Assigning new GO annotations to Protein Data Bank sequences by combining structure and sequence homology

ARTICLE *in* PROTEINS STRUCTURE FUNCTION AND BIOINFORMATICS · MARCH 2005

Impact Factor: 2.63 · DOI: 10.1002/prot.20355 · Source: PubMed

Julia Ponomarenko

University of California, San Diego

Philip Bourne

National Institutes of Health

Assigning New GO Annotations to Protein Data Bank Sequences by Combining Structure and Sequence Homology

Julia V. Ponomarenko,[1,2*] **Philip E. Bourne,**[1,3] **and Ilya N. Shindyalov**[1]
[1]*San Diego Supercomputer Center, University of California, San Diego, La Jolla, California*
[2]*Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia*
[3]*Department of Pharmacology, University of California, San Diego, La Jolla, California*

ABSTRACT **Accompanying the discovery of an increasing number of proteins, there is the need to provide functional annotation that is both highly accurate and consistent. The Gene Ontology™ (GO) provides consistent annotation in a computer readable and usable form; hence, GO annotation (GOA) has been assigned to a large number of protein sequences based on direct experimental evidence and through inference determined by sequence homology. Here we show that this annotation can be extended and corrected for cases where protein structures are available. Specifically, using the Combinatorial Extension (CE) algorithm for structure comparison, we extend the protein annotation currently provided by GOA at the European Bioinformatics Institute (EBI) to further describe the contents of the Protein Data Bank (PDB). Specific cases of biologically interesting annotations derived by this method are given. Given that the relationship between sequence, structure, and function is complicated, we explore the impact of this relationship on assigning GOA. The effect of superfolds (folds with many functions) is considered and, by comparison to the Structural Classification of Proteins (SCOP), the individual effects of family, superfamily, and fold. Proteins 2005;58:855–865.** © 2005 Wiley-Liss, Inc.

Key words: Gene Ontology annotation; Protein Data Bank; three-dimensional protein structure; structure homology; sequence homology; structure comparison; protein annotation

INTRODUCTION

The ongoing process of describing the functional properties and biological roles of all proteins represents a major task of modern molecular biology. The evolving Gene Ontology™ (GO),[1,2] which standardizes this description, is vital to this process. GO provides the vocabularies (GO terms) and relationships in the form of a directed acyclic graph (DAG) for describing molecular function, biological process, and cellular localization of gene products from multiple organisms (16,687 terms as of December 8, 2003; http://www.geneontology.org/). Importantly, GO can be easy interpreted and used by computers.

GO is widely used to annotate proteins using data derived from experiments (microarrays, 2-hybrid screens, etc.), from data already present in biological databases (usually by means of literature curation) and from data derived by theoretical approaches (e.g., Jensen et al.,[3] Lagreid et al.,[4] and Letovsky and Kasif[5]). Currently, GO annotation is provided by a number of single-species oriented databases such as *Saccharomyces* Genome Database (SGD),[6] The *Arabidopsis* Information Resource (TAIR),[7] and Mouse Genome Database (MGD),[8] as well as multi-species databases such as The Institute for Genomic Research (TIGR; http://www.tigr.org), Sanger GeneDB (http://www.genedb.org), and Gene Ontology Annotation (GOA) at the European Bioinformatics Institute (EBI; http://www.ebi.ac.uk/GOA/).[9] As of December 13, 2003, TIGR provides GO annotation for 126,556 proteins and GOA EBI for 797,117 proteins; both can be freely downloaded.

Currently, the best annotation of proteins using GO is performed by highly trained biologists who read the literature and select the appropriate GO terms to be applied. Since this manual process is time-consuming and expensive, the accurate assignment of GO terms to proteins through automated extension of manual annotation is of significance. The commonly used automated approach is to infer functional similarity by establishing the presence of sequence homology to existing functionally annotated protein(s). A number of GO tools have been created that exploit this approach; see, for example, Goblet[10] and OntoBlast.[11] Compugen, Inc. has extended this basic scheme further by developing the GO Engine, which uses sequence homology, a protein clustering procedure, and text information.[12] Here we extend the sequence relationship by adding the relationship between protein structures.

The relationship between sequence, structure, and function is complicated, yet defines whether structure can be

Grant sponsor: National Institutes of Health; Grant number: GM063208.

*Correspondence to: Julia V. Ponomarenko, San Diego Supercomputer Center, University of California, San Diego, UCSD MC 0537, 9500 Gilman Drive, La Jolla, CA 92093-0537. E-mail: jpon@sdsc.edu

Received 17 February 2004; Accepted 16 September 2004

Published online 11 January 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20355

used to extend functional annotations initially derived from sequence. The relationship of sequence to structure is well characterized and indeed can be described statistically for the relationship between sequence identity and structure similarity, where the latter is defined, for example, by the root-mean-square deviation between C-α atoms.[13] The relationship between structure and function is not so straightforward.

Structure, at the level of the fold, represents a very finite set of parts[13] capable of adopting multiple functions: Triosephosphate isomerase (TIM) barrels,[14] the immunoglobulin (Ig) fold,[15] and the cupin superfamily[16] being cases in point. In fact, folds have been shown to follow a power law distribution such that some folds are promiscuous, adopting many functions, whereas others adopt one or very few functions.[17] That being said, as will be shown here, structure similarity, when used in conjunction with high levels of sequence similarity, can extend functional annotation. Such an extension requires that the functions proposed from structure agree with the consensus view of function derived from sequence. These structure assignments are in turn dependent on the parameters used to define the structure similarity. Hence, the impact of these parameters is analyzed in detail. To our knowledge, there is no existing approach that systematically attempts to exploit structure homology in proteins to reliably extend GO annotation.

The GO annotation for the three-dimensional (3D) structures available in the Protein Data Bank (PDB)[18] was analyzed starting from GO terms provided by the European Bioinformatics Institute (GOA EBI),[9] and using protein structure similarity derived from the Combinatorial Extension (CE) algorithm.[19,20] The GOA of PDB chains provided by GOA EBI comes from 3 sources: first, manual annotation provided by both EBI experts and from the InterPro resource; second, mapping of the manually annotated Swiss-Prot key words to GO terms using manually established associations between Swiss-Prot key words and GO terms; and third, mapping of the manually annotated Enzyme Commission (EC) numbers to GO terms using manually established associations between Swiss-Prot keywords and GO terms. Therefore GOA EBI could be considered as a "gold standard" for the GOA of protein chains available in the PDB, with the limitation of not having manually curated negative examples. In this work, GOA EBI is considered a starting point for extending annotation by computational means. The CE algorithm[19] performs pairwise structure alignment of proteins based only on their 3D structures, but in this work, in addition to parameters describing protein structure similarity, we used additional parameters, including a parameter characterizing the sequence similarity of structurally aligned residues (see Materials and Methods section). As will be shown empirically, this represents a major enhancement from the sequence-only similarity traditionally calculated by methods like BLAST.

It is shown that 2371 protein chains (from 4964 chains with no annotation) in the PDB have been annotated with GO terms by assigning 13,519 new GO term–chain associations at a specificity level of 90.5–99.9% (details are provided below). Further, for 1449 chains previously annotated by GOA EBI, 3962 new GO term–chain associations were added. The GO annotation of PDB protein chains reported here is available at the website http://spdc.sdsc.edu/. These data will be updated on a regular basis.

It will be shown that the proposed functional annotation has a level of consistency that varies with the parameters used. A procedure is described that validates that we have a sample sufficient to support the notion of consistency and finds the best parameters. Like all putative annotation, it does not pretend to be 100% accurate and hence should be used with caution. Despite these limitations, there are a number of applications where putative annotation can be useful, for example, (1) as a starting point to seek experimental confirmation of protein function; (2) as shown here with several cases, to validate the consistency of annotation and thereby detect errors or incompleteness in existing annotations; (3) in putative genome annotation; and (4) in global statistical studies of protein function either within the PDB or within complete proteomes.

MATERIALS AND METHODS

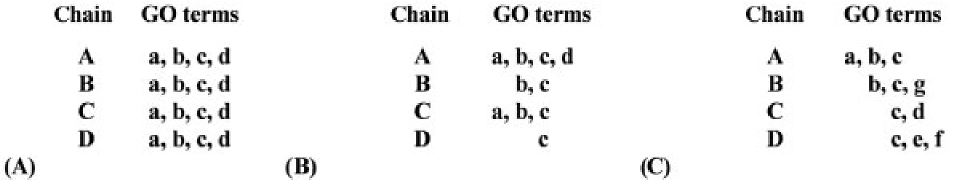
Polypeptide chains, excluding theoretical models and short chains of less than 30 C-α atoms, were taken from the PDB. An all-by-all pairwise structure alignment of these chains was calculated using CE version 2.3, which differs from the original CE algorithm (version 1.1)[19] by calculating structure alignments using not only C-α atoms but also C-β and main-chain carbonyl oxygen atoms, which provide better directionality and hence better alignments. Note that in contrast to Structural Classification of Proteins (SCOP; see Results and Discussion section), CE considers protein chains that may contain one or more SCOP domain folds. Nevertheless, as will be shown by comparison to SCOP, the process of defining extended annotation works well.

Clustering

To compare 2 protein chains using CE, 4 sequence and structure similarity parameters were used:

1. *Rmsd*—the root-mean-square deviation between 2 structurally aligned chains, which characterizes structure similarity by optimal superposition using distances between C-α, C-β, and main-chain carbonyl oxygen atoms of the aligned residues.
2. *Z-score*—the statistical score characterizing the significance of the alignment (see details[19]).
3. *Rnar*—the ratio of the number of aligned residues to the length of the shortest chain, thus measuring the overlap of the aligned protein chains relative to their length.
4. *Rseq*—the sequence identity calculated for the structurally aligned residues.

For 2 protein chains *A* and *B* with all the calculated values (*Rmsd*, *Z-score*, *Rnar*, *Rseq*) and given thresholds ($Rmsd_{threshold}$, $Z\text{-score}_{\text{threshold}}$, $Rnar_{threshold}$, $Rseq_{threshold}$),



(A)

Chain	GO terms
A	a, b, c, d
B	a, b, c, d
C	a, b, c, d
D	a, b, c, d

(B)

Chain	GO terms
A	a, b, c, d
B	b, c
C	a, b, c
D	c

(C)

Chain	GO terms
A	a, b, c
B	b, c, g
C	c, d
D	c, e, f

Fig. 1. Schematic illustrating positive clusters for the 3 specificity measures: (**A**) specificity-1; (**B**), specificity-2; (**C**), specificity-3.

we define the *sequence–structure similarity criterion* (SSC) as

$$\mathrm{SSC}_{AB} = (Rmsd < Rmsd_{threshold}) \wedge (Z\text{-}score > Z\text{-}score_{threshold}) \wedge (Rnar > Rnar_{threshold}) \wedge (Rseq > Rseq_{threshold}) \quad (1)$$

where $\wedge$ denotes a logical AND. SSC_{AB} can only be ascribed 2 values, true or false. If SSC_{AB} is true, then A and B are similar; if SSC_{AB} is false, then A and B are dissimilar.

Chains were clustered by SSC where, for every 2 chains, i and j, in the cluster, SSC_{ij} is true. Stated another way, clustering uses a combination of sequence and structure similarity as defined by SSC. GO terms for the annotated chains in a given cluster are assigned to all other chains in the cluster without GO assignments as *newly annotated chains*. These annotations were checked by a cross-validation technique described in the next section. Further, GO annotation was extended to *previously annotated chains* (chains with GO terms already assigned) by assigning new *GO term–chain* associations to those chains within a given cluster.

Specificity and Coverage

To evaluate the performance of this clustering method, the *specificity* and *coverage* of the assignments were defined. Specificity is defined here as the *positive predictive value* [$ppv = TP/(TP + FP)$] that is often applied in bioinformatics in the absence of negative examples when false and true negatives could not be measured. The number of chains with GO terms in all positive clusters is the number of true positives (TPs) and the number of chains in the negative clusters is the number of false positives (FPs). *Specificity* is calculated for chains in the cluster when there are at least 2 chains with GO assignments. We introduce 3 levels of specificity to reflect in different ways the incompleteness of GOA and the inherent limitations in the GO.

Define a set of $k(i)$ GO terms for the ith chain as $\{t_{i1}, \ldots t_{ik(i)}\}$.

Specificity-1 is the most rigorous definition, with a positive cluster defined as a cluster where every pair of chains (i, j) has the same set of GO terms:

$$t_{in} = t_{jn},\ n = 1,\ldots k(i), k(i) = k(j), \quad for\ \forall\ (i,j),\ i \in \{1,\ldots N\},\ j \in \{1,\ldots N\}. \quad (2)$$

An example of a positive cluster is given in Figure 1(A).

Specificity-2 is less rigorous that specificity-1. A positive cluster is defined such that for every pair of chains (i, j) in the cluster with a different number of GO terms, the chain with the smaller number of terms has all those terms included in the terms for the larger chain:

$$\{t_{il},..t_{ik(i)}\} \subseteq \{t_{jl},..t_{jk(j)}\},\ if\ k(i) \le k(j); \quad i \in \{1,\ldots N\}, j \in \{1,\ldots N\};\ \{t_1,..t_N\} \neq \emptyset. \quad (3)$$

An example of such a positive cluster is given in Figure 1(B).

Specificity-3 is less rigorous than specificity-2. A positive cluster is defined such that a common set of terms $\{t_1, \ldots t_L\}$ exists for all N chains within the cluster:

$$\{t_1,..t_N\} \subseteq \{t_{il},..t_{ik(i)}\},\ i = 1,\ldots N;\ \{t_1,..t_N\} \neq \emptyset. \quad (4)$$

An example of such a positive cluster is given in Figure 1(C).

Further detailing of specificity should involve estimating the semantic distance between GO terms in judging clusters to be positive or negative. The analysis of semantic distances is beyond the scope of the current work.

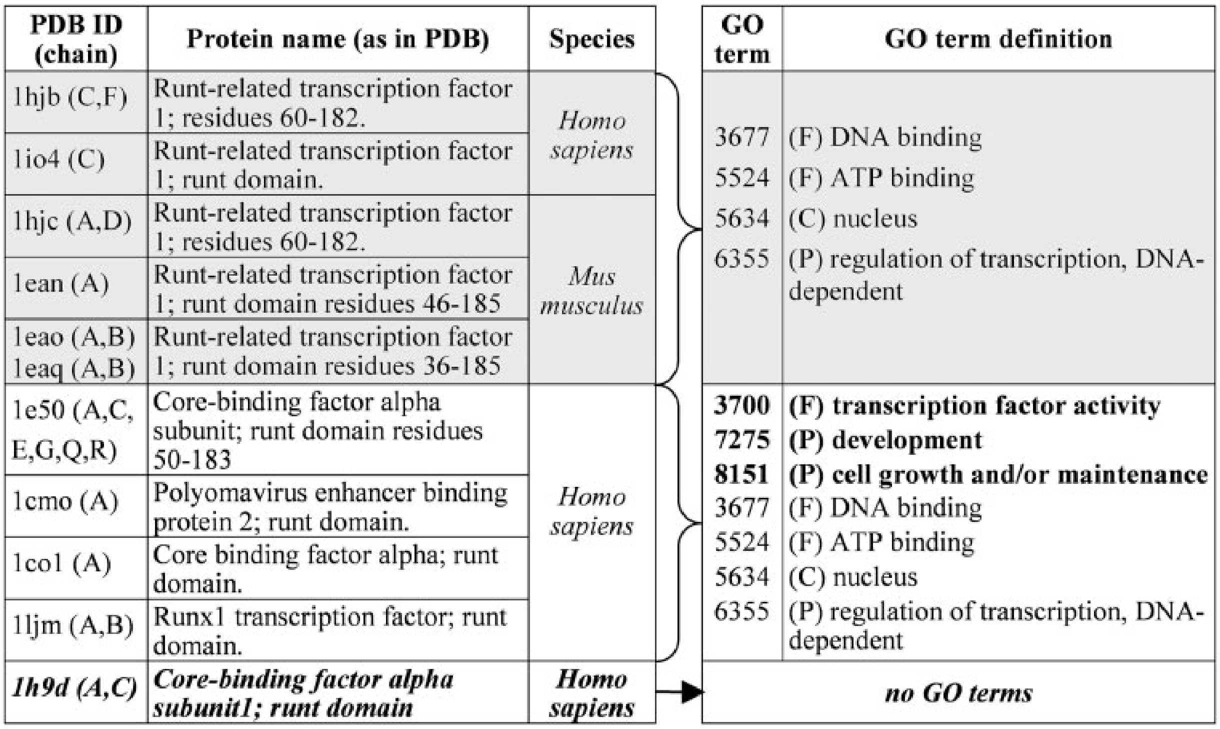
The *Coverage* of GO term assignments was defined as the ratio of newly annotated chains to all chains with no GOA EBI annotation.

Validation of the Model

The GOA EBI data set was randomly split into 2 sets without regard for the GO term assignments. The first data set was used to develop a model, that is, to select the threshold values for the 4 parameters describing structural similarity introduced above. A variant of the 10-fold cross-validation method was used[21]; that is, the first data set was randomly divided into training and validation subsets 10 different times. This variant of the 10-fold cross-validation has been used because of a limiting amount of data that precludes standard k-fold cross-validation.[22] Based on these trials, specificity and coverage of GO assignments were averaged for the 3 kinds of specificity and for the definition of coverage outlined above. The second data set was unseen and statistically independent, and was used to evaluate specificity and coverage of the chosen model.

RESULTS AND DISCUSSION

A total of 34,698 protein chains were taken from the PDB of February 2003. Existing GOAs for these PDB protein chains were taken from the GOA EBI of October



PDB ID (chain)	Protein name (as in PDB)	Species
1hjb (C,F)	Runt-related transcription factor 1; residues 60-182.	*Homo sapiens*
1io4 (C)	Runt-related transcription factor 1; runt domain.	
1hjc (A,D)	Runt-related transcription factor 1; residues 60-182.	*Mus musculus*
1ean (A)	Runt-related transcription factor 1; runt domain residues 46-185	
1eao (A,B) 1eaq (A,B)	Runt-related transcription factor 1; runt domain residues 36-185	
1e50 (A,C, E,G,Q,R)	Core-binding factor alpha subunit; runt domain residues 50-183	*Homo sapiens*
1cmo (A)	Polyomavirus enhancer binding protein 2; runt domain.	
1co1 (A)	Core binding factor alpha; runt domain.	
1ljm (A,B)	Runx1 transcription factor; runt domain.	
1h9d (A,C)	***Core-binding factor alpha subunit1; runt domain***	**Homo sapiens**

GO term	GO term definition
3677 5524 5634 6355	(F) DNA binding (F) ATP binding (C) nucleus (P) regulation of transcription, DNA-dependent
3700 **7275** **8151** 3677 5524 5634 6355	**(F) transcription factor activity** **(P) development** **(P) cell growth and/or maintenance** (F) DNA binding (F) ATP binding (C) nucleus (P) regulation of transcription, DNA-dependent
	no GO terms

Fig. 2. Example cluster. Initial annotations are based upon the GOA EBI assignment: (P), biological process; (F), molecular function and (C), cellular component. GO terms common to all chains are indicated by a gray background and in a normal font against a white background. GO terms in bold and with a gray background are new GO terms. Chains 1h9dA and 1h9dC (bold italics) are new annotations. Seven GO terms from the white background can be assigned.

15, 2003. A total of 29,734 chains from 34,698 chains had GOAs assigned. The task was to provide GOAs to as many of the protein chains as possible with no GO terms (4964 chains), and also to extend GOA to already annotated chains by assigning new GO term–chain associations.

The CE algorithm (version 2.3) was applied to calculate the pairwise structure alignments between pairs of chains. Four parameters reported by CE were considered (see Materials and Methods section). The chains were clustered such that for every 2 chains in each cluster, the SSC [Eq. (1)] was true. A single chain cluster was considered a *singleton*. A number of different empirical choices for the threshold values [Eq. (1)] were systematically considered:

$Rmsd_{threshold} \in \{2.0\ \text{Å}, 3.0\ \text{Å}, 4.0\ \text{Å}, 5.0\ \text{Å}\}$
$Z\text{-}score_{threshold} \in \{3.8, 4.0, 4.3, 4.5\}$
$Rnar_{threshold} \in \{70\%, 80\%, 90\%\}$
$Rseq_{threshold} \in \{0\%, 25\%, 35\%, 50\%, 70\%, 90\%\}$

Each choice of threshold reflects a somewhat different notion of structure or sequence similarity, and the resulting clustering shows a number of different outcomes (see section on selection of the model with respect to similarity criteria).

Figure 2 is an example of a cluster consisting of 22 protein chains based on an *Rmsd* of 5 Å, a *Z-score* of 3.8, an *Rnar* of 70%, and an *Rseq* of 90%. Ten chains are assigned 4 GO terms (gray background); another 10 chains are assigned 7 GO terms, and the remaining 2 chains have no GO term assignments (bold italics). Following the protein names (taken from the PDB), it can be seen that the cluster contains the same domain described as runt-related transcription factor 1, and which is also described in the PDB as core-binding factor alfa, polyoma virus enhancer-binding protein, and runx1 transcription factor. Four GO terms—3677, DNA-binding for function; 5524, adenosine triphosphate (ATP)-binding for function; 5634, nucleus for cellular component; and 6355, regulation of transcription—are assigned. DNA-dependence as a biological process is assigned to all chains. Another 3 GO terms—3700, transcription factor activity for function; 7275, development for process; and 8151, cell growth and/or maintenance for process—are assigned to only half of the proteins in the cluster (white background in Fig. 2). This cluster is defined as negative by the definition of specificity-1 and as positive by the definitions of specificity-2 and specificity-3. There are missed GO terms (bold in Fig. 2) for 10 chains. Seven GO terms could be assigned to chains 1h9dA and 1h9dC as newly annotated chains. Finally, GOA could be extended for 10 previously annotated chains (gray box, Fig. 2) by assigning 3 new GO terms to each of them, adding 30 new GO term–chain associations.

Selection of the Model With Respect to Similarity Criteria

The specificity and coverage values based on the method described above are presented in Table I. Consider the computation of one such value—47.38% for specificity-1

TABLE I. Specificity at Different Threshold Values

Threshold values				Performance on the training set[a]								Performance on the test set			
Rseq, %	Rnar, %	Rmsd, Å	Z-score	Av. specificity-1, %	Std. specificity-1, %	Av. specificity-2, %	Std. specificity-2, %	Av. specificity-3, %	Std. specificity-3, %	Av. coverage, %	Std. coverage, %	Av. specificity-1, %	Av. specificity-2, %	Av. specificity-3, %	Coverage, %
0	90	2.0	4.5	51.00	(1.84)	67.75	(1.68)	97.55	(0.36)	69.07	(0.43)	47.38	64.50	96.82	68.76
25	90	2.0	4.5	52.22	(1.66)	70.17	(1.25)	98.29	(0.22)	67.99	(0.46)	48.98	67.79	97.85	67.61
35	90	2.0	4.5	56.50	(1.36)	74.55	(1.00)	99.28	(0.15)	67.76	(0.69)	53.85	72.49	98.98	66.80
50	90	2.0	4.5	70.51	(1.92)	86.40	(0.94)	99.56	(0.09)	66.93	(1.71)	66.54	84.36	99.37	65.09
70	90	2.0	4.5	81.07	(1.40)	94.54	(0.41)	99.59	(0.08)	56.23	(0.95)	78.76	93.66	99.43	56.73
90	*90*	*2.0*	*4.5*	*92.17*	*(0.60)*	*97.22*	*(0.08)*	*99.95*	*(0.01)*	*38.23*	*(0.84)*	*90.85*	*97.10*	*99.94*	*39.29*
0	70	5.0	3.8	31.80	(2.47)	43.71	(2.83)	79.59	(1.25)	81.02	(1.19)	26.73	38.13	77.06	79.73
25	70	5.0	3.8	46.06	(1.85)	64.01	(1.72)	97.98	(0.29)	76.67	(1.77)	42.19	60.80	97.42	73.16
35	70	5.0	3.8	53.76	(1.48)	72.20	(1.13)	99.23	(0.16)	74.09	(1.47)	50.91	69.92	98.93	71.22
50	70	5.0	3.8	69.15	(2.15)	86.18	(1.10)	99.51	(0.08)	69.12	(1.04)	64.76	83.79	99.36	68.58
70	70	5.0	3.8	80.41	(1.32)	94.38	(0.40)	99.55	(0.08)	59.97	(0.98)	78.21	93.58	99.42	58.36
90	**70**	**5.0**	**3.8**	**91.83**	**(0.58)**	**97.09**	**(0.08)**	**99.95**	**(0.01)**	**39.76**	**(0.91)**	**90.58**	**97.02**	**99.94**	**40.25**

[a]Average (Av.) and standard deviation (Std.) values were computed over 10 random samples while dividing the whole set into 2 equal subsets for training and testing.

using the test set. This value is based on 17,349 chains (half of the sample). Only chains that have GO terms and for which the clusters have 2 or more GO terms are considered. There were 430 clusters with 2041 chains [true positives (*TP*) + false positives (*FP*)]. The number of these clusters and chains is 210 and 967, respectively, according to the *TP* criteria for specificity-1. Hence, the specificity value is [*TP*/(*TP* + *FP*)]*100 − 47.38%. The results for the test set were within 2–3 standard deviations for all combinations of the threshold values. This implies we have a sufficient and unbiased sample from which to choose the optimal threshold parameters.

The sets of parameters were selected to provide the highest specificity and the best coverage at the highest specificity levels. The maximum coverage of 79.7% (Table I, threshold values of $Rmsd \leq 5.0$ Å, $Z\text{-}score \geq 3.8$, $Rnar \geq 70\%$, $Rseq \geq 0\%$) yields a specificity of 26.7% according to the definition of specificity-1, 38.1% for specificity-2, and 77.1% for specificity-3. The best specificity-3 and coverage combination is achieved for the parameter set given in bold in Table 1, namely, a specificity-3 of 99.94% and coverage of 40.2%. For specificity-2, the highest level of specificity of 97.1% was achieved with coverage of 39.3%, and for specificity-1, 90.85% specificity at coverage of 58%. For further study we chose parameter threshold values of $Rmsd \leq 5.0$ Å, $Z\text{-}score \geq 3.8$, $Rnar \geq 70\%$, $Rseq \geq 90\%$ to provide the maximum coverage at a specificity-3 of 99.94% (Table I, bold row, "Performance on the test set" columns).

Extending GO Annotation Within the PDB: Summary of the Best Annotations

Consider the extension of GOA according to the chosen threshold values for structure–sequence similarity: $Rmsd \leq 5.0$ Å, $Z\text{-}score \geq 3.8$, $Rnar \geq 70\%$, and $Rseq \geq 90\%$ (bold row, Table I). Clustering of 34,698 protein chains produced 2762 singletons and 4269 clusters, of which 3679 clusters (27,448 chains) had more than 2 chains per cluster with assigned GOA EBI annotation. This annotation was extended to assign 13,519 GO terms to 2371 previously unannotated chains, leaving 2593 unannotated (Fig. 3). Further, 1449 previously annotated chains had 3962 new GO terms added. A total of 571 from 4964 chains (2593 + 2371) could not be annotated by this method, since they are singletons; hence, the coverage for the newly annotated chains is 54% [2371/(4964 − 571)].

Consider clusters of chains with contradictory GO annotation, that is, negative or false-positive clusters. There were only 5 such clusters from a total of 3679 according to specificity-3 (Table II). The first 3 clusters and the last cluster are examples of incomplete annotation (missing GO terms) as indicated by their PDB compound names. Hence, for 2mtaC, 2 new GO terms can be added: 16032, viral life cycle for biological process; and 16021, integral to membrane for cellular component. For the following protein chains—1mg2D, 1mg2H, 1mg2L, 1mg2P, 1mg3D, 1mg3H, 1mg3L, and 1mg3P—there are 3 new GO terms that can be added: 5489, electron transporter activity for molecular function; 6118, electron transport for biological process; and 15945, methanol metabolism for biological

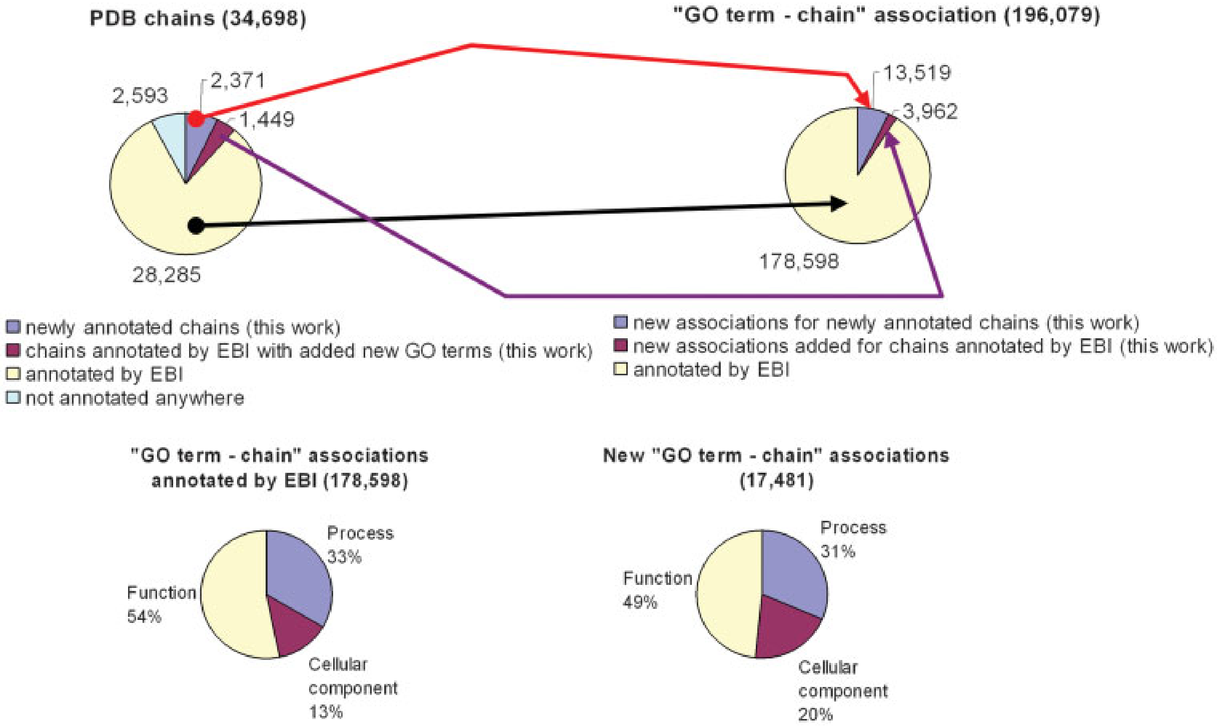


Fig. 3. Annotation results with *Rmsd* ≤ 5.0 Å, *Z-score* ≥ 3.8, *Rnar* ≥ 70%, and *Rseq* ≥ 90%.

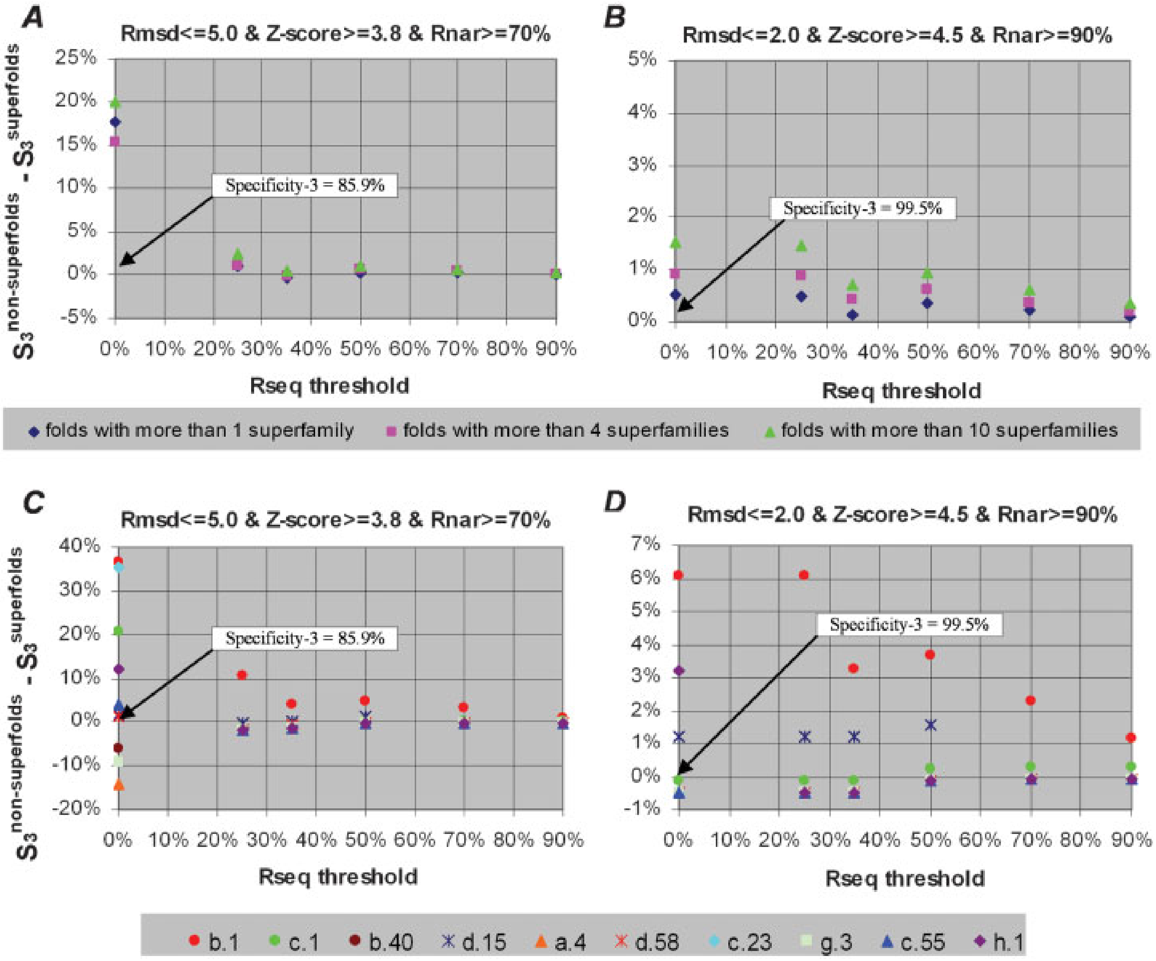


Fig. 5. Impact of superfolds. $S_3^{nonsuperfolds}$ represents SCOP nonsuperfolds (1 superfamily per fold) at specificity-3, and superfolds $S_3^{superfolds}$ represents superfolds (more than 1 superfamily per fold) at specificity-3. (**A**, **B**) Impact of all superfolds with different similarity criteria. (**C**, **D**) Impact of specific superfolds with different similarity criteria.

TABLE II. Clusters With Contradictory GO Annotations With Structure and Sequence Similarity Parameters: Rmsd ≤ 5.0 Å, Z-score ≥ 3.8, Rnar ≥ 70%, Rseq ≥ 90%

	GO terms	GO term definition[a]	PDB ID & chain	Protein name	Species
1	5489	(F) electron transporter activity	2mtaC	Cytochrome c551i	Paracoccus denitrificans
	6118	(P) electron transport			
	15945	(P) methanol metabolism			
	16032	(P) viral life cycle	1mg2D, mg2H, 1mg2L, 1mg2P	Cytochrome c-L (cytochrome c551i)	
	16021	(C) integral to membrane	1mg3D, mg3H, 1mg3L, 1mg3P	Cytochrome c-L (cytochrome c551i)	
2			1hp4A	β-*n*-acetylhexosaminidase	
	5975	(P) carbohydrate metabolism	1hp5A	β-*n*-acetylhexosaminidase complexed with intermediate analouge NAG-thiazoline	Streptomyces plicatus
	4563	(F) β-*N*-acetylhexosaminidase activity	1m03A	Mutant β-hexosaminidase (d313a)	
			1m04A	Mutant β-hexosaminidase (d313a)	
			1m01A	Wildtype β-hexosaminidase	
	5216	(F) ion channel activity	1jakA	β-*n*-acetylhexosaminidase in complex with (2r,3r,4s,5r)-2-acetamido-3,4-dihydroxy-5-hydroxymethyl-piperidinium chloride (ifg)	Streptomyces plicatus
	8200	(F) ion channel inhibitor activity			
	15269	(F) calcium-activated potassium channel activity			
	15459	(F) potassium channel regulator activity			
	6939	(P) smooth muscle contraction			
	6813	(P) potassium ion transport			
	5624	(C) membrane fraction			
	5887	(C) integral to plasma membrane			
	16020	(C) membrane			
3	3823	(F) antigen binding	1kcuL	Pc287 immunoglobulin	Mus musculus
	5576	(C) extracellular	1kc5L	Pc287 immunoglobulin	
	9405	(P) pathogenesis			
	15070	(F) toxin activity			
4	6955	(P) immune response	1ieaB, 1ieaD	Mhc class ii i-ek	Mus musculus
	19884	(P) antigen presentation, exogenous antigen	1iebB, 1iebD	Mhc class ii i-ek	
	19886	(P) antigen processing, exogenous antigen via MHC class II	1ktdB, 1ktdD	Fusion protein consisting of cytochrome C peptide, glycine rich linker, and MHC e-β-k.	
	45012	(F) MHC class II receptor activity			
	16020	(C) membrane			
	16021	(C) integral to membrane			
	5344	(F) oxygen transporter activity	1fngB, 1fngD	MHC class ii i-ek, β chain.	Mus musculus
	5833	(C) hemoglobin complex			
	6810	(P) transport	1i3rB, 1i3rD, 1i3rF, 1i3rH	Fusion protein consisting of MHC e-β-k precursor, glycine rich linker, and hemoglobin β-2 chain.	
	15671	(P) oxygen transport			
5	5507	(F) copper ion binding	1ag2_	Major prion protein, domain 121–231. Synonym: prp(121–231)	Mus musculus
	5783	(C) endoplasmic reticulum			
	5794	(C) Golgi apparatus			
	5886	(C) plasma membrane			
	6979	(P) response to oxidative stress			
	8152	(P) metabolism	1qm3A	Prion protein. Synonym: prp, major prion protein, prp27–30, prp33–35c, (ascr). Prp, residues 121–230.	Homo sapiens
	9405	(P) pathogenesis	1e1gA, 1e1jA	Prion protein variant m166v, domain 125–228.	
				Prion protein variant r220k, domain 125–228.	
			1e1uA, 1e1wA	Prion protein variant r220k, domain 125–228.	

[a](P) biological process, (F) molecular function, (C) cellular component.

process. Likewise, the annotation for protein chain 1jakA can be extended with 2 new GO terms: 5975, carbohydrate metabolism for process; and 4563, β-*N*-acetylhexosaminidase activity for molecular function. Finally, annotation of protein chains 1hp4A, 1hp5A, 1m03A, 1m04A, and 1m01A can be extended by 9 new GO terms (see cluster 2, Table II).

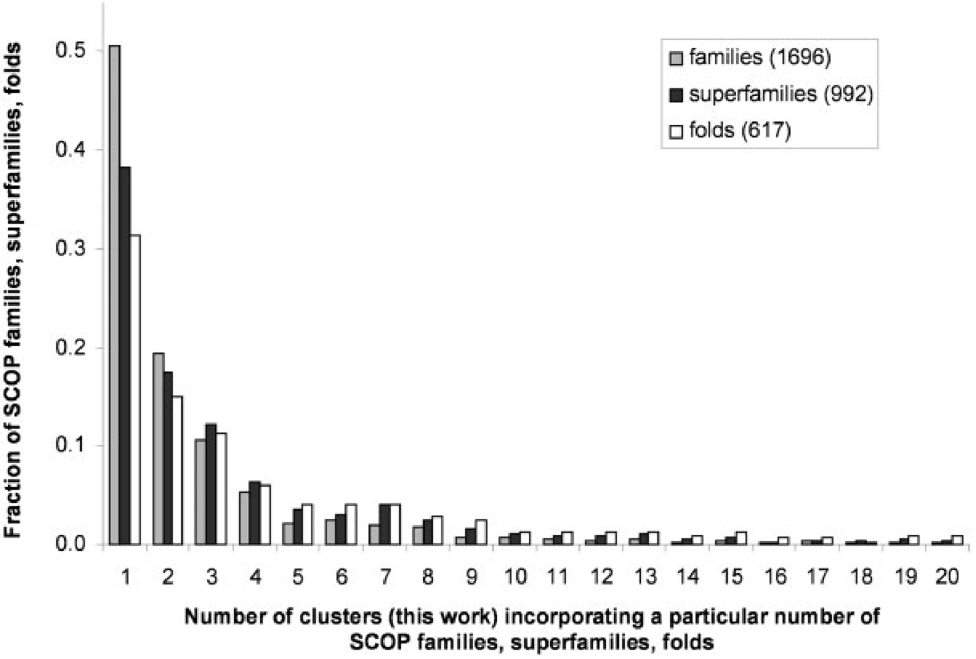


Fig. 4. Clustering of SCOP families, superfamilies, and folds with *Rmsd* ≤ 5.0 Å, *Z-score* ≥ 3.8, *Rnar* ≥ 70%, and *Rseq* ≥ 90%.

Cluster 4 in Table II represents a more curious case. While it can be assumed that proteins 1i3rB, 1i3rD, 1i3rF, and 1i3rH have been assigned GO terms—5344, oxygen transporter activity; 5833, hemoglobin complex; 6810, transport; and 15671, oxygen transport, due to the presence of the fusion proteins with a hemoglobin β-2 chain—proteins 1fngB and 1fngD have been assigned these terms by mistake: major histocompatibility complex (MHC) class II protein I-Ek is not involved in oxygen transport or in complex with hemoglobin. Meanwhile, another 6 GO terms (6955, 19884, 19886, 45012, 16020, 16021) were missed for chains 1fngB and 1fngD. This erroneous annotation provided in the version of GOA EBI used here was excluded from the next version dated December 17, 2003 (see also http://spdc.sdsc.edu).

Comparison of Our Structure Clustering With the SCOP Classification

To better understand the results from the clustering performed here, we compared these results to the well-established SCOP[23,24] resource. SCOP is a high-quality manual clustering of structures according to unique domain fold (structural similarity only), superfamily (probable common evolutionary origin), and family (structure similarity with clear evolutionary relationship). Two questions were addressed. First, what level of the SCOP hierarchy best matches the clusters detected here? Second, to what extent does the clustering presented here map to the SCOP classification? In addressing these questions, structures with low resolution, as well as peptides and designed proteins, were not considered.

Figure 4 illustrates the distribution of SCOP families, superfamilies, and folds within clusters. A total of 51% (857 of 1696) of the SCOP families, 38% (380 of 992) of the superfamilies, and 32% (194 of 617) of the domain folds occur within a single cluster; thus, the best match between clusters and the SCOP classification is achieved at the SCOP family level.

To answer the second question, the procedure used for extending GO term assignments (see Materials and Methods section, and section on selection of the model with respect to similarity criteria) was applied using the SCOP family IDs rather than GO terms. Of the PDB's 34,698 chains analyzed, 32,715 had SCOP family IDs (release 1.63). To achieve the same coverage as used for the GO terms, where 14% of chains did not have GO terms (4964 of 34,698), we randomly omitted 2981 chains as if they did not have SCOP IDs, whereupon, with equal numbers of chains, the procedure as described in the Methods section was applied, but using SCOP family IDs. Note that just as a chain can have multiple GO terms, a chain can have multiple SCOP family IDs. So, for example, chain A [Fig. 1(C)] contains 3 domains, with different SCOP family IDs denoted a, b, and c, and chain C [Fig.1(C)] contains 2 domains with different SCOP family IDs denoted c and d.

Table III illustrates specificities [Eqs. (2–4)] and coverage using SCOP family IDs. Following the definitions of specificity (see Material and Methods section), a cluster containing chains with different numbers of SCOP family IDs and at least 1 common SCOP family ID would be considered a "bad cluster" using the definition of specificity-1 and specificity-2, but a "good cluster" using the definition of specificity-3. Thus, in considering just the SCOP family classification, it is reasonable to only consider specificity-3. Table III illustrates that for the same optimum threshold values as used for chains, we observe

TABLE III. Performance at Different Threshold Values for Structure and Sequence Similarity Parameters Using SCOP Family IDs.

Threshold values				Performance on the training set[a]								Performance on the test set			
Rseq, %	Rnar, %	Rmsd, Å	Z-score	Av. specificity-1, %	Std. specificity-1, %	Av. specificity-2, %	Std. specificity-2, %	Av. specificity-3, %	Std. specificity-3, %	Av. coverage, %	Std. coverage, %	Specificity-1, %	Specificity-2, %	Specificity-3, %	Coverage, %
0	90	2.0	4.5	99.86	(0.02)	99.97	(0.01)	99.97	(0.01)	55.75	(0.31)	99.84	100	100	61.21
25	90	2.0	4.5	99.86	(0.02)	99.97	(0.01)	99.97	(0.01)	55.53	(0.30)	99.84	100	100	60.90
35	90	2.0	4.5	99.93	(0.01)	100	(0)	100	(0)	55.07	(0.32)	99.85	99.96	99.96	60.43
50	90	2.0	4.5	99.92	(0.01)	100	(0)	100	(0)	52.07	(0.37)	99.88	100	100	56.87
70	90	2.0	4.5	99.92	(0.01)	100	(0)	100	(0)	50.01	(0.41)	99.89	100	100	55.55
90	*90*	*2.0*	*4.5*	*99.95*	*(0.01)*	*100*	*(0)*	*100*	*(0)*	*47.63*	*(0.35)*	*99.90*	*100*	*100*	*54.07*
0	70	5.0	3.8	69.34	(0.25)	71.70	(0.22)	72.41	(0.22)	78.12	(0.27)	67.31	70.11	70.62	82.89
25	70	5.0	3.8	96.51	(0.06)	98.42	(0.06)	98.50	(0.06)	75.05	(0.30)	94.29	96.54	96.54	80.22
35	70	5.0	3.8	98.29	(0.06)	99.72	(0.03)	99.80	(0.01)	72.25	(0.21)	98.30	99.81	99.81	77.93
50	70	5.0	3.8	98.59	(0.06)	99.92	(0.02)	100	(0)	69.53	(0.33)	98.68	100	100	76.22
70	70	5.0	3.8	98.62	(0.07)	99.91	(0.02)	100	(0)	65.40	(0.36)	98.79	100	100	72.46
90	**70**	**5.0**	**3.8**	**98.84**	**(0.04)**	**99.99**	**(0.01)**	**100**	**(0)**	**61.72**	**(0.39)**	**98.88**	**100**	**100**	**69.86**

[a] Average (Av) and standard deviation (Std) values were computed over 10 random samples while dividing the whole set into 2 equal subsets for training and testing.

69.86% annotation coverage at 100% specificity (row in bold, Table III). Thus, the clustering presented here maps well onto the SCOP classification at the family level. Stated another way, almost 70% of the SCOP annotations at the family level assigned manually by the SCOP authors could be automatically derived by the method presented here.

Annotation of Proteins Structurally Classified as SCOP Superfolds

As noted, similar structure alone does not necessarily define similar function and hence shared GO terms. Our premise is that when structure is used in concert with a suitable *Rseq,* false annotation that would result from structure comparison alone does not occur. In other words, the incorporation of the sequence relationship defined by structure alignment prevents the occurrence of incorrect annotations; yet, alone, the sequence relationship is insufficient to unambiguously identify the correct annotation. This premise is tested by the analysis of a number of SCOP folds referred to as "superfolds"—structurally similar domain folds with multiple functions. Thus, we address the question, given the selection criteria used: Is our ability to functionally annotate proteins by this method adversely impacted by the presence of folds with multiple functions? Figure 5(A and B) plots the difference for specificity-3 between SCOP folds and SCOP superfolds (we consider as superfolds all folds that are present in more than one superfamily) using different comparison parameters. It is clear that superfolds contribute substantially to the number of FP annotations when the sequence identity in the structure alignment (parameter *Rseq*) is less than 25%, regardless of other parameters, thus decreasing overall specificity. However, the effect of superfolds is not noticeable when the sequence identity is above 90%. For the final annotation provided by this method, an *Rseq* threshold of greater that 90% was used and hence inhibits any adverse influence from superfolds.

Which superfolds have the most adverse effects upon annotation? Figure 5(C and D) illustrates results for the 10 most represented superfolds in SCOP for different comparison parameter sets. For *Rseq* values below 90%, the major contribution of FPs comes from the Ig-like β-sandwich [b.1, red circle, Fig. 5(C and D)]. In the final analysis, only 43% of the Ig-like fold proteins from SCOP are annotated by this method, and 80% of proteins from the other 9 superfolds.

Biological Relevance and the problems of Annotation of 3D Protein Chains in the PDB

Consider examples of new GO annotation assigned to protein chains in the PDB (Fig. 2 and Table II). Some of the associations are obvious; some need to be validated through analysis of the literature or by experiment. For example, mouse major prion protein (domain 121–231) was manually annotated by GOA EBI using 5 GO terms: (1) 5507, copper ion binding for function; (2) 6979, response to oxidative stress for process; (3) 5783, endoplasmic reticulum; (4) 5794, Golgi apparatus; and (5) 5886, plasma membrane for cellular component (cluster 5, Table II). According to our approach, 5 PDB chains, 1qm3A, 1e1gA, 1e1jA, 1e1uA, and 1e1wA, representing the same domain of human major prion protein should also be assigned these GO terms. Assignment requires consideration of posttranslationally modified and orthologous proteins. Does a protein found in humans possess the same function, and is it found in the same cellular location as the one from mouse? This question requires experimental work to be answered unambiguously.

In another, similar case, the mutated variants of human prion protein m166v and r220k (Table II, cluster 5) had similar conformations to the orthologous proteins from other species.[25] Further investigation of their functionality is required. It is known that point mutations could lead to loss of protein functionality either in part or in full. For example, we assigned 3 mutant proteins (1gs6X, 1gs7A, and 1gs8A) of nitrite reductase from *Alcaligenes xylosoxidans* 2 GO terms: (1) 5507, copper ion binding for function; and (2) 6807, nitrogen metabolism for process, according to their structural similarity to the same wild-type proteins (1bq5, 1hauA). It was experimentally shown that D92N (1gs8A) and H254F (1gs7A) mutants had negligible or no activity, while the M144A mutant (1gs6X) had 30% of the native enzyme activity.[26] Thus, while the native protein had the functional activity, the post-translationally modified proteins did not, yet they were assigned the same GO terms.

Assigning GO terms to a particular PDB chain is problematic when the chain only represents part of natural protein. For example, the PDB entry 1cxwA represents one domain of the human matrix metalloproteinase-2, which is a multidomain protein, whereas 2fn2 represents the structure of the human fibronectin type II domain. The CE alignment of 1cxwA against 2fn2 results in an *Rmsd* of 1.3 Å and a sequence identity of 51% for structurally aligned regions. However, the proteins containing these domains have completely different biological functions. Matrix metalloproteinase-2 is an enzyme (EC 3.4.24.24) involved in development, inflammation, wound healing, tumor invasion, metastasis, and other processes. The enzyme cleaves several types of collagen, elastin, fibronectin, and laminin proteins bound by 3 fibronectin type II domains. Conversely, fibronectin is not an enzyme; it binds to cell surfaces and various compounds, including collagen, fibrin, heparin, DNA, and actin. Fibronectins are involved in cell adhesion, cell motility, opsonization, wound healing, and maintenance of cell shape (Swiss-Prot: COG2_HUMAN, FINC_HUMAN). Thus, the extension of GO terms assigned to one protein is not always applicable to the other. GO terms for these PDB entries were assigned by GOA EBI for the whole protein. However, it has not been shown experimentally that the particular domain has the same function when separated from the whole protein. It is suggested that the GO include such terms as "part of protein" or "protein domain" to describe cases where the protein domain or a part of protein does not possess the function of the complete protein.

CONCLUSIONS

A fully automated method for the extension of protein annotation using GO terms is described. The method introduces the notion of using 3D structure homology to extend available annotation provided by direct evidence or through inferred sequence homology. The method is tuned by examining the impact on coverage versus specificity when using different parameters that describe sequence and structure relationships. Exploiting structure homology in addition to sequence homology allows the derivation of reliable annotation for previously unannotated proteins, thereby achieving 40% coverage at specificity levels 90.5–99.9% depending on the definition of true positives. This level of performance is not achieved from sequence homology alone.[12] The method also allows new GO terms to be added to previously annotated proteins and reveals chains with contradictory GO terms.

When applied to "superfolds" (i.e., structurally similar proteins with different functions), the method showed that superfolds contribute substantially to the number of FP annotations at low sequence identity, thus decreasing specificity. However, at high levels of sequence identity used for the annotation introduced here, specificity is high. In attempting to reproduce SCOP annotation by the method introduced here, the best performance was achieved at the SCOP family level classification, and approximately 70% of the SCOP associations could be found automatically.

The complete list of PDB protein chains annotated by this method is available at the website http://spdc.sdsc.edu/ and can also be download as a formatted text file. The site provides periodic updates of the resource based on the current PDB, CE database,[20] and GOA EBI annotation. Possible uses of the resource include the following:

- Improving target selection for structural genomics by selecting sequences with GO terms not represented in the PDB, and hence increasing the likelihood that the protein structure determined from the target sequence will have a new function, and less likely, a new fold.
- Enabling the selection of structures for further experimental study, since they have no GO annotation and hence are functionally unclassified.
- Assisting in the error detection within existing GO annotation resources, thus provoking further review of the experimental or theoretical evidence for a particular GO term being assigned to a specific protein.

In this work, GOA EBI was used as a source of GOA for PDB chains. However, any source of annotation for proteins could be used.[27] as well as any ontology describing protein function.

GO enables researchers to systematize the study of protein function, and this is a welcome step forward; however, limitations remain. For example, multifunction proteins are not well described, nor is a function that is meaningful only given some specific conditions—regulation, participation in a protein complex, requirement of a cofactor, and so on.[28] Of particular issue going forward is the problem of semantic similarity: What is the relative functional similarity of proteins described by different GO terms? While this has been investigated through relationships between semantics and protein sequence similarity,[29] the relationship between semantic similarities of GO terms versus protein sequence *and* structure similarity needs to be addressed as a follow up to the work presented here.

REFERENCES

1. The Gene Ontology Consortium. Creating the gene ontology resource: design and implementation. Genome Res 2001;11:1425–1433.
2. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, Richter J, Rubin GM, Blake JA, Bult C, Dolan M, Drabkin H, Eppig JT, Hill DP, Ni L, Ringwald M, Balakrishnan R, Cherry JM, Christie KR, Costanzo MC, Dwight SS, Engel S, Fisk DG, Hirschman JE, Hong EL, Nash RS, Sethuraman A, Theesfeld CL, Botstein D, Dolinski K, Feierbach B, Berardini T, Mundodi S, Rhee SY, Apweiler R, Barrell D, Camon E, Dimmer E, Lee V, Chisholm R, Gaudet P, Kibbe W, Kishore R, Schwarz EM, Sternberg P, Gwinn M, Hannick L, Wortman J, Berriman M, Wood V, de la Cruz N, Tonellato P, Jaiswal P, Seigfried T, White R. The Gene Ontology (GO) database and informatics resource [Database issue]. Nucleic Acids Res 2004;32:D258–D261.
3. Jensen LJ, Gupta R, Staerfeldt HH, Brunak S. Prediction of human protein function according to Gene Ontology categories. Bioinformatics 2003;19:635–642.
4. Lagreid A, Hvidsten TR, Midelfart H, Komorowski J, Sandvik AK. Predicting gene ontology biological process from temporal gene expression patterns. Genome Res 2003;13:965–979.
5. Letovsky S, Kasif S. Predicting protein function from protein/protein interaction data: a probabilistic approach. Bioinformatics 2003;19(Suppl 1):197–204.
6. Dwight SS, Harris MA, Dolinski K, Ball CA, Binkley G, Christie KR, Fisk DG, Issel-Tarver L, Schroeder M, Sherlock G, Sethuraman A, Weng S, Botstein D, Cherry JM. Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). Nucleic Acids Res 2002;30:69–72.
7. Rhee SY, Beavis W, Berardini TZ, Chen G, Dixon D, Doyle A, Garcia-Hernandez M, Huala E, Lander G, Montoya M, Miller N, Mueller LA, Mundodi S, Reiser L, Tacklind J, Weems DC, Wu Y, Xu I, Yoo D, Yoon J, Zhang P. The *Arabidopsis* Information Resource (TAIR): a model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community. Nucleic Acids Res 2003;31:224–228.
8. Blake JA, Richardson JE, Bult CJ, Kadin JA, Eppig JT. MGD: the Mouse Genome Database. Nucleic Acids Res 2003;31:193–195.
9. Camon E, Barrell D, Lee V, Dimmer E, Apweiler R. The Gene Ontology Annotation (GOA) Database—an integrated resource of GO annotations to the UniProt Knowledgebase. In Silico Biol 2004;4:5–6.
10. Groth D, Lehrach H, Hennig S. GOblet: a platform for Gene Ontology annotation of anonymous sequence data [Web server issue]. Nucleic Acids Res 2004;32:W313–W317.
11. Zehetner G. OntoBlast function: From sequence similarities directly to potential functional annotations by ontology terms. Nucleic Acids Res 2003;31:3799–3803.
12. Xie H, Wasserman A, Levine Z, Novik A, Grebinskiy V, Shoshan A, Mintz L. Large-scale protein annotation through gene ontology. Genome Res 2002;12:785–794.
13. Gerstein M, Hegyi H. Comparing genomes in terms of protein structure: surveys of a finite parts list. FEMS Microbiol Rev 1998;22:277–304.
14. Nagano N, Orengo CA, Thornton JM. One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. J Mol Biol 2002;321:741–765.
15. Halaby DM, Poupon A, Mornon J. The immunoglobulin fold family: sequence analysis and 3D structure comparisons. Protein Eng 1999;12:563–571.
16. Dunwell JM, Culham A, Carter CE, Sosa-Aguirre CR, Goodenough PW. Evolution of functional diversity in the cupin superfamily. Trends Biochem Sci 2001;26:740–746.
17. Qian J, Luscombe NM, Gerstein M. Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model. J Mol Biol 2001;313:673–681.
18. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
19. Shindyalov IN, Bourne PE. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng 1998;11:739–747.
20. Shindyalov IN, Bourne PE. A database and tools for 3-D protein structure comparison and alignment using the Combinatorial Extension (CE) algorithm. Nucleic Acids Res 2001;29:228–229.
21. Mitchell TM. Machine learning. New York: McGraw-Hill Science/Engineering/Math; 1997.
22. Efron B, Tibshirani RJ. An introduction to the bootstrap. Boca Raton, FL: Chapman & Hall/CRC; 1993. 436 p.
23. Lo Conte L, Brenner SE, Hubbard TJ, Chothia C, Murzin AG. SCOP database in 2002: refinements accommodate structural genomics. Nucleic Acids Res 2002;30:264–267.
24. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
25. Calzolai L, Lysek DA, Guntert P, von Schroetter C, Riek R, Zahn R, Wüthrich K. NMR structures of three single-residue variants of the human prion protein. Proc Natl Acad Sci USA 2000;97:8340–8345.
26. Ellis MJ, Prudencio M, Dodd FE, Strange RW, Sawers G, Eady RR, Hasnain SS. Biochemical and crystallographic studies of the Met144Ala, Asp92Asn and His254Phe mutants of the nitrite reductase from *Alcaligenes xylosoxidans* provide insight into the enzyme mechanism. J Mol Biol 2002;316:51–64.
27. Krebs WG, Bourne PE. Statistically rigorous automated protein annotation. Bioinformatics 2004;20:1066–1073.
28. Lan N, Montelione GT, Gerstein M. Ontologies for proteomics: towards a systematic definition of structure and function that scales to the genome level. Curr Opin Chem Biol 2003;7:44–54.
29. Lord PW, Stevens RD, Brass A, Goble CA. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics 2003;19:1275–1283.