamily specific sites (Non sub site; black bar). *X*-axes represent the range of scores while *Y*-axes show the percentage of sites.

determinants and other sites, the combined score clearly has power to discriminate between these two site populations, which suggests the complementarity of the proposed scoring schemes. Indeed, the correlation matrix calculated for different scoring terms shows that correlation coefficients are low (Supplementary Data, Table SM1).

Prediction of subfamily specific sites

We have used multiple alignments of thirteen protein families to predict subfamily specific sites. The combined SPEER score was calculated for each gapless column of the alignment where no amino acid type was represented more than 80% of the times (see Materials and Methods). The prediction sensitivities at 1, 5 and 15% error rate together with the receiver operating characteristic (ROC) statistics and their standard deviations are given in Table 1. Prediction sensitivities (at 1 and 15% error rate) for individual families are also provided in Table 2. As can be seen from these tables and the overall ROC curve (Figure 3), for the majority of families (62%, 8 out of 13) the SPEER method outperforms other methods such as SDP-pred,[44] SPEL[47] and SH.[45] The difference in prediction performance between SPEER and other methods is also statistically significant as suggested by ROC_{500}, ROC_{total} and the Wilcoxon signed-ranked test *p*-values (p-value<0.004). For 3 out of 13 families (cd00120, LacI and GST) other methods yield better predictions at certain error rates. SPEL and SDP-pred, overall, yield similar performance, although SPEL seems to show somewhat higher sensitivities at low error rates compared to SDP-pred. The SH algorithm can not be compared with other methods across all the families as it can not make predictions for families with more than two subfamilies. It should be mentioned that this comparison does not take into account certain strong points of the other methods which are not directly associated with the problem being solved here. For example, SPEL can simultaneously define subfamilies and predict specificity determinants, and SDP-pred takes full advantage of orthologous–paralogous groupings in defining the

Table 1. Comparison of overall prediction sensitivities

	Prediction sensitivity (%)		
Error rate	SPEER	SPEL	SDP-pred
1%	12	6	3
5%	47	34	19
15%	68	54	59
ROC_{500}	0.538±0.012	0.446±0.011	0.408±0.013
ROC_{total}	0.820±0.006	0.783±0.007	0.784±0.007

SDP-pred[44] and SPEL[47] are other methods for prediction of subfamily specific sites.

Table 2. Comparison of prediction sensitivities for individual families

Name of families	Number of specific sites	SPEER	SDP-Pred	SPEL	SH
cd00120	3	0; 45	0; 50	0; **100**	0; 13
cd00264	3	0; **67**	0; 33	0; **67**	0; 0
cd00333	12	**37; 75**	30; **75**	33; 71	0; 67
cd00363	6	**27; 83**	15; **83**	0; 67	0; 0
cd00365	10	**10; 60**	7; **60**	0; 40	0; 0
cd00423	4	**25; 100**	**25; 100**	**25**; 75	0; 50
cd00985	3	**67; 100**	**67**; 67	0; 67	0; 67
Gprotein	7	**73; 100**	61; **100**	36; 92	–
GST	9	54; **67**	43; **67**	**56; 67**	–
LacI	14	36; **95**	52; 86	**57**; 84	–
Ricin	21	0; **19**	0; 10	–	–
CNMyc	11	**9; 27**	0; 20	0; 9	0; 0
CBM9	7	0; **71**	0; 43	0; 38	0; 43

Sensitivity values for predicting correct subfamily specific sites are shown at 1% and 15% error rates. Maximum values at each error rate are marked in bold. Cases where no results were obtained are marked (–).

subfamily specific sites. We further examine the performance of methods in predicting different types of subfamily specific sites (Figure 4). It is clear from the Figure that SPEER performs well for all three categories including the most difficult type I and marginally conserved (MC) sites, which pose a significant challenge for computational identification of subfamily determinants. Likewise, SPEL makes very good predictions for the MC category as well. We have also observed that, overall, the prediction accuracy depends on the level of conservation of physico-chemical properties within the subfamilies (Pearson correlation coefficient is 0.61) as well as between them (Pearson correlation coefficient is −0.66) (Supplementary Data, Figure SM1).

Examples of successful predictions

We illustrate the performance of the SPEER method on different examples (Figure 5). Figure 5(a) shows a representative structure of dihydropteroate synthase (1AJ0) taken from the pterin binding enzymes domain family (cd00423). This family includes two subfamilies, dihydropteroate synthase (DHPS) and cobalamin-dependent methyltransferases. DHPS catalyzes the condensation of *p*-aminobenzoic acid (pABA) in the biosynthesis of folate, which is an essential cofactor in both nucleic acid and protein biosynthesis. DHPS represents a very important subfamily as it can be targeted by sulfonamide drugs, which are substrate analogs of pABA. Both DHPS and cobalamin-dependent methyltransferases bind to pterin substrates while sulfonamide (pABA) acts as a specific ligand to DHPS. SPEER and SDP-pred methods successfully identified all four (Lys220, Arg221, Arg255 and His257; marked as space-filling model) sites for pABA/sulfonamide binding in DHPS.[49–51] In addition to that SPEER was able to predict three additional sites (Ile20, Gly187 and Gly189) that could be important in specific interaction and reside within 5 Å from the specific pABA ligand.

Another example shows a representative structure of a novel NTPase from *Methanococcus jannaschii* (2MJP, chain A) belonging to Maf_Ham1 domain family (cd00985; Figure 5(b)). Ham1-related protein is a novel NTPase that has been shown to hydrolyze non-standard nucleotides, such as hypoxanthine/xanthine NTP. The Maf subfamily includes nucleotide-binding proteins which have been implicated in inhibition of septum formation in eukaryotes, bacteria and archaea. Despite the fact that proteins from both subfamilies share structural similarities

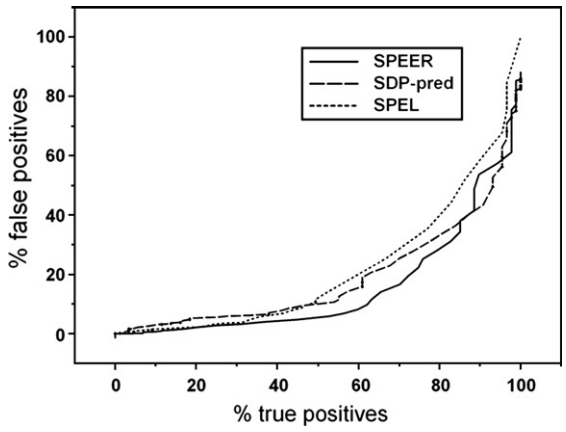

Figure 3. Comparison of prediction performances. ROC-curves for prediction of subfamily specific sites are shown for SPEER, SDP-pred[44] and SPEL[47] methods.

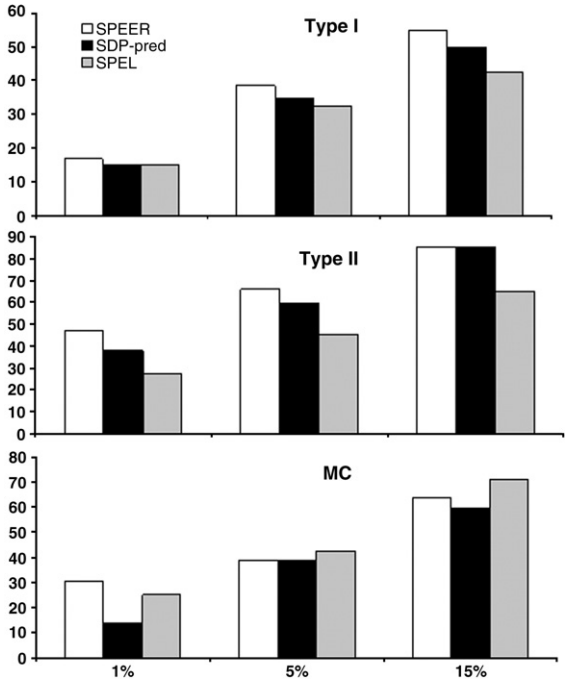

Figure 4. Comparison of prediction performances for different categories of subfamily specific sites. Percentage of sites (*Y*-axes) predicted by SPEER, SDP-pred and SPEL at 1, 5 and 15% error rates are shown for typeI, typeII and marginally conserved (MC) sites.

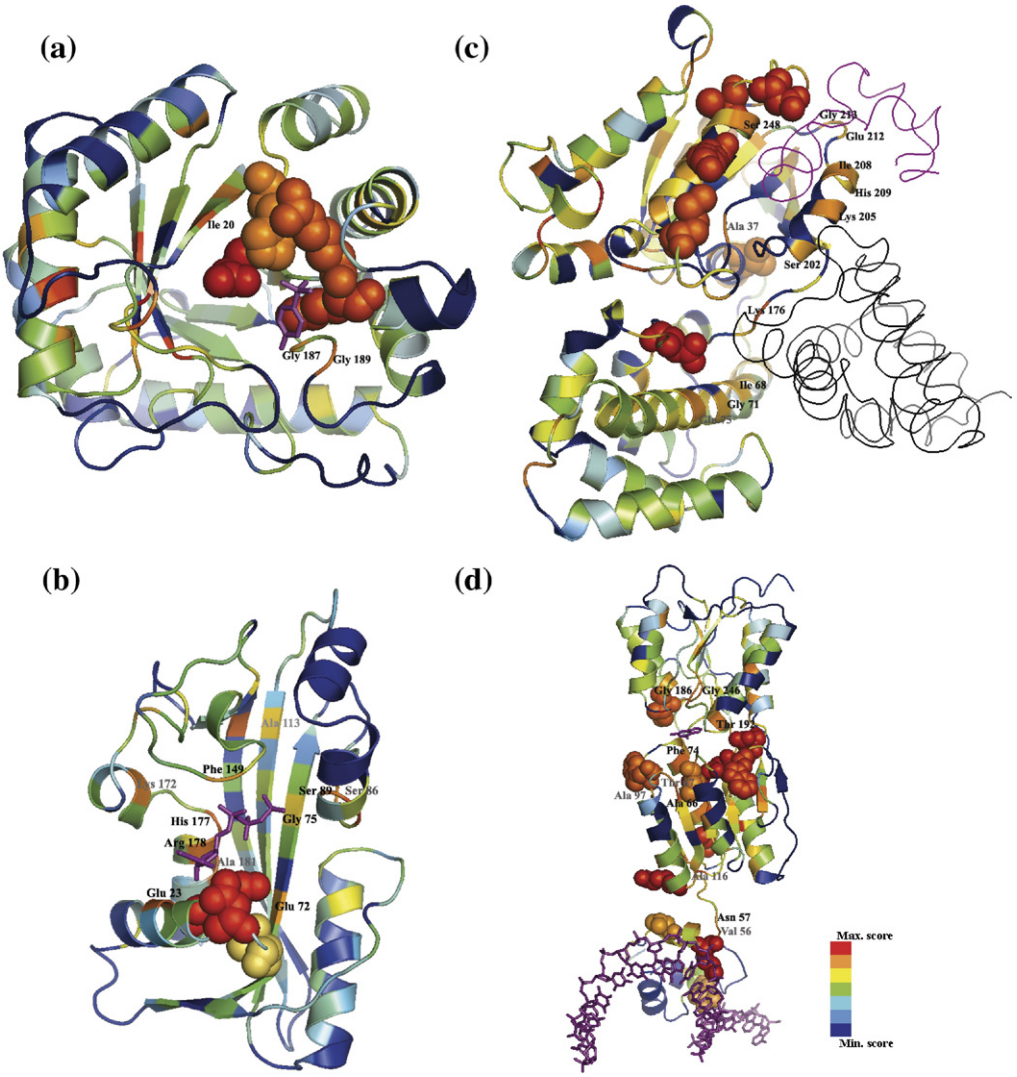

Figure 5. Examples of successful predictions. (a) Cartoon representation of dihydropteroate synthase (1AJ0) taken from the pterin binding enzymes domain family. SPEER scores for this family are projected onto the 3D structure where correctly predicted known subfamily specific sites are shown in space filling model. Three additional sites (Gly187, Gly189 and Ile20) that could be important in specific interaction are also labeled as they have high scores by SPEER and reside within 5 Å of the specific ligand (shown in purple). (b) A representative structure of a novel NTPase from *Methanococcus jannaschii* (2MJP, chain A) belonging to Maf_Ham1 domain family. Known specific sites (Thr15, Asn17 and Lys20) are shown in space filling model while additional high scoring sites residing closely to the specific ligand are mapped (within 5 Å are in black font; within 10 Å are in grey font) onto the 3D structure. (c) 3D structure representative (1FQJ, cartoon representation) of G protein α subunit family bound to PDE (purple ribbon) and RGS (black ribbon). Known specific sites (Met236, Lys244, Cys250 and Asn252 for PDE binding; Val46, Asn145 and Thr336 for RGS binding) are shown in space filling model while additional high scoring sites residing closely to the specific binding partners are mapped (within 5 Å are in black; within 10 Å are in grey) onto the 3D structure. (d) Representative 3D structure (1WET) from the PurR subfamily complexed with the effector, guanine and the ligand, DNA (shown in purple).

in the nucleotide-binding cleft, the locations and nature of conserved residues differ, which could lead to adoption of different functions and ligand binding properties. Three such conserved residues (Ser9/Thr15, Ser11/Asn17 and Arg14/Lys20 in Maf and Ham1, respectively) could be important for binding to different nucleotides and therefore can be regarded as subfamily specific.[52,53] SPEER successfully identified all three binding sites (shown in space filling model in Figure 5(b)) at 15% false positive rate. Additionally, we predict seven extra potential specificity determinants, which have high scores and reside within 5A° (Glu23, Glu72, Gly75, Ser89, Phe149, His177 and Arg178) from the ligand (Figure 5(b)).

The third example constitutes the G protein α subunit (G_α, Gprotein), which controls important cellular signaling processes involving G protein coupled receptors through a regulated cycle of GTPase activity. G_α subunits can be divided into four main subtypes where each of the subtypes performs different biological functions through

specific interactions with the effectors (e.g. cyclic GMP phosophodiesterase (PDE)) and regulators (e.g. Regulator of G protein signaling (RGS) domains).[47] Figure 5(c) shows the predicted subfamily specific sites mapped onto the 3D structure representative (1fqj, cartoon representation) bound to PDE (purple ribbon) and RGS (black ribbon). Potential specificity determinant sites that reside within 5 Å or 10 Å from the effector and/or regulator molecules are marked in black and grey correspondingly.

The LacI/PurR family is a large family of bacterial transcription factors (15 subfamilies) that are regulated by small molecules, such as sugars and nucleotides. In addition to available experimental and structural information, the LacI/PurR family has been widely used by researchers for prediction of subfamily specific sites. Specific sites predicted by SPEER are mapped onto a representative 3D structure (PDB code 1WET) from the PurR subfamily complexed with the guanine (effector) and the ligand, DNA (Figure 5(d)). All the predicted sites are color-coded based on their SPEER prediction score (see color scale in Figure 5) and the known subfamily specific sites are marked in space filling model.

Conclusion

The problem of identifying specificity determinants is both challenging and captivating as its solution would point to the evolutionary and physico-chemical mechanisms producing a wide variety of specific functional activities based on the same fold and overall function of a protein family. Since proteins with similar specificities use similar amino acids, specificity prediction methods look for the specific distribution patterns (that could be directly related to the biochemical function or be characteristic of a given subfamily) of amino acids across the subgroups or with respect to the overall family and try to identify those sites where such a subfamily specific distribution is observed.

Here we investigated the factors which can distinguish between different subfamilies of the same family. First we found that it is important to encode the conservation of amino acids' properties within each subfamily and differences between subfamilies (ED term). Second, we showed that the conservation of subfamily specific features can be successfully described in terms of amino acid substitution rates (ER term) which are calculated from the phylogenetic trees and reflect the evolutionary history of family divergence. Finally, we noticed that amino acid properties can be very similar between different subfamilies at specificity determining sites, although their amino acid usage can vary. Consequently, the difference in amino acid usage between and within subfamilies should also be encoded explicitly (CRE term).

We note that variations of many measures employed in our cost function have been used[8,23,28] for characterization and prediction of specificity determinants for selected families. Here we present a more general approach tested on a benchmark encompassing a diverse set of protein families, which showed that the simple combination of seemingly redundant but in fact complementary terms performs well in prediction. Comparison with other sensitive methods of specificity prediction showed that although SPEER in many cases yields better results, the methods' sensitivities are still moderate. On the other hand, many examples of successful predictions have been found by our method. Considering the difficulty level and the current state of the field, the prediction sensitivity provided by our combinatorial approach is very much acceptable and encouraging enough for further future investigations. Therefore, the present study provides a platform for future endeavors to understand the critical issue of protein subfamily specificity determination.

Materials and Methods

Benchmark for prediction validation

We have performed an extensive analysis to collect reliable alignments of protein families, for which experimental evidence is available on subfamily specific sites. Our benchmark includes seven families that have been used for validation of previously published prediction methods and six families from the version 2.10 of the Conserved Domain Database (CDD[54]). Subfamily specific sites for six CDD families were assigned based on an extensive literature search (see Supplementary Data for details). A complete list of the test set families together with their subfamily specific site locations is provided in Table 3. Highly conserved positions within the overall family alignment (where any amino acid type was represented more than 80% of the time) were not regarded as subfamily specific and excluded from the analysis. The resulting test set covers a wide range of families with different functions, types of functional sites, number of subfamilies and sequence diversity (Supplementary Data, Table SM2). To our knowledge this is the most comprehensive benchmark used so far for validation of subfamily specific site prediction. These alignments and subfamily specific sites information can be obtained through e-mail request or can be downloaded *via* ftp†.

All specificity determining sites were categorized into three groups, type I, type II and marginally conserved (MC). Type I functional sites were defined as those conserved for one subfamily and variable in another while type II sites were defined as those where different types of amino acids were conserved across different subfamilies. Here we considered a site to be conserved for one subfamily if any amino acid type is represented more than 75% of the time. The sites that failed to satisfy the above criteria are marked as MC (none of the subfamilies are conserved in this site). For families with more than two subgroups, sites were categorized into different types based on the category assigned to the majority of subfamily pairs.

† ftp://ftp.ncbi.nih.gov/pub/chakraba/SPEER

Table 3. Description of the dataset

Code	Description	No. of subgroups	No. of subgroup specific site	No. of family member	Avg. sequence identity (%)
cd00120	MADS: MCM1, Agamous, Deficiens, and SRF box family.	2	3	90	12
cd00264	Bactericidal permeability-increasing protein, lipopolysaccharide-binding protein and cholesteryl ester transfer protein domains	2	3	31	8
cd00333	Major intrinsic protein (MIP) family	2	12	27	20
cd00363	Phosphofructokinase	2	6	11	35
cd00365	Hydroxymethylglutaryl-coenzyme A (HMG-CoA) reductase	2	10	30	24
cd00423	Pterin binding enzymes	2	4	33	16
cd00985	Maf_Ham1 family	2	3	180	17
Gprotein	G protein alpha subunit	11	7	105	47
GST	Glutathione *S*-transferase family	11	9	107	20
LacI	LacI/PurR family	15	12	54	27
Ricin	RICIN domain family	3	21	47	37
CNMyc	C and N terminal domain of Myc family	2	11	34	60
CBM9	Family 9 carbohydrate-binding module	2	7	19	37

Cost function to distinguish subfamily specific sites

In our approach we devise a cost function, which represents a linear combination of Euclidean distances based on amino acids' physico-chemical properties, evolution rate and combined relative entropy. All three terms account for the variability of sites within the subfamilies in terms of their physico-chemical properties, evolutionary rates and amino acid types. The first and the third terms also approximate the variability of physico-chemical properties and amino acid types between the subfamilies.

Euclidean distance based on amino acids' physico-chemical properties (ED)

Comparison of amino acids' physico-chemical properties can be very useful to characterize subtle variations in stereochemistry of subfamily specific sites. Matrices/indices containing quantitative values for amino acid physico-chemical properties (such as hydrophobicity, polarity, charge, etc.) scaled between 0 and 1 were obtained from the UMBC AAIndex database[55] (Supplementary Data, Table SM3). To quantify the variability between different amino acid properties within or between subfamilies, we employed different distance metrics that to various extents encoded the distance between subfamilies and conservation of properties within them (Supplementary Data, Figure SM1). We found that the ED-score performs best among the various metrics and is calculated as shown below. To quantify the difference between any two sequences i and j at a given site we use a weighted Euclidean distance:

$$d_{ij} = \frac{w_i w_j}{w_i + w_j} \sqrt{\sum_{m=1}^{N_m} (x_i^m - x_j^m)^2}. \quad (1)$$

Here x_i and x_j are the normalized values of the physico-chemical properties of amino acids at a given site from sequences i and j; N_m is the number of different amino acid property indices; w_i and w_j are the sequence weights of corresponding sequences.[56] The average variability of properties within the subfamilies in a given column referenced to the background variability of the whole column is estimated as follows:

$$\mathrm{ED} = \frac{\mathrm{SED}}{\mathrm{GED}}; \quad (2)$$

$$\mathrm{SED} = \sum_{s=1}^{N_s} \left\{ 1/N_p^s \sum_{i,j=1}^{N_p^s} d_{ij} \right\} \quad (3)$$

$$\mathrm{GED} = \frac{1}{N_{all}} \sum_{i,j=1}^{N_{all}} d_{ij}. \quad (4)$$

N_p^s is the number of all possible pair combinations of residues within each subfamily, N_{all} is the overall number of residue combinations in a given column and N_s is the number of subfamilies. It should be mentioned that using sequence weights together with the reference distribution of all sites in the alignment attempts to decouple subfamily specificity from the overall phylogenetic similarity of proteins in the subfamilies. The ED score is positive and its low values correspond to the situation where amino acid properties are very well conserved within the subfamilies (low SED values) and vary in between them (large GED values). The ED score equals 0 if all residues are absolutely conserved within each subfamily but different in between. For absolutely conserved (AC) columns the ED scores become undefined and such columns are excluded from the prediction procedure. Alignment columns that contain gaps are also excluded from the prediction procedure.

Evolutionary rate

Functional divergence can be inferred from the changes in the evolutionary rate at a particular site and evolutionary rate in turn can be estimated using probabilistic evolutionary models. A maximum likelihood approach allows one to estimate evolutionary rates taking into account the topology and branch lengths of the phylo-

genetic tree as well as the rate heterogeneity over different sites in a protein family. In our study we used the ML approach implemented in the rate4Site[25] program to calculate the evolution rate at each site separately for each subfamily and then average it among all subfamilies. The low average ER value would indicate that there is a slowly evolving site in certain subfamilies.

Combined relative entropy (CRE)

Relative entropy or Kullback–Leibler divergence is a very important concept in information theory and has been successfully implemented to distinguish the distributions of amino acid types between two different subfamilies.[23,28] We calculated relative entropy for each pair of subfamilies and took an average over these values at a given site:

$$\mathrm{CRE} = \frac{1}{N_{\mathrm{sp}}} \sum_{\mathrm{k,m}=1}^{N\mathrm{sp}} \sum_{x} p_k(x) \log \frac{p_k(x)}{p_m(x)}. \quad (5)$$

Here $p_{\mathrm{k}}(x)$ and $p_{\mathrm{m}}(x)$ are the probabilities to find amino acid type x in the subfamilies k and m, respectively and N_{sp} is the number of all possible combinations of subfamily pairs. The CRE is equal to zero if all distributions of p_{k} and p_{m} are the same while large values of CRE correspond to large differences between amino acid distributions of subfamilies. The relative entropy cannot be calculated if a particular type of amino acid is absent from the subfamily, such singularity is taken into account by adding pseudo counts to the calculation of probabilities p.[23]

Normalization of scores and their statistical significance

As the background conservation levels may vary substantially between different protein families we normalize each of the three scores by subtracting the mean value and dividing by the standard deviation of the score distribution obtained for all columns in a given alignment. As a pilot project we wanted to stick to equal weighing instead of putting arbitrary weights to three component terms. Determination of differential weights for ED score, CRE and ER may require a much more detailed investigation and a larger dataset to deal family specific biases for individual terms. The linear combination of three normalized scores is used to predict the specificity determinants. To calculate the statistical significance of our predictions we shuffled a given column of the alignment 100 times disregarding subfamily annotations (the procedure is similar to the one described by Mirny and Gelfand[43]). Assuming this distribution to be normal we estimate the probability that a site without the specific functional constraints would have a score equal to or higher than the observed score (P-value). The P-value assigns statistical confidence to blind predictions but the ranking of predictions with respect to P-value has not improved considerably the performance of our method (Supplementary Data, Table SM4). We think it is partially because the cost function employed in the study uses the reference distribution of non-specific sites (equation (4)).

Evaluation of prediction accuracy

We tested the performance of our method using the alignments of 13 families (Table 3) by calculating the receiver operating characteristics (ROC) curves and ROC statistics. For a given alignment, we estimated the sensitivity and error rate based on the number of true positives (known specificity sites) and false positives (non-specificity sites) found at each score cutoff. Sensitivity was defined as the fraction of true positives found at each score threshold over the overall number of true positives in the family alignment and error rate was estimated as the fraction of false positives found at same score threshold over all false positives in the alignment (difference between the total number of sites in the alignment and the number of subfamily specific sites). True positives were defined as those sites annotated as being subfamily specific based on literature and previous studies. We have evaluated the method's performance by estimating the sensitivity at 1, 5 and 15% of false positive or error rates and by calculating ROC statistics and their standard deviations.[57] A ROC_n statistic was calculated as the sum of the number of true positives found at 1,2,3, ... n false positive levels (t_{i}) divided by the overall number of true positives (T): $\mathrm{ROC}_n = (_{I=1}, \ldots, _n\, t_{\mathrm{i}})/nT$. To compare sets of ROC statistics produced by different methods we used the Wilcoxon signed rank test and calculated p-values under the null hypothesis that the medians of two distributions are equal.[58] We compared the performance of our method with three other independent methods, which predict specificity determinants: SDP-pred‡,[44] Sequence-Harmony server§[45] and the SPEL program from Pei *et al.*[47] We also analyzed and compared the performance of each method in predicting different types of subfamily specific sites (e.g. type I, type II and MC).

Acknowledgements

We thank Michael Galperin and Oishee Chakrabarti for helpful discussions and Thomas Madej for critically reading the manuscript. This work was supported by the Intramural Research Program of the National Library of Medicine at National Institutes of Health/DHHS.

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2007.08.036

References

1. Kimura, M. (1983). *The Neutral Theory of Molecular Evolution.* Cambridge University Press, Cambridge.
2. Gu, X. (1999). Statistical methods for testing functional divergence after gene duplication. *Mol. Biol. Evol.* **16**, 1664–1674.
3. Ohno, S. (1970). *Evolution by Gene Duplications.* Springer-Verlag, Berlin.
4. Doolittle, R. F. (1981). Similar amino acid sequences: chance or common ancestry? *Science*, **214**, 149–159.
5. Aharoni, A., Gaidukov, L., Khersonsky, O., S.McQ. Gould, C. & Roodveldt, D. S. (2005). The 'evolvability'

‡ http://math.genebee.msu.ru/~psn/
§ http://www.ibi.vu.nl/programs/seqharmwww

of promiscuous protein functions. *Nature Genet.* **37**, 73–76.
6. Glasner, M. E., Gerlt, J. A. & Babbitt, P. C. (2006). Evolution of enzyme superfamilies. *Curr. Opin. Chem. Biol.* **10**, 492–497.
7. Yoshikuni, Y., Ferrin, T. E. & Keasling, J. D. (2006). Designed divergent evolution of enzyme function. *Nature*, **440**, 1078–1082.
8. Gu, X. (2001). Maximum-likelihood approach for gene family evolution under functional divergence. *Mol. Biol. Evol.* **18**, 453–464.
9. Valdar, W. S. (2002). Scoring residue conservation. *Proteins: Struct. Funct. Genet.* **48**, 227–241.
10. Lockless, S. W. & Ranganathan, R. (1999). Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, **286**, 295–299.
11. Cover, T. M. & Thomas, J. A. (1991). *Elements of Information Theory.* Wiley Series in Telecommunications, New York.
12. Sander, C. & Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins: Struct. Funct. Genet.* **9**, 56–68.
13. Schneider, T. D. (1997). Information content of individual genetic sequences. *J. Theor. Biol.* **189**, 427–441.
14. Baczkowski, A. J., Joanes, D. N. & Shamia, G. M. (1998). Range of validity of alpha and beta for a generalized diversity index H (alpha, beta) due to Good. *Math. Biosci.* **148**, 115–128.
15. Taylor, W. R. (1986). The classification of amino acid conservation. *J. Theor. Biol.* **119**, 205–218.
16. Zvelebil, M. J., Barton, G. J., Taylor, W. R. & Sternberg, M. J. (1987). Prediction of protein secondary structure and active sites using the alignment of homologous sequences. *J. Mol. Biol.* **195**, 957–961.
17. Livingstone, C. D. & Barton, G. J. (1996). Identification of functional residues and secondary structure from protein multiple sequence alignment. *Methods Enzymol.* **266**, 497–512.
18. Williamson, R. M. (1995). Information theory analysis of the relationship between primary sequence structure and ligand recognition among a class of facilitated transporters. *J. Theor. Biol.* **174**, 179–188.
19. Mirny, L. A. & 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.
20. Tatusov, R. L., Altschul, S. F. & Koonin, E. V. (1994). Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks. *Proc. Natl. Acad. Sci. USA*, **91**, 12091–12095.
21. Brooks, D. J., Fresco, J. R., Lesk, A. M. & Singh, M. (2002). Evolution of amino acid frequencies in proteins over deep time: inferred order of introduction of amino acids into the genetic code. *Mol. Biol. Evol.* **19**, 1645–1655.
22. Soyer, O. S. & 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.
23. Abhiman, S., Daub, C. O. & Sonnhammer, E. L. (2006). Prediction of function divergence in protein families using the substitution rate variation parameter alpha. *Mol. Biol. Evol.* **23**, 1406–1413.
24. Mayrose, I., Graur, D., Ben-Tal, N. & Pupko, T. (2004). Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. *Mol. Biol. Evol.* **21**, 1781–1791.
25. Pupko, T., Bell, R. E., Mayrose, I., Glaser, F. & Ben-Tal, N. (2002). Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics*, **18**, S71–S77.
26. Panchenko, A. R., Kondrashov, F. & Bryant, S. H. (2004). Prediction of functional sites by analysis of sequence and structure conservation. *Protein Sci.* **13**, 884–892.
27. Lichtarge, O., Bourne, H. R. & Cohen, F. E. (1996). An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* **257**, 342–358.
28. Sjolander, K. (1998). Phylogenetic inference in protein superfamilies: analysis of SH2 domains. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **6**, 165–174.
29. Aloy, P., Querol, E., Aviles, F. X. & Sternberg, M. J. (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.
30. del Sol Mesa, A., Pazos, F. & Valencia, A. (2003). Automatic methods for predicting functionally important residues. *J. Mol. Biol.* **326**, 1289–1302.
31. Krishnamurthy, N., Brown, D. & Sjolander, K. (2007). FlowerPower: clustering proteins into domain architecture classes for phylogenomic inference of protein function. *BMC Evol. Biol.* **7**(Suppl. 1), S12–S22.
32. Jones, S. & Thornton, J. M. (1997). Analysis of protein–protein interaction sites using surface patches. *J. Mol. Biol.* **272**, 121–132.
33. Tsai, C. J., Lin, S. L., Wolfson, H. J. & Nussinov, R. (1997). Studies of protein–protein interfaces: a statistical analysis of the hydrophobic effect. *Protein Sci.* **6**, 53–64.
34. Elcock, A. H. (2001). Prediction of functionally important residues based solely on the computed energetics of protein structure. *J. Mol. Biol.* **312**, 885–896.
35. Bartlett, G. J., Porter, C. T., Borkakoti, N. & Thornton, J. M. (2002). Analysis of catalytic residues in enzyme active sites. *J. Mol. Biol.* **324**, 105–121.
36. Landgraf, R., Xenarios, I. & Eisenberg, D. (2001). Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. *J. Mol. Biol.* **307**, 1487–1502.
37. Amitai, G., Shemesh, A., Sitbon, E., Shklar, M., Netanely, D., Venger, I. & Pietrokovski, S. (2004). Network analysis of protein structures identifies functional residues. *J. Mol. Biol.* **344**, 1135–1146.
38. Rossi, A., Marti-Renom, M. A. & Sali, A. (2006). Localization of binding sites in protein structures by optimization of a composite scoring function. *Protein Sci.* **15**, 2366–2380.
39. Casari, G., Sander, C. & Valencia, A. (1995). A method to predict functional residues in proteins. *Nature Struct. Biol.* **2**, 171–178.
40. Andrade, M. A., Casari, G., Sander, C. & Valencia, A. (1997). Classification of protein families and detection of the determinant residues with an improved self-organizing map. *Biol. Cybern.* **76**, 441–450.
41. Mihalek, I., Res, I. & Lichtarge, O. (2006). Evolutionary trace report_maker: a new type of service for comparative analysis of proteins. *Bioinformatics*, **22**, 1656–1657.
42. Hannenhalli, S. S. & Russell, R. B. (2000). Analysis and prediction of functional sub-types from protein sequence alignments. *J. Mol. Biol.* **303**, 61–76.

43. Mirny, L. A. & Gelfand, M. S. (2002). Using orthologous and paralogous proteins to identify specificity-determining residues in bacterial transcription factors. *J. Mol. Biol.* **321**, 7–20.
44. Kalinina, O. V., Novichkov, P. S., Mironov, A. A., Gelfand, M. S. & Rakhmaninova, A. B. (2004). SDPpred: a tool for prediction of amino acid residues that determine differences in functional specificity of homologous proteins. *Nucl. Acids Res.* **32**, W424–W428.
45. Pirovano, W., Feenstra, K. A. & Heringa, J. (2006). Sequence comparison by sequence harmony identifies subtype-specific functional sites. *Nucl. Acids Res.* **34**, 6540–6548.
46. Donald, J. E. & Shakhnovich, E. I. (2005). Predicting specificity-determining residues in two large eukaryotic transcription factor families. *Nucl. Acids Res.* **33**, 4455–4465.
47. Pei, J., Cai, W., Kinch, L. N. & Grishin, N. V. (2006). Prediction of functional specificity determinants from protein sequences using log-likelihood ratios. *Bioinformatics*, **22**, 164–171.
48. Marttinen, P., Corander, J., Toronen, P. & Holm, L. (2006). Bayesian search of functionally divergent protein subgroups and their function specific residues. *Bioinformatics*, **22**, 2466–2474.
49. Achari, A., Somers, D. O., Champness, J. N., Bryant, P. K., Rosemond, J. & Stammers, D. K. (1997). Crystal structure of the anti-bacterial sulfonamide drug target dihydropteroate synthase. *Nature Struct. Biol.* **4**, 490–497.
50. Smith, A. E. & Mathews, R. G. (2000). Protonation state of methyltetrahydrofolate in a binary complex with cobalamin-dependent methionine synthase. *Biochemistry*, **39**, 13880–13890.
51. Hampele, I. C., D'Arcy, A., Dale, G. E., Kostrewa, D., Nielsen, J., Oefner, C. *et al.* (1997). Structure and function of the dihydropteroate synthase from Staphylococcus aureus. *J. Mol. Biol.* **268**, 21–30.
52. Minasov, G., Teplova, M., Stewart, G. C., Koonin, E. V., Anderson, W. F. & Egli, M. (2000). Functional implications from crystal structures of the conserved *Bacillus subtilis* protein Maf with and without dUTP. *Proc. Natl. Acad. Sci. USA*, **97**, 6328–6333.
53. Hwang, K. Y., Chung, J. H., Kim, S. H., Han, Y. S. & Cho, Y. (1999). Structure-based identification of a novel NTPase from *Methanococcus jannaschii*. *Nature Struct. Biol.* **6**, 691–696.
54. Marchler-Bauer, A., Anderson, J. B., DeWeese-Scott, C., Fedorova, N. D., Geer, L. Y., He, S. *et al.* (2003). CDD: a curated Entrez database of conserved domain alignments. *Nucl. Acids Res.* **31**, 383–387.
55. Bulka, B., desJardins, M. & Freeland, S. J. (2006). An interactive visualization tool to explore the biophysical properties of amino acids and their contribution to substitution matrices. *BMC Bioinformat.* **7**, 329–338.
56. Henikoff, S. & Henikoff, J. G. (1994). Position-based sequence weights. *J. Mol. Biol.* **243**, 574–578.
57. Schaffer, A. A., Aravind, L., Madden, T. L., Shavirin, S., Spouge, J. L., Wolf, Y. I. *et al.* (2001). Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucl. Acids Res.* **29**, 2994–3005.
58. Sokal, R. R. & Rohlf, F. J. (1995). *Biometry: The Principles and Practice of Statistics in Biological Research.* 3rd edit., W.H. Freeman and Co, New York.