Analysis and prediction of functionally important sites in proteins

SAIKAT CHAKRABARTI AND CHRISTOPHER J. LANCZYCKI

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

(RECEIVED August 18, 2006; FINAL REVISION October 12, 2006; ACCEPTED October 13, 2006)

Abstract

The rapidly increasing volume of sequence and structure information available for proteins poses the daunting task of determining their functional importance. Computational methods can prove to be very useful in understanding and characterizing the biochemical and evolutionary information contained in this wealth of data, particularly at functionally important sites. Therefore, we perform a detailed survey of compositional and evolutionary constraints at the molecular and biological function level for a large set of known functionally important sites extracted from a wide range of protein families. We compare the degree of conservation across different functional categories and provide detailed statistical insight to decipher the varying evolutionary constraints at functionally important sites. The compositional and evolutionary information at functionally important sites has been compiled into a library of functional templates. We developed a module that predicts functionally important columns (FIC) of an alignment based on the detection of a significant “template match score” to a library template. Our template match score measures an alignment column’s similarity to a library template and combines a term explicitly representing a column’s residue composition with various evolutionary conservation scores (information content and position-specific scoring matrix-derived statistics). Our benchmarking studies show good sensitivity/specificity for the prediction of functional sites and high accuracy in attributing correct molecular function type to the predicted sites. This prediction method is based on information derived from homologous sequences and no structural information is required. Therefore, this method could be extremely useful for large-scale functional annotation.

Keywords: functionally important sites; function prediction; active sites; metal binding sites; protein binding sites; ligand binding sites; evolutionary conservation; compositional pattern

Supplemental material: see www.proteinscience.org

In recent decades numerous approaches have been applied to the difficult problems posed by functional site detection in proteins. Several databases, such as PROSITE (Hofmann et al. 1999), InterPro (Mulder et al. 2005), Gene Ontology (Ashburner et al. 2000), and MIPS (Mewes et al.

1999) identify and annotate protein function based on sequentially conserved residues and information extracted from experimental studies and literature searches. This information has, in turn, been used to annotate various protein alignment databases, such as Pfam (Bateman et al. 2002), HOMSTRAD (Mizuguchi et al. 1998), and CDD (Marchler-Bauer et al. 2002).

Several efforts have also been undertaken to derive functional insights using three-dimensional structural information. Authors have extensively analyzed protein structures in search of three-dimensional descriptors for functional sites (Fetrow and Skolnick 1998; Zhang et al.

Reprint requests to: Saikat Chakrabarti, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA; e-mail: chakraba@ncbi.nlm.nih.gov; fax: (301) 480-2288.

Article and publication are at <http://www.proteinscience.org/cgi/doi/10.1110/ps.062506407>.

1999; Fetrow et al. 2001) and to correlate specific structural characteristics (e.g., steric hindrance, side chain conformation, cleft size, hydrophobicity, solvent accessibility) with specific classes of protein function (Herzberg and Moulton 1991; Honig and Nicholls 1995; Laskowski et al. 1996; Heringa and Argos 1999). But these studies require extensive manual intervention and are limited by the availability of the structural information.

Other investigators have predicted functional sites using multiple sequence alignments (Casari et al. 1995; Andrade et al. 1997; Hannenhalli and Russell 2000; Li et al. 2003). Some methods of functional site prediction use spatial clustering and phylogenetic analysis to identify residues associated with functional divergence (Lichtarge et al. 1996; Sjolander 1998; Aloy et al. 2001; Madabushi et al. 2002; Chelliah et al. 2004; Innis et al. 2004; Panchenko et al. 2004). The evolutionary trace method, for example, delineates the invariant residues responsible for subgroup specificity by partitioning a dendrogram into an increasing number of subgroups of similar sequences with subsequent analysis of their three-dimensional structures (Lichtarge et al. 1996; Aloy et al. 2001; Madabushi et al. 2002). Other authors have removed the dependence on absolute conservation, using a substitution table to allow for replacement by physio-chemically similar residues (Armon et al. 2001; Pupko et al. 2002) and to allow for nonuniform rates of evolution at each site.

Despite the numerous efforts made in this field, the accuracy of functional site prediction remains low. In addition to the resourcefulness of evolutionary forces, there could be several reasons for this. One likely reason is the lack of a sufficiently extensive characterization of evolutionarily conserved patterns of amino acid residues at known functionally important sites. This is limited by additional factors, including the relative paucity of high quality multiple alignments, resources that carefully organize domain structure, and accurate identification of distantly homologous sequences.

It is generally believed that functional sites are the most highly conserved regions of a protein and are favored evolutionarily. The function of a protein can be described from complementary perspectives. The molecular function of a protein can be broadly described based on, for example, the presence of active sites (enzymes), ligand binding sites (substrate binding, nucleic acid binding, etc.) or protein-protein interaction sites. At another level, the biological function of a protein often depends on its temporal and spatial expression in the cell. Uncovering correlations between the physio-chemical properties and evolutionary conservation patterns at functionally important sites known to participate in a protein's molecular or biological function can be expected to yield important clues into characterizing similar sites in as yet uncharacterized proteins.

The present study performs such a concurrent, detailed analysis of physio-chemical properties and evolutionary conservation patterns at molecular and biological function levels for a large set of known functionally important sites extracted from a wide range of protein families. In particular, we take advantage of the manually curated multiple sequence alignment models of ancient protein families that exist in the Conserved Domain Database (CDD) (Marchler-Bauer et al. 2002). For the annotated functionally important columns (FICs) in those alignments, we can define patterns of functional group conservation within each of six molecular function categories and 16 biological functional categories, using a reduced 10-letter amino acid alphabet (Innis et al. 2004). We have used these patterns as a means to quantify the degree of conservation across different functional categories and provide specific statistical measures as a means to disentangle evolutionary constraints on functionally important sites. For example, our analysis indicates that metal binding sites and active sites are significantly more conserved than protein binding and post-translational modification sites. Similarly, we also provide statistical evidence of a variable evolutionary selection pressure across a broader set of biological functional categories.

The high-quality CDD alignments have allowed us to compute various measures of conservation that we, in turn, used to construct a library of functional templates. This library of functional templates is then used to predict functionally important sites in other families, and the molecular function associated with the best match from the library is transferred to predicted sites as a putative functional annotation. Our benchmarking studies show good sensitivity/specificity profiles for the prediction of functional sites. More interestingly, the prediction module also was able to assign the correct molecular function to predicted functional sites with high (~85%) accuracy. Consequently, as this prediction method utilizes information derived from homologous sequences, it could be extremely useful for large-scale functional annotation.

Results

Conservation at the molecular functional level

The distributions of the median conservation scores (m_c), information content (I_c) and substitution rate (R_c) for FICs are found to be distinctively different from the distribution observed for all alignment columns (including nonfunctional and functional columns) in set_210 (Supplemental Fig. SM1).

We compared the degree of conservation across six broad molecular functional categories (see Table 1) as measured by median conservation score, information

Table 1. *Categories of molecular functions*

Molecular function	Subdivision	Percentage of occurrence (%)	
		Training set of 4130 FICs (set_210)	Test set of 2918 FICs (set_96)
Active site	All active site, catalytic site residues	14	15.6
Ligand binding site	a. Nucleotide binding sites: DNA and RNA binding sites	32.3	32.9
	b. Lipid binding sites: cholesterol, glycerol, ganglioside, etc.		
	c. Carbohydrate binding sites: glucose, fructose, lactose, maltose, disaccharides, trisaccharides, etc.		
	d. Small organic ligand binding sites: ATP, ascorbate, benzamidine, butyramide, citrate, cyclosporin, FAD, FMN, hapten, heme, NAD, NADP, NADPH, oxalate, pterin, pyruvate, tetracycline, etc.		
Protein binding site	e. Inorganic ligand binding sites: CO, phosphate, and sulfur binding sites		
	a. Protein–protein interaction sites extracted from heterocomplexes	46.3	43.5
	b. Peptide binding sites		
	c. Specific protein binding sites: actin, tubulin, heparin, etc.		
Metal binding site	All metal binding residues	4.4	4.8
Post-translational modification site	a. Acetylation sites	0.6	0.8
	b. Cleavage sites		
	c. Glycosylation sites		
	d. Lipoylation sites		
	e. Phosphorylation sites		
Miscellaneous site	Sites involving structural changes, disulfide bonding residues, hinge regions, etc.	1.4	2.3

content, and substitution rate. It has been noted that active sites are among the most strongly conserved (Zvelebil et al. 1987), although some studies have pointed out exceptions where that generalization fails to hold (Todd et al. 2001, 2002). By the three quantitative estimates of conservation shown in Figure 1, the degree of conservation is, in general, found to be quite similar in active sites (A) and metal binding sites (D). Furthermore, 26% of metal binding sites in our data set appear in conjunction with at least one active site residue in a protein. Similar evolutionary selection pressures on active sites and metal binding sites has been observed in specific enzymes (Aravind et al. 2002; Boeggeman and Qasba 2002; Anantharaman et al. 2003). In aggregate, our findings encourage a more detailed investigation of the possibility of co-evolution of active sites and metal binding sites.

Post-translational modification sites and protein–protein interaction sites are found to be less well conserved by the current method. This could be due to a reduction of selection pressure in that class of functional sites, allowing them to retain broader flexibility and variability. The degree of conservation observed at nonmetal ligand binding sites, encompassing inorganic (phosphates, sulfur, carbon monoxide, etc.) and organic (nucleic acids, lipid, carbohydrate, smaller organic molecules) ligands, appears to be intermediate on our measurement scale.

Conservation at the biological functional level

Every functional site in our FIC template library has also been categorized as one of 16 broadly defined biological functions outlined in Table 2. Comparing the degree of conservation shows two rough groupings of biological function. The first group consists of DNA synthesis and processing, cellular transport and transport mechanism, intracellular communication or signal transduction, RNA processing, protein fate, and energy and metabolism functions. These biological functions are more conserved than the remaining categories, as indicated by higher median conservation scores and information content and a lower substitution rate (Fig. 2).

It is surprising to find that functional sites involved with signal transduction are among the more strongly conserved group. Considering the diversity of potential interaction partners, one would assume that the selection pressures would be mitigated to allow modification and/or substitution throughout the signaling cascade more readily. However, the high number of active site and metal binding residues (~28%, see Supplemental Fig. SM2) in this category could bias it toward an overall higher degree of conservation. Functional sites involved in cellular transport and transport mechanisms seem to have a reasonable fraction of metal binding sites (~10%), which may account for their overall higher degree of conservation.

Functions such as cellular structural organization, inter-cellular communication or cellular environment, transcription,

Figure 1. Degree of conservation across molecular functional categories. Degree of conservation of FICs is compared across six molecular functional categories (see Table 1 and text for details) based on its median conservation score m_c (a), rate of substitution R_c (b), and information content I_c (c). The central line in each box shows the median value, the upper and lower boundaries of an individual box show the upper and lower quartiles, and the vertical lines extend to a value of 1.5 times the interquartile range. Outlier values are shown outside the whiskers.

protein synthesis, development, defense stress and detoxification, cell death, and “unknown” seem to have a generally lower degree of conservation score and higher substitution rate. This is likely because sites with these functions are mainly comprised of protein binding and ligand interaction sites.

Prediction of functionally important sites

We have used our FIC library of the 4130 known functional sites extracted from set_210 to predict functionally important sites in a query multiple alignment. Further, we transfer the molecular function annotation from the most similar template FIC to the predicted site as a putative annotation. As suggested by the data in Figure 2, we note that conservation patterns across our biological functional categories possess much less specificity and, therefore, we do not transfer a putative biological function annotation as part of the prediction.

To ensure the reliability and generalization of the functional site attribution using our prediction algorithm we carried out a fivefold cross-validation test as described in Materials and Methods. The cross-validation analysis results in Table 3 suggest a high accuracy (~74%) in functional site attribution.

Next, the sensitivity and specificity of the prediction algorithm was evaluated using set_96, a separate test set

of 96 CDs (2918 FICs) not used in building the FIC template library. This set represents an unbalanced test set in that the numbers of nonfunctional sites are seven times higher than the functional sites (see test set list in Supplemental materials). The multiple sequence alignments from these 96 CDs were used as the input for the function prediction algorithm, and each column of the multiple sequence alignment is matched against the template library. The sensitivities of predicting correct functional sites at 1%, 5%, and 15% false positive rates (error rate) are given in Table 4.

Table 4 summarizes the performance of the function prediction algorithm. In the first four columns, we isolate the behavior when only a single term in the FIC template match score ($M_{s,c}$) is used. The last column provides results for the optimal combination parameter weights. The upper half of Table 4 (unshaded area) shows the sensitivity values for correctly predicting a site as functionally important regardless of the molecular function of the most similar FIC template. The lower half of the table (shaded area) provides the sensitivity values for the correct prediction of a functional site and correctly assigning its molecular function based on that of the most similar FIC template. For comparison, we show in parentheses the values obtained by random assignment of a molecular function category to a predicted functional site, based on the observed distribution of the

Table 2. *Categories of biological functions*

Serial no.	Biological functions	Percentage of occurrence (%) in training set of 4130 FICs (set_210)
1	DNA synthesis and processing	6.0
2	Cellular transport and transport mechanism	5.0
3	Intracellular communication/signal transduction	13.4
4	RNA processing	2.8
5	Protein fate	4.3
6	Energy	9.8
7	Metabolism	18.8
8	Cellular structural organization	2.5
9	Intercellular communication/cellular environment	8.2
10	Transcription	6.5
11	Protein synthesis	3.2
12	Development	1.5
13	Defense stress and detoxification	4.9
14	Cell death	1.9
15	Miscellaneous (mixed functions)	9.9
16	Unknown	0.6

six molecular function categories within our template library. Hence, compared to random molecular function assignment given a predicted functionally important site, our prediction method does well when assigning a molecular function category.

It is clear from Table 4 that the prediction sensitivity for functionally important sites improves when having contributions from all five compositional and evolutionary terms in the match score. Although the improvements of prediction sensitivity achieved by the combined strategy is moderate, they are significant and consistent across different specificity (false positive) cutoffs. Furthermore, these results indicate that locating a functional site from among a sea of nonfunctional sites in a multiple alignment is harder for our method than the prediction of the molecular function for the putative functional sites it selects: There is an $\sim 85\%$ accuracy in attributing correct function type to the predicted functional sites at a 15% error rate.

Examples of successful function predictions

Figure 3 shows several examples of the accuracy and performance of the function prediction module. In these examples, a successful prediction means that the site and its molecular function assignment were made correctly. Figure 3a shows a representative structure (response regulator PleD, PDB code 1W25, chain A) of diguanylate-cyclase (DGC) from the GGDEF domain family. Our prediction method successfully predicted eight (marked in cyan) out of 10 active sites and failed to predict two (marked in orange) of the inhibitor/ligand binding sites.

For this case, our prediction module obtains $\sim 70\%$ sensitivity while maintaining a low 12% false positive rate.

Figure 3b shows a representative structure (Phob effector domain, PDB code 1GXP, chain A) of the C-terminal effector domain (DNA binding domain) of a response regulator family of proteins, and Figure 3c shows a structure of a Ran-binding protein Mog1p (PDB code 1EQ6, chain A). Our prediction module successfully predicted 77% and 100% of the ligand and protein binding sites, again with low false positive rates of 10% and 5%, respectively.

Figure 3d provides a structure of human long-[Arg3] insulin-like growth factor 1 (PDB code 3LRI, chain A, region 17–74), which is part of the insulin/insulin-like growth factor/relaxin domain family (CDD accession no. CD00101). FICs had not yet been annotated on the

Figure 2. Degree of conservation across biological functional categories. Degree of conservation of FICs is compared across 16 biological functional categories (see Table 2 and the text for details) based on its median conservation score m_c (upper panel), rate of substitution R_c (middle panel), and information content I_c (lower panel).

Table 3. Performance of the function prediction in cross-validation test

FIC template match score (M_{sc}) threshold	Accuracy (%)
1.0	74 (2.5)
0.9	64 (3.9)
0.8	56 (1.5)
0.7	55 (1.5)
0.6	54 (1.1)
0.5	53 (2.5)
0.0	53 (3.8)

Best performing parameter weights are used: $W_g = 0.4$, $W_l = 0.2$, $W_m = 0.2$, $W_f = 0.1$, and $W_w = 0.1$.

Conserved Domain Database alignment for this family. Our function prediction module identified a number of putative functionally important protein binding sites, marked in cyan in Figure 3d. Other protein resources such as PDBsum (Laskowski et al. 1997) independently confirmed our prediction that residues Phe36, Tyr37, and Phe38 (marked in stick model in Fig. 3d) in human long-[Arg3] insulin-like growth factor 1 are indeed involved in binding to insulin-like growth factor (IGF) receptors.

Discussion

This study provides an analysis and characterization of compositional and evolutionary constraints for different molecular and biological functional categories. Despite continuing efforts in this field, a detailed understanding of the general conservation patterns at functionally important sites remains elusive. It is likely that this state of affairs adversely affects the accuracy and utility of functional site prediction methods. Similarly, the successful prediction of functional sites could be aided by the availability of high quality multiple alignments, careful organization of domain structure, and the identification of distant sequence homologies.

The curated alignments from CDD (Marchler-Bauer et al. 2002) provide an accurate representation of the conserved core of ancient protein domain families and can be used for this purpose. Importantly, each curated CDD alignment also records aligned conserved features of the family; the alignments at these functionally important sites have undergone particularly careful scrutiny by the CDD curators. Therefore, the CDD alignments are a good resource for large-scale analysis of conservation patterns across different protein functions. As demonstrated by the current study, conservation criteria extracted from this data set can be used to train successful function prediction software tools.

We have compared the degree of conservation across six different molecular functions as well as 16 different biological functions. Certainly, the definition and categorization of protein functions continues to be a debatable issue, influenced as it is by subjective decisions. On the other hand, a systematic, consistently applied categorization of protein functions does provide hope for a tractable model from which we may gain understanding of the conservation patterns displayed for different biological and/or molecular functions. Understanding the physiochemical distribution and evolutionary conservation pattern across homologous sequences at functionally important sites will help us gain insights into protein interactions with other molecules and, ultimately, improve function prediction in proteins.

A comparative analysis of conservation criteria across our six molecular functional categories indicates that metal binding sites and active sites have a notably similar evolutionary conservation pattern across a broad range of protein families and are significantly more conserved than protein binding and post-translational modification sites. However, we find that conservation patterns across our biological functional categories are much less distinctive (data not shown). This suggests a greater level of complexity and is likely due to the fact that a mixture of

Table 4. Sensitivities values estimated from ROC curves at 1%, 5%, and 15% error rates

Error rate	% TP using negative conservation scores ($W_f = 0.5$, $W_w = 0.5$)	% TP using median conservation score ^a ($W_m = 1.0$)	% TP using information content ($W_l = 1.0$)	% TP using frequency of functional group ($W_g = 1.0$)	% TP using combined method
1%	18	—	21	23	33
5%	26	—	30	38	44
15%	36	—	50	54	64
1%	10 (2 ± 0.5)	12 (—)	13 (2 ± 0.2)	18 (6 ± 1.2)	23 (4 ± 0.3)
5%	22 (5 ± 2.6)	24 (—)	27 (7 ± 0.8)	32 (10 ± 3.1)	36 (13 ± 2.7)
15%	30 (9 ± 5.4)	38 (—)	43 (10 ± 2.2)	48 (17 ± 5.3)	54 (19 ± 5.3)

The top, unshaded rows are sensitivities for correct functional site identification; lower, shaded rows are sensitivities for correct prediction of the functional site and its molecular function. Values and standard deviations within the parentheses indicate percentage of true positives obtained when molecular functions are assigned randomly based on their distribution in the template library, averaged over five independent trials. The best performing parameter weights are used for the combined method (for details, see Supplemental Material): $W_g = 0.4$, $W_l = 0.2$, $W_m = 0.2$, $W_f = 0.1$, and $W_w = 0.1$.

^a(—) Cases where sensitivity values could not be estimated from the ROC curve.

Figure 3. Examples of functionally important site prediction. (a) Cartoon representation of response regulator protein PleD (PDB code 1W25). Correctly predicted active sites and inhibitor binding sites are colored in cyan; functional sites missed by our prediction module are marked either by purple (active sites) or orange (inhibitor binding sites). (b) Structure of DNA (ligand) binding Phob effector Domain (PDB code 1GXP, chain A). (c) Structure of a Ran-binding protein Mog1p. Correctly predicted functional sites for ligand binding (b) and protein binding (c) are colored in cyan, whereas functional sites that are missed by our prediction module are marked by purple. (d) Structure of Human Long-[Arg3] insulin-like growth factor 1 (PDB code 3LRI, chain A, region 17–74). All predicted functionally important sites are colored in cyan, whereas correctly predicted sites are marked in stick representation.

molecular functions is often required to carry out a biological function.

In the bulk of this work, we utilized several quantitative measures of sequence conservation at FICs that were suggested by the initial comparative analysis to provide a measure of distinction between FICs of different molecular function. From this, we formed a library of functional templates that was used to successfully predict functionally important sites in a query alignment while maintaining a relatively low rate of false prediction. The heuristic template match score we developed evaluates potential matches between library templates and query alignment columns in terms of a “composition pattern” vector (composed of the fraction of residues in each functional group) and several scalar statistics computed from the PSSM scores at the query alignment column. Molecular functions are transferred from those of the best predicted functional template.

Our results indicated that a combination of compositional and evolutionary conservation features can lead to improved prediction of functionally important sites com-

pared to that achieved using any individual feature alone. As shown in the upper half of Table 4, we correctly identified 64% of the functionally important sites in our test set at an acceptable false positive rate of 15%. The joint identification of a functionally important site and its correct molecular function assignment at the same error rate only fell slightly below this level, to 54%. This implies that our prediction algorithm is very successful (~85%) at assigning molecular functions.

Data set incompatibility between existing function prediction programs makes it difficult to do a thorough comparison study using a single data set. Therefore, we qualitatively compare the performance of our function prediction module with some of the existing methods. The performance (in terms of sensitivity) of our prediction algorithm at a given error rate closely resembles the results of Panchenko et al. (2004), who describes a strategy of protein function prediction utilizing sequence conservation supplemented by structural information such as spatial contacts and solvent accessibility. At a 15% error rate we obtained slightly better sensitivity

(54%) compared to a weighted average of sensitivity (~52%) for the overall data set described by Panchenko et al. (2004). Our prediction sensitivity at a 1% error rate is also comparable to results obtained by other sequence and structure based methods (Chelliah et al. 2004; Vinayagam et al. 2004; Cheng et al. 2005). Overall, our benchmarking studies show good sensitivity/specificity for the prediction of functional sites and high accuracy in attributing the correct molecular function type to the predicted sites. Although the template library was constructed using curated, structure-based multiple alignments from CDD as a source, this prediction method only uses information derived from homologous sequences in a query alignment; no structural information from the query is used. Therefore, this method could be extremely useful for automated, large-scale functional annotation.

Materials and methods

Data set

We selected 210 curated domain alignments (set_210) from version 2.03 of the Conserved Domain Database (CDD): The most current version of CDD is available at <http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml> (Marchler-Bauer et al. 2002). Multiple alignments in CDD have been manually curated to reconcile sequence alignments with protein 3D structures and structure-structure alignments. A wide range of protein domain families appear in set_210, and all have at least three structural representatives. Functionally important sites (e.g., residues that participate at active sites, make contacts with ligands or macromolecules, etc.) for these families have been annotated manually by CDD curators who base their decisions on evidence available in the literature and other relevant scientific sources (Marchler-Bauer et al. 2002).

To test the performance of the function prediction algorithm, we also used a separate test set of 96 curated alignments (set_96) from CDD version 2.03 that does not overlap with set_210. Detailed lists and selection criteria for the datasets are provided in the Supplemental materials.

Categorization of the functional sites

A total of 4130 functional sites were extracted from 210 CDD multiple alignments that have been categorized by us into six molecular functions and 16 biological functions (provided in Table 1 and Table 2) based on the three-dimensional structures, literature, and experimental data annotations available for each CDD domain. These sites cover a broad range of molecular functional categories, including 534 active sites, 1303 ligand binding sites, 1985 protein-protein binding sites, 192 metal binding sites, 30 post-translational binding sites, and 86 sites with miscellaneous or mixed functions. For comparison, the corresponding breakdown by assigned molecular function for the 2918 known functionally important sites extracted from the CDs in the test set (set_96) is also provided in Table 1.

Assignment of functional groups

We assigned each of the 20 standard amino acids to one of 10 functional groups based on their physio-chemical properties. We

follow the functional group classification of Innis et al. (2004), as summarized in Supplemental Table SM1. The fraction of amino acids in functional group i at alignment position c is denoted as $g_{i,c}$ and has been calculated at all functionally important sites in our data set. If 50% or more of the sequences at one column in the alignment have an amino acid belonging to a single functional group, we assigned that as the representative functional group for the site. An eleventh group, MISC, was defined solely as the representative for those sites where none of the 10 functional groups represents at least 50% of the residues.

Calculation of sequence conservation

We used several quantitative measures to estimate the degree of conservation at each functionally important column (FIC) of the CDD alignments. Information content (I_c) of the FIC c was calculated based on counting the number of different amino acid types per aligned column and comparing with the number expected based on Robinson and Robinson (1991) background frequencies. Second, the substitution rate (R_c) for each FIC was calculated using the PAML3.12 package (Yang 1997) and its implementation of the Jones, Taylor, and Thornton amino acid substitution model (Jones et al. 1992), where the variable substitution rates across sites were described with the γ -model. Phylogenetic trees required for this analysis were constructed by the neighbor-joining method (Saitou and Nei 1987) with the PHYLIP package (Felsenstein 1997).

We also analyzed the distribution of position specific scoring matrix (PSSM) scores for FICs and non-FICs in aligned, gapless columns of a CDD alignment. At each such alignment column site c we compute the median (m_c) of the PSSM score (Chakrabarti et al. 2006) for all the amino acid residues at that position of the alignment, the frequency of occurrence of negative PSSM scores (f_c), and the relative weight of negative PSSM scores (w_c) (Chakrabarti et al. 2006). It is reasonable to expect that for FICs the frequency and relative weight of negative scores in the PSSM, which reflect unfavorable and nonconservative substitutions, should be minimized. For each site, f_c is computed simply as the ratio of the number of sequences with a negative PSSM score to the number of alignment rows, whereas w_c is computed as the absolute sum of PSSM scores for negatively-scoring residues in a column divided by the sum of the absolute value of the PSSM scores for all residues in that column (Chakrabarti et al. 2006). Although not independent measures, f_c and w_c differ in how they treat nonconserved columns: f_c penalizes PSSM columns with many slightly-negative-scoring substitutions while w_c can penalize those columns with a small number of highly-negative-scoring substitutions.

Building of the functional template library

A library of functional templates has been developed by compiling all 4130 functionally important sites from the set_210 CDD domains. For each site we define a “compositional pattern” as the set of $g_{i,c}$ for the 10 possible functional group categories i from Supplemental Table SM1, which we also represent by the data vector g_c . The quantitative portion of the functional template for FIC c is composed of compositional and evolutionary components. The former is represented by functional group composition pattern g_c , whereas the latter consists of median PSSM score m_c , information content I_c , frequency of occurrence of negative PSSM scores (f_c), and the relative weight

of negative PSSM scores (w_c). Each template is also associated with a qualitative portion, comprised of the representative functional group and the assigned molecular and biological function categories (Tables 1 and 2). Distribution of each of the representative functional groups within the six molecular function and 16 biological functions categories are provided in Supplemental materials.

To develop a good classifier of FICs, one seeks the property that a FIC from a functional category X be most similar to other FICs in category X , while being distinguishably dissimilar to FICs from another functional category Y . Therefore, we examined the relative specificity of the quantitative portion of the molecular function categories within our functional template library and found that metal binding sites, post-translational modification sites, and active site categories are more specific and discriminating as compared to protein binding sites or ligand binding sites (for details, see Supplemental materials).

Prediction of functionally important sites

An algorithm has been developed that utilizes a heuristic FIC template match score to predict functionally important sites for a given multiple sequence alignment by finding the significantly similar FICs in the template library. To compare a FIC c from the template library with a column s of an input alignment, the FIC template match score $M_{s,c}$ is computed as a linear combination of scoring terms, defined as

$$M_{s,c} = W_g D_{s,c} + \sum_{x \in \{I, m, f, w\}} W_x \sigma_{s,c}^x.$$

In the first term $D_{s,c} = 1 - \|g_s/\|g_s\| - g_c/\|g_c\|\|$ gives a high score when the Euclidean distance between the normalized compositional pattern (functional group) vectors g_s and g_c is small. The four scalar quantities I , m , f , and w described above also contribute in the second term of the match score. Each scalar quantity x makes a contribution proportional to the relative similarity between c and s , defined as $\sigma_{s,c}^x = 1 - |x_s - x_c|/\max(x_s, x_c)$.

The various W coefficients are weights constrained to sum to one, and individually may take values between zero and one. Therefore, a perfect match to a functional template has the maximal score of one. Extensive tests have been performed to determine the relative weights W_g , W_m , W_f , W_j , and W_w that provide the best accuracy for the functional site prediction. Simple sampling of the four-dimensional parameter subspace of combinations of relative parameter weights on a grid proved to be an adequate technique: The top grid points grouped together and no competing regions were found. At each of the sampled grid points a fivefold cross-validation analysis using random subset of the FIC library as test set against the remaining FICs within the library (for details, see Supplemental Table SM2) was performed.

Because multiple alignment columns that contain a single residue type (invariant sites) have roles across a wide spectrum of biological and molecular functions, we did not attempt a prediction when a single amino acid residue represented >90% of the total number in a column.

Noise estimation in match score

As noted above, $M_{s,c} = 1$ for a perfect match between alignment column s and template c . However, a simple threshold value of

$M_{s,c}$ that differentiates true and false positives is not available. For example, while one expects the information content and mean-column score terms in the definition of $M_{s,c}$ to go to zero in the case of bad matches, even for randomly aligned artificial sequences the first term involving the composition pattern vectors typically makes a positive contribution. This is because we consider only gapless columns, so there must always be at least one non-zero component in g_s .

To discount the null hypothesis of a “good” match score by chance alone, we scanned a set of 2000 columns, each containing 500 residues randomly designed according to the Robinson and Robinson (1991) background frequencies, against all FIC templates. By setting each weight W in turn equal to one, an estimate of the noise due to randomness in each term of $M_{s,c}$ can be made (for details, see Supplemental Table SM3). Match scores significantly exceeding the average background match score to the random set can be considered to fail the null hypothesis and thus identify templates with potential functional similarity to the query column s .

Evaluation of prediction accuracy

The performance of our prediction algorithm was measured by computing the average fraction of predicted true positives in a fivefold cross-validation analysis, where the 4130 FICs compiled from the set_210 CDD alignments were divided into five parts: four parts were used as template library and the remaining part as a test set. This procedure was repeated five times by randomly generating different subsets. The accuracy of the prediction was calculated as the number of correctly matched functional sites divided by the total number of predictions at a given match score threshold.

We also tested the performance of the function prediction algorithm using set_96, an entirely separate test set of 96 curated alignments from CDD version 2.03. In this case, the sensitivity–specificity analysis was performed by calculating the receiver operating characteristics (ROC) curves and ROC statistics. For a given alignment from set_96, we calculated the sensitivity and specificity based on the fraction of detected true positives and false positives at each match score cutoff. True positives were identified as those functionally important sites that had scores higher than a given match score threshold. False positives, in turn, were identified as sites with scores higher than a given threshold, but unrelated to the functional activity of a given domain family. We have evaluated the function prediction module’s performance by estimating the sensitivity at 1%, 5%, and 15% of false positive rates (specificity).

Acknowledgments

We thank Dr. Anna Panchenko for critically reading the manuscript and for providing useful suggestions, and Dr. Stephen H. Bryant for useful discussions. This work was supported by the Intramural Research Program of the National Library of Medicine at National Institutes of Health/DHHS.

References

- Aloy, P., Querol, E., Aviles, F.X., and Sternberg, M.J.E. 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.

- Anantharaman, V., Aravind, L., and Koonin, E.V. 2003. Emergence of diverse biochemical activities in evolutionarily conserved structural scaffolds of proteins. *Curr. Opin. Chem. Biol.* **7**: 12–20.
- Andrade, M.A., Casari, G., Sander, C., and Valencia, A. 1997. Classification of protein families and detection of the determinant residues with an improved self-organizing map. *Biol. Cybern.* **76**: 441–450.
- Aravind, L., Mazumder, R., Vasudevan, S., and Koonin, E.V. 2002. Trends in protein evolution inferred from sequence and structure analysis. *Curr. Opin. Struct. Biol.* **12**: 392–399.
- Armon, A., Graur, A., and Ben-Tal, N. 2001. ConSurf: An algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J. Mol. Biol.* **307**: 447–463.
- Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.P., et al. 2000. Gene ontology: Tool for the unification of biology. *Nat. Genet.* **25**: 25–29.
- Bateman, A., Birney, E., Cerruti, L., Durbin, R., Eweller, L., Eddy, S.R., Griffiths-Jones, S., Howe, K.L., Marshall, M., and Sonnhammer, E.L. 2002. The Pfam protein families database. *Nucleic Acids Res.* **30**: 276–280.
- Boeggeman, E. and Qasba, P.K. 2002. Studies on the metal binding sites in the catalytic domain of β 1,4-galactosyltransferase. *Glycobiology* **12**: 395–407.
- Casari, G., Sander, C., and Valencia, A. 1995. A method to predict functional residues in proteins. *Nat. Struct. Biol.* **2**: 171–178.
- Chakrabarti, S., Lanczycki, C.J., Panchenko, A.R., Przytycka, T.M., Thiessen, P.A., and Bryant, S.H. 2006. Refining multiple sequence alignments with conserved core regions. *Nucleic Acids Res.* **34**: 2598–2606.
- Chelliah, V., Chen, L., Blundell, T.L., and Lovell, S.C. 2004. Distinguishing structural and functional restraints in evolution in order to identify interaction sites. *J. Mol. Biol.* **342**: 1487–1504.
- Cheng, G., Qian, B., Samudrala, R., and Baker, D. 2005. Improvement in protein functional site prediction by distinguishing structural and functional constraints on protein family evolution using computational design. *Nucleic Acids Res.* **33**: 5861–5867.
- Felsenstein, J. 1997. An alternating least squares approach to inferring phylogenies from pairwise distances. *Syst. Biol.* **46**: 101–111.
- Fetrow, J.S. and Skolnick, J. 1998. Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases. *J. Mol. Biol.* **281**: 949–968.
- Fetrow, J.S., Siew, N., Di Gennaro, J.A., Martinez-Yamout, M., Dyson, H.J., and Skolnick, J. 2001. Genomic-scale comparison of sequence- and structure-based methods of functional prediction: Does structure provide additional insight? *Protein Sci.* **10**: 1005–1014.
- Hannenhalli, S.S. and Russell, R.B. 2000. Analysis and prediction of functional sub-types from protein sequence alignments. *J. Mol. Biol.* **303**: 61–76.
- Heringa, J. and Argos, P. 1999. Strain in protein structures as viewed through non-rotameric side chains: II. Effects upon ligand binding. *Proteins* **37**: 44–55.
- Herzberg, O. and Moulton, J. 1991. Analysis of the steric strain in the polypeptide backbone of protein molecules. *Proteins* **11**: 223–229.
- Hofmann, K., Bucher, P., Falquet, L., and Bairoch, A. 1999. The PROSITE database, its status in 1999. *Nucleic Acids Res.* **27**: 215–219.
- Honig, B. and Nicholls, A. 1995. Classical electrostatics in biology and chemistry. *Science* **268**: 1144–1149.
- Innis, C.A., Anand, A.P., and Sowdhamini, R. 2004. Prediction of functional sites in proteins using conserved functional group analysis. *J. Mol. Biol.* **337**: 1053–1068.
- Jones, D.T., Taylor, W.R., and Thornton, J.M. 1992. The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* **8**: 275–282.
- Laskowski, R., Luscombe, N.M., Swindells, M.B., and Thornton, J.M. 1996. Protein clefts in molecular recognition and function. *Protein Sci.* **5**: 2438–2452.
- Laskowski, R.A., Hutchinson, E.G., Michie, A.D., Wallace, A.C., Jones, M.L., and Thornton, J.M. 1997. PDBsum: A Web-based database of summaries and analyses of all PDB structures. *Trends Biochem. Sci.* **22**: 488–490.
- Li, L., Shakhnovich, E.I., and Mirny, L.A. 2003. Amino acids determining enzyme-substrate specificity in prokaryotic and eukaryotic protein kinases. *Proc. Natl. Acad. Sci.* **100**: 4463–4468.
- Lichtarge, O., Bourne, H.R., and Cohen, F.E. 1996. An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* **257**: 342–358.
- Madabushi, S., Yao, H., Marsh, M., Kristensen, D.M., Philippi, A., Sowa, M.E., and Lichtarge, O. 2002. Structural clusters of evolutionary trace residues are statistically significant and common in proteins. *J. Mol. Biol.* **316**: 139–154.
- Marchler-Bauer, A., Panchenko, A.R., Shoemaker, B.A., Thiessen, P.A., Geer, L.Y., and Bryant, S.H. 2002. CDD: A database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res.* **30**: 281–283.
- Mewes, H.W., Heumann, K., Kaps, A., Mayer, K., Pfeiffer, F., Stocker, S., and Frishman, D. 1999. MIPS: A database for genomes and protein sequences. *Nucleic Acids Res.* **27**: 44–48.
- Mizuguchi, K., Deane, C.M., Blundell, T.L., and Overington, J.P. 1998. HOMSTRAD: A database of protein structure alignments for homologous families. *Protein Sci.* **7**: 2469–2471.
- Mulder, N.J., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bradley, P., Bork, P., Bucher, P., Cerutti, L., et al. 2005. InterPro, progress and status in 2005. *Nucleic Acids Res.* **33**: D201–D205.
- Panchenko, A.R., Kondrashov, F., and Bryant, S. 2004. Prediction of functional sites by analysis of sequence and structure conservation. *Protein Sci.* **13**: 884–892.
- Pupko, T., Bell, R.E., Mayrose, I., Glaser, F., and 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.
- Robinson, A.B. and Robinson, L.R. 1991. Distribution of glutamine and asparagine residues and their near neighbors in peptides and proteins. *Proc. Natl. Acad. Sci.* **88**: 8880–8884.
- Saitou, N. and Nei, M. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**: 406–425.
- Sjolander, K. 1998. Phylogenetic inference in protein superfamilies: Analysis of SH2 domains. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **6**: 165–174.
- Todd, A.E., Orengo, C.A., and Thornton, J.M. 2001. Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.* **307**: 1113–1143.
- Todd, A.E., Orengo, C.A., and Thornton, J.M. 2002. Plasticity of enzyme active sites. *Trends Biochem. Sci.* **27**: 419–426.
- Vinayagam, A., Konig, R., Moormann, J., Schubert, F., Eils, R., Glatting, K.H., and Suhai, S. 2004. Applying support vector machines for gene ontology based gene function prediction. *BMC Bioinformatics* **5**: 116–129.
- Yang, Z. 1997. PAML: A program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13**: 555–556.
- Zhang, B., Rychlewski, L., Pawlowski, K., Fetrow, J.S., Skolnick, J., and Godzik, A. 1999. From fold predictions to function predictions: Automation of functional site conservation analysis for functional genome predictions. *Protein Sci.* **8**: 1104–1115.
- Zvelebil, M.J., Barton, G.J., Taylor, W.R., and Sternberg, M.J. 1987. Prediction of protein secondary structure and active sites using the alignment of homologous sequences. *J. Mol. Biol.* **195**: 957–961.