Large-Scale Prediction of Function Shift in Protein Families with a Focus on Enzymatic Function

Saraswathi Abhiman and Erik L.L. Sonnhammer*
Center for Genomics and Bioinformatics, Karolinska Institutet, Stockholm, Sweden

ABSTRACT **Protein function shift can be predicted from sequence comparisons, either using positive selection signals or evolutionary rate estimation. None of the methods have been validated on large datasets, however. Here we investigate existing and novel methods for protein function shift prediction, and benchmark the accuracy against a large dataset of proteins with known enzymatic functions. Function change was predicted between subfamilies by identifying two kinds of sites in a multiple sequence alignment: Conservation-Shifting Sites (CSS), which are conserved in two subfamilies using two different amino acid types, and Rate-Shifting Sites (RSS), which have different evolutionary rates in two subfamilies. CSS were predicted by a new entropy-based method, and RSS using the Rate-Shift program. In principle, the more CSS and RSS between two subfamilies, the more likely a function shift between them. A test dataset was built by extracting subfamilies from Pfam with different EC numbers that belong to the same domain family. Subfamilies were generated automatically using a phylogenetic tree-based program, BETE. The dataset comprised 997 subfamily pairs with four or more members per subfamily. We observed a significant increase in CSS and RSS for subfamily comparisons with different EC numbers compared to cases with same EC numbers. The discrimination was better using RSS than CSS, and was more pronounced for larger families. Combining RSS and CSS by discriminant analysis improved classification accuracy to 71%. The method was applied to the Pfam database and the results are available at http://FunShift.cgb.ki.se. A closer examination of some superfamily comparisons showed that single EC numbers sometimes embody distinct functional classes. Hence, the measured accuracy of function shift is underestimated. Proteins 2005;60:758–768.**
© 2005 Wiley-Liss, Inc.

Key words: protein evolution; adaptive evolution; enzyme; protein function;

INTRODUCTION

Homologous proteins belonging to diverse protein families frequently evolve slightly different functions, such as different substrate specificities and activities. For example, the enzymes trypsin, chymotrypsin, and elastase are homologs that belong to the serine protease family, with a conserved catalytic triad of Asp-His-Ser. These proteins catalyze the same reaction, that is, hydrolysis of a peptide bond, but recognize and bind to different substrates and thus differ in function even though they have sequence/structural similarity and conserved catalytic residues.[1] Gene duplications giving rise to multigene families are known to create opportunities for functional divergence by allowing one gene copy to freely evolve a novel function, while the other maintains the original function.[2]

Several methods exist to predict that a gene has undergone a change in function (i.e., positive selection) using the protein-coding DNA sequence.[3–5] These methods look for positive selection along specific branches of a phylogenetic tree by estimating the ratio of nonsynonymous to synonymous nucleotide substitution rates (Ka/Ks). Ratios >1 indicate positive selection. This kind of analysis has been done on various smaller datasets, as well as on a large dataset of chordate gene families called "The Adaptive Evolution Database" (TAED).[6] However, Ka/Ks-based methods use pairwise sequence alignments[7] and are limited to closely related species, because silent substitutions become saturated over longer evolutionary timescales, in the order of 100 million years.[8]

Alternative methods exist for detecting function shifts, which use protein sequence multiple alignments to identify amino acid sites that have undergone a rate change between two subfamilies.[9–11] These methods are based on the fact that a significant rate difference at a given site between two subgroups of a protein family indicates that the function constraints at this position are different in the two groups. Detecting a large number of such positions, termed Rate Shifting Sites (RSS) in this article, suggests that the overall protein function has diverged.

Perfectly conserved positions in a family, like binding site residues, are normally essential for maintaining the function or structure. Some positions, however, may show a subfamily-specific conservation pattern, that is, conserved in all subfamilies, but using different amino acids in different subfamilies. In such cases, it is possible that the subfamilies have different substrate specificities or

Grant sponsors: the Pfizer Corporation and the Swedish Knowledge Foundation

*Correspondence to: Erik L. L. Sonnhammer, Center for Genomics and Bioinformatics, Karolinska Institutet, S-17177 Stockholm, Sweden. E-mail: erik.sonnhammer@cgb.ki.se

Received 6 July 2004; Revised 5 January 2005; Accepted 11 February 2005

Published online 6 July 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20550

© 2005 WILEY-LISS, INC.

have undergone some change in function. Positions that exhibit this conservation pattern can thus also be used as indicators of function shift. We call these positions Conservation Shifting Sites (CSS), and present a method to detect these. Previous efforts have been made to identify such functional specificity determining sites in a handful of protein families.[12–15] In this article we use the same concept but employ a simplified and normalized method that can be used to compare CSS levels of different subfamily comparisons with each other.

To apply these methods on protein families, it is necessary to first divide them into subfamilies. Often this is achieved by simple phylogenetic analysis, where subfamily membership of proteins is decided based on the inspection of a phylogenetic tree. However, it is difficult to come up with an objective cutoff that optimally divides the tree into subfamilies. Because this is also time consuming for large-scale analysis, attempts have been made to automate the process. Sjolander presented a method called BETE[14] to this end, in which subfamilies are automatically defined in a tree where each node is represented as a sequence profile of the sequences under that node. There is a cost that increases with the number of subfamilies, although there is a benefit in keeping subfamilies apart that have high relative entropy (i.e., many CSS). Starting from the leaves, nodes are joined to form subfamilies until the optimal balance between these costs is found.

Here we present an approach where we use enzyme families derived from the Pfam database and apply the BETE method for defining subfamilies. We then analyze these subfamilies with the CSS and RSS methods mentioned above as indicators of function shift between the subfamilies. The Enzyme Commission (EC) number assigned to protein sequences was used as a token of function in this study.[16] The methodology was evaluated in a large-scale test based on the Pfam database, from which we derived subfamilies pairs of two categories: (1) subfamilies with the same EC number, indicating no functional change, and (2) subfamilies with different EC numbers, indicating a functional shift. We measured the capacity of the CSS and RSS methods to separate these two categories, and also explored joining the two methods. From these results we derived optimal cutoffs for predicting function shift. These were used to predict function shifts in subfamilies in the entire Pfam database. Many previously unknown cases of function divergence were detected by this approach. The results are available at http://FunShift.cgb.ki.se.

METHODS AND DATA

Rate-Shifting Sites (RSS)

RSS were identified using the LRT program.[9] In this method the positions are analyzed individually and the program generates U-values that indicate the likelihood that there is a rate change for each alignment position between the subfamilies under consideration. A threshold cutoff of U-value 4.0 was considered significant at 5% significance level,[9] and a site is regarded as a rate-shifting site only if it is equal to or above this threshold cutoff. The RSS value for a subfamily comparison equals the percentage of rate shifting sites per alignment position. In both RSS and CSS calculations, positions that contained only gaps in a subfamily were not counted; these were detected by the hmmbuild program.[17]

Conservation-Shifting Sites (CSS)

CSS were identified using the method described by Sjolander.[14] The amino acid distribution at each position in an alignment is computed by using pseudo counts.

$$p(x) = \frac{n_x + AK_x}{N + A}$$

where n_x are the observed counts of amino acid x in a column, N is the total number of amino acids observed in the column, A is a weighting factor equal to 20, and K_x is the frequency of the amino acid x derived from the Swissprot protein sequence database.[16]

The relative entropy in one position between two subfamilies p and q is computed as

$$\mathrm{REp} = \sum_x p(x)\log\frac{p(x)}{q(x)}$$

$$\mathrm{REq} = \sum_x q(x)\log\frac{q(x)}{p(x)}$$

where $p(x)$ and $q(x)$ are the probabilities of amino acid x in subfamily p and q, respectively.

The cumulative relative entropy (CRE) between two subfamily alignments for a position then becomes

$$\mathrm{CRE} = \mathrm{REp} + \mathrm{REq}$$

The CRE metric for each position was then converted into a Z-score, computed as

$$Z = \frac{\mathrm{CRE} - \mu}{\sigma}$$

where μ is the arithmetic mean and σ is the standard deviation of the CRE values observed in all positions of the subfamily comparison.

The Z-score is a normalized method to examine the similarity between two distributions of amino acids. Smaller Z-score values are associated with similar amino acid distributions in both subfamilies, while larger Z-score values are associated with very different distributions. The absolute levels of CRE may vary substantially between families due to different conservation levels, but by using the Z-score normalization we can treat different families in the same scale. The CSS value for a subfamily comparison equals the percentage of sites with Z-score exceeding 0.5 per alignment position. The total number of positions were counted as in the RSS calculation.

Pfam Subfamilies

Data were derived from "full" alignments of Pfam protein domain families (Version 9.0)[19] with an upper limit of 500 sequences and a lower limit of eight sequences per

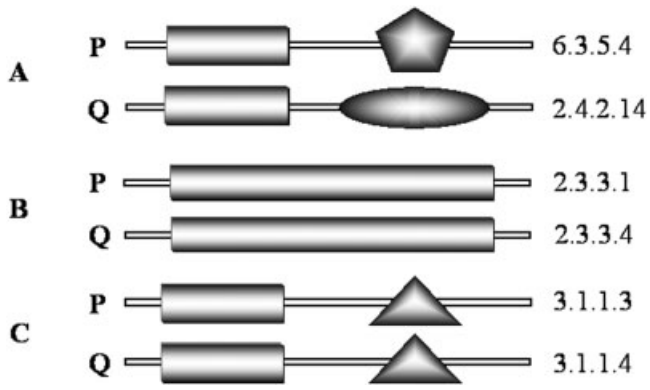

Fig. 1. Different scenarios of how different EC numbers can be present in a single Pfam family. P and Q represent two subfamilies and the Pfam domain under consideration is the rectangular domain. (**A**) subfamily P has a polygon-shaped domain responsible for EC number 6.3.5.4 and subfamily Q has an oval shaped domain responsible for EC number 2.4.2.14. All such comparisons were not considered for analysis. (**B**) Subfamilies P and Q have a single domain but have different EC numbers. (**C**) Subfamilies P and Q have more than one domain and same domain architecture but different EC numbers. Only comparisons belonging to cases B and C were analyzed.

family. This left a total of 4225 Pfam families, which were divided into 36,446 subfamilies by BETE. Only subfamilies with at least four members were analyzed (13,599). RSS and CSS values were calculated for all subfamily pairs in each family.

Subfamilies in which protein sequences had Enzyme Commission (EC) numbers in the description lines were identified, which were in turn derived from the Enzyme database.[16] A difference in EC number was used as a token for difference in function and comparisons were divided into two categories termed "Same_EC" and "Diff_EC," depending on whether both subfamilies have the same EC number or two different EC numbers for all the constituent sequences, respectively. For example, if subfamily A has EC 1.1.1.1 for all members, and Subfamily B has EC 1.1.1.2 for all its members, then such a comparison belongs to Diff_EC. Likewise, if both subfamilies A and B have EC 1.1.1.1 for all the constituent members then that comparison belongs to Same_EC category. If two different EC numbers were present within one subfamily, then it was not considered for the analysis. This was observed in 179 subfamilies.

The subfamily comparisons had to satisfy the following conditions to be considered for this analysis. (1) Both subfamilies should have a minimum of four sequences with EC numbers each, because rate-shift analysis is not meaningful for fewer sequences. (2) The EC number information should be available at all four levels. (3) All sequences should have the same domain composition. This last condition was used to eliminate potentially wrong EC numbers. The member proteins in a Pfam family can have many kinds of domains in them, and this can be the reason why more than one EC number is present in the family. Hence, we discarded all comparisons that involved members with different domain compositions, both within and across subfamilies. This is illustrated in Figure 1.

Phylogenetic Analysis of the Galactose-1-Phosphate Uridylyltransferase Family

A multiple alignment for the C terminal domain (PF02744) of Galactose-1-phosphate uridylyltransferase was obtained from Pfam, followed by manual editing to remove partial sequences and sequences not belonging to both the Gal7 and GalT subfamilies. A phylogenetic tree was inferred using the Neighbor-Joining method with observed divergence distances and 500 bootstrap replicates by the program Phylo_win.[20]

RESULTS

The aim of this analysis was to analyze the ability of the CSS and RSS methods for predicting function shift between subfamilies within a larger family. The evidence for function shift is based on the EC number annotation, assuming that any change in EC number (Diff_EC) represents a function shift while proteins with the same EC number (Same_EC) have identical functions.

A total of 206 subfamily comparisons in the Diff_EC category from 49 Pfam families, and 791 comparisons in the Same_EC category from 189 Pfam families were considered for the analysis. The subfamilies were generated from 216 Pfam families using the BETE method. In each category the comparisons were further divided into single domain and multidomain comparisons. The number of comparisons, Pfam families, subfamilies, and sequences in each of these categories are given in Table I.

RSS Distributions in the Same_EC and Diff_EC Categories

We wanted to test if there are any differences between Same_EC comparisons and Diff_EC comparisons with respect to CSS and RSS. Figure 2(a) shows the cumulative distributions of RSS values for all comparisons in the test sets. We chose to plot the cumulative distributions because it smoothes out local variations and because it provides direct information of how many comparisons are above a certain cutoff.

The maximum separation between the Same_EC and Diff_EC distributions was observed at 11% RSS. In Figure 2(a) it can be seen that 44% of the Diff_EC comparisons and 70% of the Same_EC comparisons fall below this cutoff. In other words, we can detect 56% of the function shift cases (false negative rate of 44%) while having to accept a false positive rate of 30%. Using another cutoff, say 16% RSS, we can detect 26% true positives at a 9% false positive ratio.

A possible reason for the lack of clear separation may be the presence of multidomain and small subfamilies. Therefore, the data was divided into single domain comparisons and multidomain comparisons and analyzed them separately. We observed that in the case of single domain comparisons [Fig. 2(b)] the maximum separation between the Same_EC and Diff_EC distributions increased from 26 to 35%, also at 11% RSS. In case of multidomain comparisons, a maximum separation of 19% was observed at 9% RSS.

TABLE I. Number of Comparisons, Pfam Families, Subfamilies, and Sequences for Each of the Categories Analyzed

				Sequences	
Category	Comparisons	Pfam Families	Subfamilies	Total	Unique
Diff_EC Single domain	97	29	98	4616	2029
Diff_EC Multi domain	109	20	74	3492	1216
Diff_EC combined	206	49	172	8108	3245
Same_EC Single domain	518	139	428	24283	10540
Same_EC Multi domain	273	52	192	9485	3771
Same_EC combined	791	189	620	33768	14311

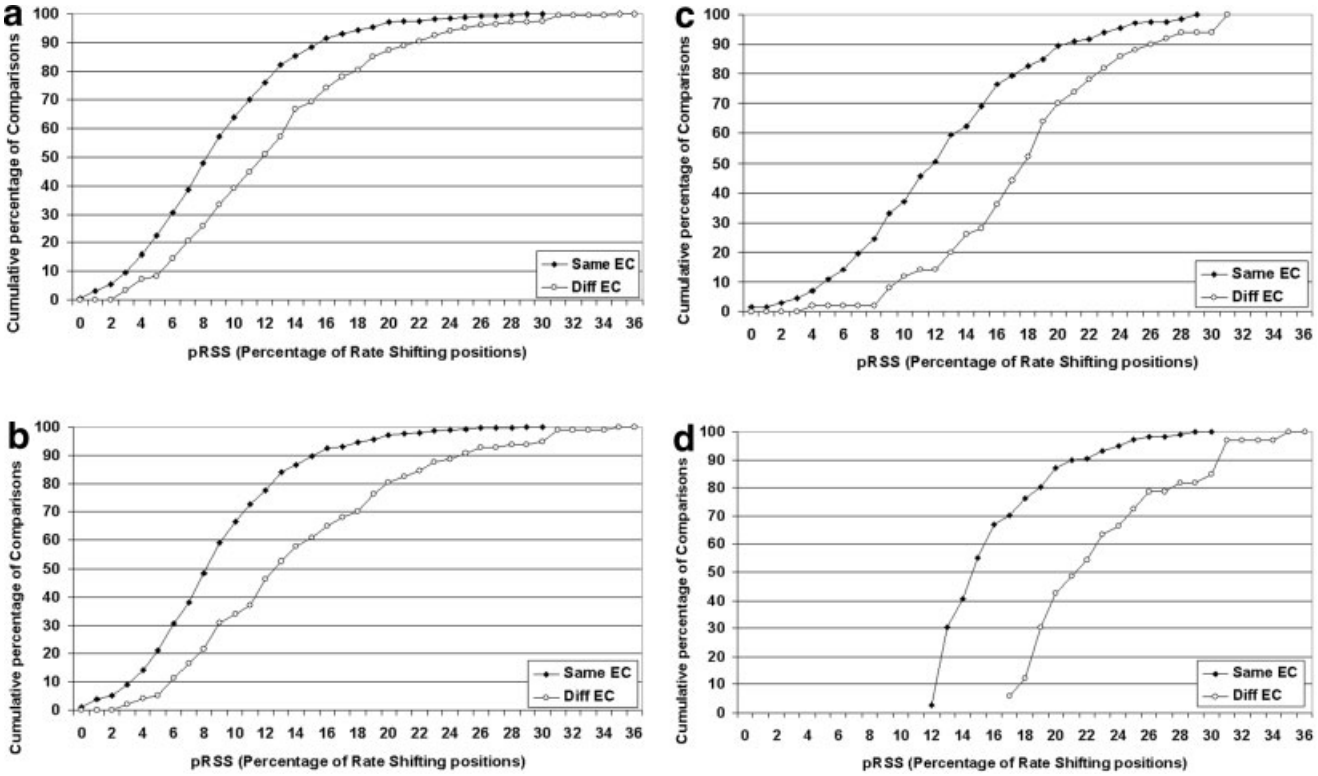

Fig. 2. Percentage of Rate-shifting sites plotted as a cumulative percentage of comparisons (**a**) for all comparisons, (**b**) for single domain comparisons, (**c**) for comparisons with larger subfamilies, and (**d**) for comparisons with large subfamilies and having single domains.

One possible interpretation for this could be that the domains in the multidomain category comparisons have higher functional constraints to resist changes because they carry out a function in combination with the domain on the same chain. In contrast, the domains in the single domain comparison category may have more relaxed functional constraints because they do not depend on a partner domain.

To investigate whether the size of subfamilies had any effect on the distribution of RSS, we took out the 25% largest subfamilies comparisons (based on the size of the smaller subfamily, which is a minimum of 14 sequences) and plotted the RSS distribution for this quartile, shown in Figure 2(c). Here the maximum separation between Diff_EC and Same_EC comparisons is further increased to 41%, at 15% RSS. For the smallest quartile the maximum separation was 23% at 5% RSS. These results show that large subfamily comparisons give better separation between Diff_EC and Same_EC categories than small subfamily comparisons. A more complete picture of the separation's dependency on the subfamily size is given in Table II. Although smaller subfamilies tend to yield smaller separation, it was still 34% for subfamilies with only four members. Thus, using four as the minimum subfamily size seems reasonable.

We then looked at the distribution of large subfamily comparisons with single domains. Figure 2(d) shows the distribution of RSS for the large subfamily comparisons with single domains. The maximum separation is here further increased to 64% at 17% RSS.

CSS Distributions in the Same_EC and Diff_EC Categories

The data for conservation shifting sites is shown in Figure 3(a). In general, it can be observed that the percentage of CSS per alignment position is higher than

TABLE II. The Maximum Separation (in Percentage Units) between the Same_EC and Diff_EC Categories as a Function of Subfamily Size for RSS and CSS

Smallest Subfamily Size	Maximum Separation for RSS	% RSS	Maximum Separation for CSS	% CSS
4	34	6	9	15
5	13	5	17	15
6	40	12	31	23
7	45	10	38	15
8	55	11	48	7
9	27	10	25	37
10	38.5	16	22	20
11	31	11	30	36
12	37.5	20	22	13
13	43.5	12	35.5	33
14	19	16	23	33
≥15	44.5	15	28	41

The columns "%RSS" and "%CSS" indicate the point where the maximum separation occurred.

RSS for a particular comparison. The maximum separation obtained was 25% at 29% CSS. The distributions of single domain category, shown in Figure 3(b), showed a slight increase in maximum separation which was 27% also at 29% CSS. Again, the multidomain category showed less separation between Same_EC and Diff_EC comparisons: 14% at 29% CSS.

Figure 3(c) shows the distribution of CSS between Same_EC and Diff_EC categories for comparisons of the 25% largest subfamilies. The maximum separation was improved to 30% at 23% CSS. For the 25% smallest subfamily comparisons, the CSS distributions of Diff_EC and Same_EC could not be differentiated. Figure 3(d) shows the distribution of CSS for large subfamily comparisons having single domains. Here the separation improved to 40% at 30% CSS. The CSS method thus yielded overall somewhat less separation between Diff_EC and Same_EC category comparisons than RSS. The same pattern was observed that large subfamilies and single domain comparisons contained most of the signal. The separation for comparisons with only four sequences in the smaller subfamily was 9% (Table II), which is not much but still adds value.

This analysis shows that there is a clear difference in the distribution of RSS and CSS between Same_EC and Diff_EC categories, with a maximum observed separation of 64% for RSS and 40% for CSS, indicating that the predictors derived from these distributions can be used to predict function change.

Predictors for Classifying New Cases of Function Shift

If the RSS and CSS methods complement each other, it may be possible to combine them to get even better separation between the Same_EC and Diff_EC comparisons. The correlation between RSS and CSS is shown in Figure 4(a) and (b) for all comparisons and large subfamily comparisons, respectively. The plots show that there is almost no correlation between RSS and CSS, yet there is an enrichment of Same_EC comparisons in the upper right area, corresponding to high RSS and CSS values. A combined approach should thus yield better separation.

To obtain a combined predictor using both RSS and CSS we used linear discriminant analysis and derived classification functions, which can be used to determine the most likely group a given case belongs to.

Two classification functions were derived, one for each group (with equal weighting), in the following form:

$$S_i = C_i + (w_{i1}v_1) + (w_{i2}v_2)$$

The subscript i denotes the group (Same_EC or Diff_EC); the subscripts 1 and 2 denote the variables (RSS or CSS). C_i is a constant; w_{i1} and w_{i2} are the weight factors; v_1 and v_2 are the arcsin transformed values of the square root[21] of RSS and CSS respectively; S_i is the resultant classification score.

The classification functions thus derived using all the comparisons are:

$$S_{Same\_EC} = (-28.45) + (19.53 \times RSS) + (38.26 \times CSS)$$

$$S_{Diff\_EC} = (-33.24) + (23.22 \times RSS) + (40.28 \times CSS)$$

The classification functions derived using large subfamily comparisons are:

$$S_{Same\_EC} = (-18.26) + (14.87 \times RSS) + (22.66 \times CSS)$$

$$S_{Diff\_EC} = (-23.12) + (19.09 \times RSS) + (24.01 \times CSS)$$

These classification functions are then used to classify a comparison into the group that obtains the highest resultant classification score. A user should choose the classification function depending on the subfamily size; for subfamilies larger than eight members, the second set of functions should be employed.

The comparisons in each category (Same_EC and Diff_EC) were divided randomly into training and test sets. The training set was used to derive classification functions, which were used on the test set to determine the accuracy of prediction, that is, what fraction of the predictions were correct. The procedure was iterated 10 times; the average accuracy in each dataset is given in Table III. A prediction accuracy of 66% was observed for the Same_EC category and 64–71% for Diff_EC category.

Database of Subfamily Alignments, Comparisons, and Function Shift Predictions

The classification functions derived were used to identify new cases of function shift by applying them to the 4225 Pfam families with multiple subfamilies containing at least four sequences. In 1280 families at least one subfamily pair was detected as function shifted (756 if using the discriminant function from large subfamily comparisons). These families include the enzyme families used as test set in this analysis. In addition, 1117 other Pfam families contained cases where function shifting was detected.

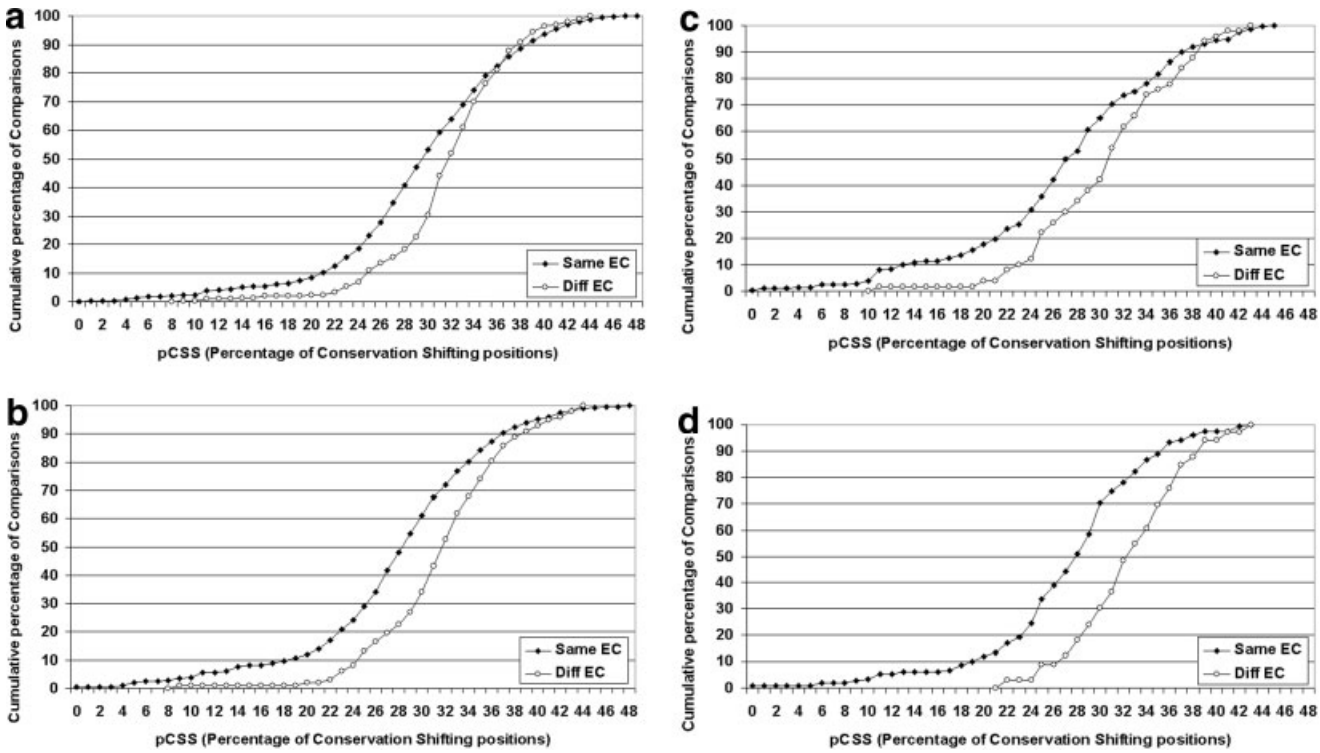

Fig. 3. Percentage of Conservation-shifting sites plotted as a cumulative percentage of comparisons (**a**) for all comparisons, (**b**) for single domain comparisons, (**c**) for comparisons with larger subfamilies, and (**d**) for comparisons with large subfamilies and having single domains.

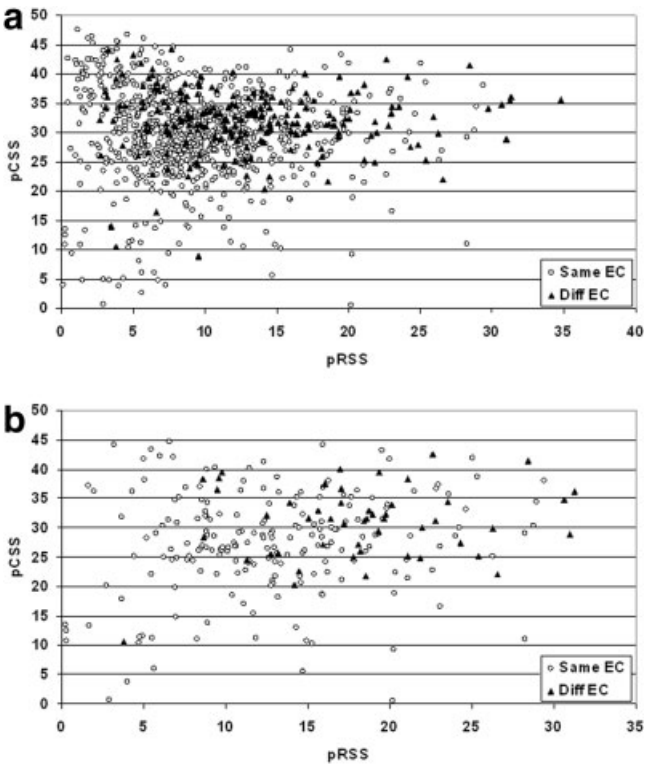

Fig. 4. Correlation between Rate-shifting sites and conservation-shifting sites shown (**a**) for all comparisons (**b**) for comparisons with larger subfamilies.

TABLE III. Average Prediction Accuracies with Standard Deviations Using Classification Functions Based on Both RSS and CSS

Category	Training Set	Test Set
All Comparisons		
Same_EC	66.2% ± 1.9	66.7% ± 2.7
Diff_EC	63.0% ± 2.4	64.4% ± 3.9
Large Subfamily Comparisons		
Same_EC	66.8% ± 4.4	65.6% ± 5.9
Diff_EC	77.3% ± 4.3	71.5% ± 6.5

For each subfamily comparison, positions were marked up as RSS or CSS when these parameters exceeded the cutoffs according to the methods described earlier. The subfamily alignments along with predictions of functional shift and RSS/CSS markup are available for browsing and download at http://FunShift.cgb.ki.se.

Analysis of Potentially Erroneous EC Classifications

As can be seen in the above results, the separation between the Same_EC and Diff_EC categories is rather poor. A particular worry is that some Same_EC comparisons had very high RSS and CSS values, in some cases as high as the highest values for Diff_EC comparisons. To investigate the reasons for this, we analyzed in detail the top 20 Same_EC comparisons with the highest RSS or CSS value from the large subfamily category in detail (see Table IV). Only 5% of the Diff_EC comparisons had a

TABLE IV. Top Ranking Same_EC Comparisons Based on Both RSS and CSS

Comparison	Domain	EC Number	Enzyme Name	Subfamily A	Subfamily B	Comments/Reference	% RSS	% CSS
PF00068_fam18-fam1	Phospholipase A2	3.1.1.4	Phospholipase A2	Subfam-18: GroupII sPLA2s of Viperidae snakes	Subfam-1: GroupI sPLA2s of Elapidae snakes	See text for details	5	42
PF00068_fam5-fam17	Phospholipase A2	3.1.1.4	Phospholipase A2	Subfam-5: Elapidae snakes (GroupI sPLA2s)	Subfam-17: Viperidae snakes and Mammals (GroupII sPLA2s)	See text for details	24	33
PF00068_fam5-fam18	Phospholipase A2	3.1.1.4	Phospholipase A2	Subfam-5: Elapidae snakes (GroupI sPLA2s)	Subfam-18: Viperidae snakes (GroupII sPLA2s)	See text for details	25	38
PF00113_fam2-fam1	Enolase, C-terminal TIM barrel domain	4.2.1.11	Phosphopyruvate hydratase	Subfam-2: Plants, Fungi, Insects and Eukaryotes	Subfam-1: Bacteria and Archaea	Enolase superfamily[31,32]	28	11
PF00148_fam5-fam37	Nitrogenase component 1 type Oxidoreductase	1.18.6.1	Nitrogenase	Subfam-5: Bacteria (Nif K family)	Subfam-37: Bacteria (Nif D family)	Ancient gene duplication event giving rise to paralogous gene family.[33]	26	25
PF00148_fam9-fam37	Nitrogenase component 1 type Oxidoreductase	1.18.6.1	Nitrogenase	Subfam-9: Bacteria (Nif K family)	Subfam-37: Bacteria (Nif D family)	Ancient gene duplication event giving rise to paralogous gene family.[33]	28	29
PF00161_fam38-fam13	Ribosome inactivating protein (RIP domain)	3.2.2.22	rRNA N-glycosylase	Subfam-38: Bacteria	Subfam-13: Viridiplantae (green plants)	Comparison between Shiga toxins of Bacteria and Type I RIP subfamily (Ricin toxins) of plants	6	44
PF00161_fam38-fam15	Ribosome inactivating protein (RIP domain)	3.2.2.22	rRNA N-glycosylase	Subfam-38: Bacteria	Subfam-15: Viridiplantae (green plants)	Comparison between Shiga toxins of Bacteria and Type I RIP subfamily (Ricin toxins) of plants	6	42
PF00161_fam39-fam13	Ribosome inactivating protein (RIP domain)	3.2.2.22	rRNA N-glycosylase	Subfam-39: Bacteria	Subfam-13: Viridiplantae (green plants)	Comparison between Shiga toxins of Bacteria and Type I RIP subfamily (Ricin toxins) of plants	5	43
PF00161_fam39-fam15	Ribosome inactivating protein (RIP domain)	3.2.2.22	rRNA N-glycosylase	Subfam-39: Bacteria	Subfam-15: Viridiplantae (green plants)	Comparison between Shiga toxins of Bacteria and Type I RIP subfamily (Ricin toxins) of plants	3	44
PF00180_fam23-fam22	Isocitrate/isopropylmalate dehydrogenase	1.1.1.42	Isocitrate dehydrogenase (NADP+)	Subfam-23: Eukaryotes and Bacteria (Gamma Proteobacteria)	Subfam-22: Bacteria (Alpha Proteobacteria) and Archaea	[34]	25	41
P F00401_fam3-fam19	ATP synthase, Delta/Epsilon chain, long alpha-helix domain	3.6.3.14	H(+)-transporting two-sector ATPase.	Subfam-3: Bacteria	Subfam-19: Viridiplantae (green plants)	Ancient duplication[35]	29	38
PF00719_fam4-fam10	Inorganic pyrophosphatase	3.6.1.1	Inorganic diphosphatase	Subfam-4: Bacteria	Subfam-10: Animals and Fungi	Differences between Plant, Bacterial and Animal/Fungi IPPases discussed in[36]	4	41

TABLE IV. (Continued)

Comparison	Domain	EC Number	Enzyme Name	Subfamily A	Subfamily B	Comments/Reference	% RSS	% CSS
PF01192_fam2-fam12	RNA polymerase Rpb6	2.7.7.6	DNA-directed RNA polymerase.	Subfam-2: Bacteria (RNAP ω chain)	Subfam-12: Eukaryotes (RNAP Rpb6) and Archaea (RNAP Rpok)	Two different subunits in different lineages. Interpro defines two separate entries.[37]	28	30
PF02502_fam3-fam2	Ribose/Galactose Isomerase	5.3.1.26	Galactose-6-phosphate isomerase	Subfam-3: Bacteria (Firmicutes)	Subfam-2: Bacteria (Firmicutes)	Subfam-3 is LacA subunit sequences and Subfam-2 is LacB subunit sequences.	19	43
PF02744_fam4-fam1	Galactose-1-phosphate uridyl transferase, C-terminal domain	2.7.7.12	UDP-glucose-hexose-1-phosphate uridylyltransferase	Subfam-4: Bacteria (Firmicutes)	Subfam-1: Archaea and Bacteria (Actinobacteria, Thermatogae, and Deinococcusthermus)	See text for details	26	26
PF02744_fam4-fam2	Galactose-1-phosphate uridyl transferase, C-terminal domain	2.7.7.12	UDP-glucose--hexose-1-phosphate uridylyltransferase	Subfam-4: Bacteria (Firmicutes)	Subfam-2: Eukaryotes and Bacteria (Proteobacteria)	See text for details	28	34
PF03118_fam9-fam11	Bacterial RNA polymerase, alpha chain C terminal domain	2.7.7.6	DNA-directed RNA polymerase	Subfam-9: Bacteria	Subfam-11: Viridiplantae (green plants).	Probable functional divergence between Bacteria and higher plants.[38]	15	44

higher RSS/CSS value than the lowest of these. When analyzing these cases we found that almost all the comparisons have an underlying difference in function between the subfamilies. Each case is different, but the common observation was that the two subfamilies stem from distinct groups that have diverged a very long time ago, often dating back to early prokaryotic evolution. One can imagine this as a protein family with two parallel groups that perform functions so similar that the EC nomenclature can not distinguish them, yet on the molecular level the differences are notable. To illustrate the nature of these differences, we will discuss the families Galactose-1-phosphate uridylyltransferase and Phospholipase A2 enzyme.

Galactose-1-phosphate uridylyltransferase

Galactose-1-phosphate uridylyltransferase catalyzes the reversible transfer of the uridine 5′-monophosphoryl moiety of UDP-glucose to the phosphate group of galactose 1-phosphate to form UDP-galactose. This enzyme (EC 2.7.7.12) participates in the Leloir pathway of galactose metabolism, and its absence is the primary cause of the potentially lethal disease galactosemia.[23] In Pfam, this enzyme is divided in two domain families, namely galactose-1-phosphate uridyl transferase, N-terminal domain (PF01087) and C-terminal domain (PF02744). Most proteins have both the N-terminal and C-terminal domain.

The proteins in these domain families fall into two distinct subfamilies, the GalT subfamily and Gal7 subfamily[24] (see Fig. 5). The GalT subfamily is limited to the

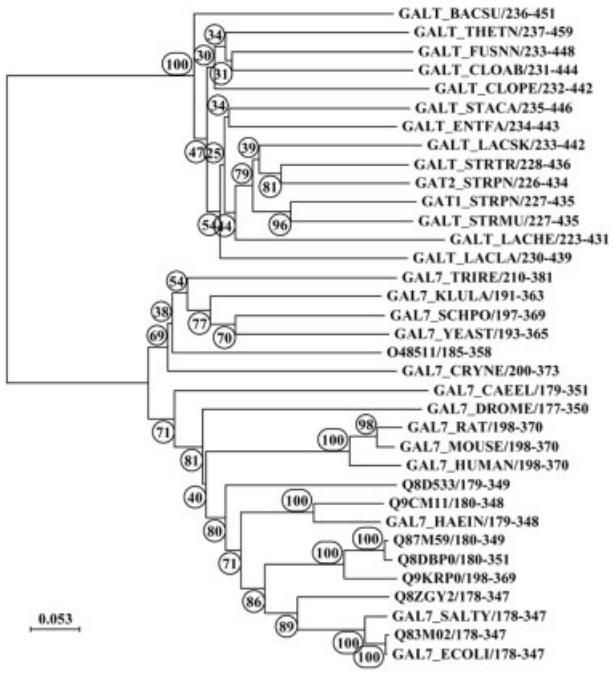

Fig. 5. Phylogeny for the galactose-1-phosphate uridylyltransferases based on the C-terminal domain (PF02744). The subfamilies of Gal7 and GalT are shown here. SWISSPROT accession numbers for the sequences are given along with the coordinates of domain boundaries. Numbers on the branches indicate bootstrap support and bootstrapping was performed with 500 replicates.

TABLE V. Conservation of Ligand and Metal Binding Residues of Galactose-1-phosphate Uridylyltransferases between Gal7 Subfamily and GalT Subfamily Are shown

S. No.	Position in 1GUP	Residue in 1GUP	Residue in Gal7 Subfamily	Residue in GalT Subfamily	Position in Pfam Full Alignment	Type of Position
N-terminal domain (PF01087)						
Ligand Binding sites						
1	28	R	R	X	44	RSS & CSS
2	31	R	R	X	47	RSS & CSS
3	32	P	P	X	48	RSS
4	33	W	W	X	71	RSS & CSS
5	54	F	X	X	115	NONE
6	60	R	R	P	125	CSS
7	61	V	V	X	126	RSS
8	75	F	F	A	149	CSS
9	77	N	N	X	151	RSS & CSS
10	78	D	D	X	152	RSS
11	79	F	F	F	153	CONSERVED
12	151	F	F	I	275	CSS
13	159	G	G	W	284	CSS
14	160	C	C	G	285	RSS & CSS
15	161	S	S	F	286	RSS & CSS
16	162	N	N	Q	287	RSS & CSS
17	168	Q	Q	I	299	CSS
18	170	W	W	L	301	CSS
Metal Binding sites (ZINC)						
19	52	C	X	X	114	NONE
20	55	C	C	M	120	CSS
21	115	H	H	K	212	CSS
22	164	H	H	S	289	CSS
C- terminal domain (PF02744)						
Ligand binding sites						
23	311	K	K	L	195	CSS
24	312	F	F	I	196	CSS
25	314	V	V	V	198	CONSERVED
26	315	G	G	G	200	CONSERVED
27	316	Y	Y	X	201	NONE
28	317	E	E	A	202	CSS
Metal Binding sites (IRON)						
29	182	E	E	X	5	RSS
30	281	H	H	D	151	CSS
31	296	H	H	H	173	CONSERVED
32	298	H	H	H	175	RSS

The ligand/metal binding residues of 3D structure for GAL7_ECOLI (PDB: 1GUP) were mapped on to the combined alignment of the subfamilies and their conservation was observed. Only the most conserved amino acid (> 80% identity) for each position is provided. X denotes that the position is variable (more than two types of amino acids exist with <30% identity).

phylum firmicutes (Gram-positive bacteria), whereas the Gal7 subfamily is a mixture of sequences belonging to the proteobacteria (purple bacteria), Gram-negative bacteria, fungi, and metazoa. It thus appears as if two types of the enzyme existed already at the split between early forms of prokaryotes.

Looking closer at the Gal7 subfamily in the tree, it is clear that it does not correspond to regular divergent evolution of species. Either it must have been transferred to some bacteria from a metazoan ancestor, or there was a duplication early in the Gal7 lineage from which one copy was selectively lost in Gram-negative bacteria and human/fly, while the other one was lost selectively in fungi/elegans.

The 3D structure is known for GAL7_ECOLI (PDB: 1GUP), a member of the Gal7 subfamily, which shows that it binds one zinc and one iron ion per subunit.[25] Table V shows the conservation pattern of the ligand binding sites for the Gal7 and GalT subfamilies for both the N-terminal and the C-terminal domains. Of the 32 binding sites, 30 are conserved (more than 80% identity) in the Gal7 subfamily, while only 21 are conserved in the GalT subfamily. Only four sites are conserved the same way in both subfamilies. This indicates that a majority of the residues required for the catalytic activity of the enzyme are under different evolutionary constraints in the GalT subfamily, and that the function is likely to be different in the two subfamilies.

Phospholipase A2

Another striking example are Phospholipases A2 (PLA2) (EC 3.1.1.4) that affect the hydrolysis of sn-2 fatty acid acyl ester bond of phopholipids (Pfam:PF00068). They are divided into two major classes, secreted form (sPLA2) and cytoplasmic form, and are further subclassified as Group I sPLA2s found in mammalian pancreas and in Elapid Snake venom, and as Group II sPLA2s found in human inflammatory fluids and in Viperid Snake venom.[26]

Comparisons between subfamilies belonging to Group I sPLA2s and Group II sPLA2s often generate very high RSS and CSS values. It has been observed that Group I sPLA2s bind calcium for activity, while Group II sPLA2s can not bind calcium.[27] They thus lack the enzymatic PLA2 activity but can serve as inhibitors or chaperones for normal sPLA2. Another observed function difference is that Group I sPLA2s are mostly neurotoxins, whereas Group II sPLA2s are myotoxic in nature.

Davidson and Dennis (1990) proposed an evolutionary scheme for sPLA2. This scheme entails a series of duplication events starting with a progenitor sPLA2, which undergoes gene duplication to produce a Group I/II precursor and a Group III (bee venom sPLA2) precursor. The Group I/II precursor is again duplicated to generate the Group I and II enzymes. As both reptiles and mammals possess various sPLA2s of both groups, all these duplications have occurred before their divergence. It also has been shown previously that PLA2s undergo accelerated rate of evolution[28] in branches leading to Viperidae and Elapidae snakes, with protein coding regions evolving faster than intronic regions. Hence, this functional divergence is supposed to have occurred before the divergence of reptiles and mammals.

The above cases demonstrate that in some families, subfamilies exist that have diverged functionally and accommodate subfamily specific functions even though they still perform the general function of the family. In general, the different forms are specific to a particular lineage, and may be related to environmental conditions. These cases also demonstrate the power of the function shift detection approach, as it identifies families where there is evidence for function divergence between subfamilies.

DISCUSSION

Previous to this study, several methods had been developed to identify positions in protein sequence alignments responsible for functional divergence, and had been applied to either single families or small datasets. Most of these studies were demonstrated on cases where a 3D structure was available and the functional divergence between the subfamilies was already known. The positions responsible for such functional divergence were then predicted and confirmed, often with the help of 3D structural information. This is the first time a large-scale study has been done on function shift between subfamilies of a protein family using only conservation signals derived from multiple sequence alignments of protein families.

In this study we have tested the hypothesis that the RSS and CSS levels between two subfamilies are correlated with function shifting. This was tested by deriving subfamilies with EC functional annotation from Pfam. It was shown that RSS and CSS levels differ substantially between subfamily comparisons in which both subfamilies have the same EC number compared to where they differ. RSS and CSS levels can thus be used as indicators of functional shift. We have selected cutoffs that perform optimally for this purpose, and applied the method to the entire Pfam database. This produced a new dataset called FunShift of subfamilies that belong to the same family but are predicted to have different function.

It was observed that the optimally discriminating CSS levels were much higher than the optimal RSS levels. We could have reduced the CSS values by raising the cutoff parameter in the CSS calculation. However, rather than striving to synchronize RSS and CSS levels, we strived to optimize the discrimination between the Same_EC and Diff_EC categories; this produced the disparity of the levels where the optimum was found.

The distributions of RSS and CSS between Diff_EC and Same_EC were not separated enough to accurately classify all the known cases. As shown in the results, one reason for this is the presence of functionally distinct subfamilies with the same EC number. We also observed that comparisons where each subfamily belongs to a different kingdom (e.g., Bacteria vs. Eukaryotes) have relatively higher percentages of RSS and CSS when compared to comparisons where both subfamilies belong to the same kingdom. A general correspondence between the subfamily size and the RSS and CSS levels was also observed. The fact that many Same_EC cases involve both mixed kingdom comparisons and large subfamilies might thus also contribute to poorer separation between the Diff_EC and Same_EC categories.

Collecting a large dataset of functionally divergent families and conserved families is a difficult task because the term “function” itself is loosely defined. We wish to emphasize that functional divergence with respect to biochemical function has only been considered, while many other types like structural or phenotypic exist. The type of reaction catalyzed by an enzyme forms the basis for assigning an EC number to the enzyme, which was used as a token of function in this study. As a consequence, all enzymes carrying out the same reaction are often grouped together under the same EC number, even if they operate by different chemical mechanisms, or occur as very different proteins in different species or cell types.[29] It is also possible that a particular protein can act as more than one type of enzyme but not all roles have been experimentally validated yet. This might be an additional reason why some of the Same_EC comparisons have very high RSS and CSS levels. Despite these and many other shortcomings in the EC numbering system, we have used this for our analysis, as no other system exists (to our knowledge) where function has been experimentally defined and can be used for systematic analysis of function.

One of the main results of this study is the FunShift database of protein subfamilies annotated with predicted function shifting sites and predicted functionally distinct subfamilies. This dataset may be used for a number of other studies. For instance, investigating the distribution of RSS and CSS residues on the 3D structure of the protein,[10] identifying function subtypes,[15] or function divergence principles.[11,30] Many of these studies have only been carried out on single protein families and will be of more general value using our dataset. Furthermore, the RSS and CSS can be used as primary candidates for site directed mutagenesis in function elucidation of proteins from laboratory experiments.

ACKNOWLEDGMENTS

We thank Bjarne Knudsen for providing the Rate-shift program, Kimmen Sjolander for providing the BETE program, and for helpful discussions. We thank David A Liberles for suggestions about our research, Markus Wistrand, and other members of Sonnhammer group for discussions.

REFERENCES

1. Krem MM, Rose T, Di Cera E. Sequence determinants of function and evolution in serine proteases. Trends Cardiovasc Med 2000;10: 171–176.
2. Lynch M, Force A. The probability of duplicate gene preservation by subfunctionalization. Genetics 2000;154:459–473.
3. Yang Z. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol 1998;15: 568–573.
4. Gu X. Statistical methods for testing functional divergence after gene duplication. Mol Biol Evol 1999;16:1664–1674.
5. Suzuki Y, Gojobori T, Nei M. ADAPTSITE: detecting natural selection at single amino acid sites. Bioinformatics 2001;17:660–661.
6. Liberles DA, Schreiber DR, Govindarajan S, Chamberlin SG, Benner SA. The adaptive evolution database (TAED). Genome Biol 2001;2:RESEARCH0028.
7. Kimura M. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. Nature 1997;267:275–276.
8. Smith JM, Smith NH. Synonymous nucleotide divergence: what is "saturation"? Genetics 1996;142:1033–1036.
9. Knudsen B, Miyamoto MM. A likelihood ratio test for evolutionary rate shifts and functional divergence among proteins. Proc Natl Acad Sci USA 2001;98:14512–14517.
10. Gu X, Vander Velden K. DIVERGE: phylogeny-based analysis for functional–structural divergence of a protein family. Bioinformatics 2002;18:500–501.
11. Knudsen B, Miyamoto MM, Laipis PJ, Silverman DN. Using evolutionary rates to investigate protein functional divergence and conservation. A case study of the carbonic anhydrases. Genetics 2003;164:1261–1269.
12. Casari G, Sander C, Valencia A. A method to predict functional residues in proteins. Nat Struct Biol 1995;2:171–178.
13. Lichtarge O, Bourne HR, Cohen FE. An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol 1996;257:342–358.
14. Sjolander K. Phylogenetic inference in protein superfamilies: analysis of SH2 domains. Proc Int Conf Intell Syst Mol Biol 1998;6:165–174.
15. Hannenhalli SS, Russell RB. Analysis and prediction of functional sub-types from protein sequence alignments. J Mol Biol 2000;303: 61–76.
16. Bairoch A. The ENZYME database in 2000. Nucleic Acids Res 2000;28:304–305.
17. Eddy SR. Profile hidden Markov models. Bioinformatics 1998;14: 755–763.
18. Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res 2000;28:45–48.
19. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ, Yeats C, Eddy SR. The Pfam protein families database. Nucleic Acids Res 2004;32D138–D141.
20. Galtier N, Gouy M, Gautier C. SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. Comput Appl Biosci. 1996;12:543–548.
21. Hogg, C. Mathematical statistics. Englewood Cliffs, NJ: Prentice Hall; 1995. p 251–252.
22. Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A. PANTHER: a library of protein families and subfamilies indexed by function. Genome Res 2003;13:2129–2141.
23. Tyfield L, Reichardt J, Fridovich-Keil J, Croke DT, Elsas LJ, 2nd, Strobl W, Kozak L, Coskun T, Novelli G, Okano Y, Zekanowski C, Shin Y, Boleda MD. Classical galactosemia and mutations at the galactose-1-phosphate uridyl transferase (GALT) gene. Hum Mutat 1999;13:417–430.
24. Mollet B, Pilloud N. Galactose utilization in Lactobacillus helveticus: isolation and characterization of the galactokinase (galK) and galactose-1-phosphate uridyl transferase (galT) genes. J Bacteriol 1991;173:4464–4473.
25. Thoden JB, Ruzicka FJ, Frey PA, Rayment I, Holden HM. Structural analysis of the H166G site-directed mutant of galactose-1-phosphate uridylyltransferase complexed with either UDP-glucose or UDP-galactose: detailed description of the nucleotide sugar binding site. Biochemistry 1997;36:1212–1222.
26. Dennis EA. Diversity of group types, regulation, and function of phospholipase A2. J Biol Chem 1994;269:13057–13060.
27. Davidson FF, Dennis EA. Evolutionary relationships and implications for the regulation of phospholipase A2 from snake venom to human secreted forms. J Mol Evol 1990;31:228–238.
28. Ohno M, Menez R, Ogawa T, Danse JM, Shimohigashi Y, Fromen C, Ducancel F, Zinn-Justin S, Le Du MH, Boulain JC, Tamiya T, Menez A. Molecular evolution of snake toxins: is the functional diversity of snake toxins associated with a mechanism of accelerated evolution? Prog Nucleic Acid Res Mol Biol 1998;59:307–364.
29. Tipton K, Boyce S. History of the enzyme nomenclature system. Bioinformatics 2000;16:34–40.
30. Gaucher EA, Miyamoto MM, Benner SA. Function–structure analysis of proteins using covarion-based evolutionary approaches: Elongation factors. Proc Natl Acad Sci USA 2001;98:548–552.
31. Gerlt JA, Babbitt PC. Mechanistically diverse enzyme superfamilies: the importance of chemistry in the evolution of catalysis. Curr Opin Chem Biol 1998;2:607–612.
32. Gerlt JA, Babbitt PC. Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies. Annu Rev Biochem 2001;70:209–246.
33. Fani R, Gallo R, Lio P. Molecular evolution of nitrogen fixation: the evolutionary history of the nifD, nifK, nifE, and nifN genes. J Mol Evol 2000;51:1–11.
34. Dean AM, Golding GB. Protein engineering reveals ancient adaptive replacements in isocitrate dehydrogenase. Proc Natl Acad Sci USA 1997;94:3104–3109.
35. Blair A, Ngo L, Park J, Paulsen IT, Saier MH, Jr. Phylogenetic analyses of the homologous transmembrane channel-forming proteins of the F0F1-ATPases of bacteria, chloroplasts and mitochondria. Microbiology 1996;142:17–32.
36. Sivula T, Salminen A, Parfenyev AN, Pohjanjoki P, Goldman A, Cooperman BS, Baykov AA, Lahti R. Evolutionary aspects of inorganic pyrophosphatase. FEBS Lett 1999;454:75–80.
37. Minakhin L, Bhagat S, Brunning A, Campbell EA, Darst SA, Ebright RH, Severinov K. Bacterial RNA polymerase subunit omega and eukaryotic RNA polymerase subunit RPB6 are sequence, structural, and functional homologs and promote RNA polymerase assembly. Proc Natl Acad Sci USA 2001;98: 892–897.
38. Sheveleva EV, Giordani NV, Hallick RB. Identification and comparative analysis of the chloroplast alpha-subunit gene of DNA-dependent RNA polymerase from seven Euglena species. Nucleic Acids Res 2002;30:1247–1254.