Predicting Gene Ontology Functions from ProDom and CDD Protein Domains

Jonathan Schug,[1] Sharon Diskin, Joan Mazzarelli, Brian P. Brunk, and Christian J. Stoeckert, Jr.

Center for Bioinformatics, Computational Biology and Informatics Laboratory, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA

A heuristic algorithm for associating Gene Ontology (GO) defined molecular functions to protein domains as listed in the ProDom and CDD databases is described. The algorithm generates rules for function-domain associations based on the intersection of functions assigned to gene products by the GO consortium that contain ProDom and/or CDD domains at varying levels of sequence similarity. The hierarchical nature of GO molecular functions is incorporated into rule generation. Manual review of a subset of the rules generated indicates an accuracy rate of 87% for ProDom rules and 84% for CDD rules. The utility of these associations is that novel sequences can be assigned a putative function if sufficient similarity exists to a ProDom or CDD domain for which one or more GO functions has been associated. Although functional assignments are increasingly being made for gene products from model organisms, it is likely that the needs of investigators will continue to outpace the efforts of curators, particularly for nonmodel organisms. A comparison with other methods in terms of coverage and agreement was performed, indicating the utility of the approach. The domain-function associations and function assignments are available from our website http://www.cbil.upenn.edu/GO.

An important early step in the postgenomic era is the characterization of the biochemical functions of gene products. Accurate computational predictions are a useful resource for both the community at large and the curators that eventually assign function to gene products. The Gene Ontology (GO) (Gene Ontology Consortium 2001) is an ontology, that is, a database of agreed-to terms for molecular functions, biological processes and cellular components. GO also includes relationships between terms such as specialization or part-whole relations. GO was developed to facilitate effective use of this information. We present an automatic method for leveraging curated GO function annotation of proteins to associate GO terms with protein domains that can then be applied to proteins that contain any of the domains.

We make the assumption that the functions that a protein is capable of performing are determined by the protein domains that it contains. We use the simplest possible model of this kind; each domain contributes a function independently of any other domain in the protein. The basis of the algorithm, illustrated in Figure 1, is to identify, using an intersection procedure, the GO functions common to a set of proteins that each contain a domain. We determine whether or not a protein has a domain using sequence similarity. As part of the process of associating GO functions with a domain, we determine a p-value threshold for each domain that indicates the level of similarity needed to confer function. As demonstrated by Hegyi and Gerstein (2001), this is an important consideration. We contrast this approach with the comparison of entire proteins where similarities may be found between regions that are not responsible for the transferred annotation (Myers et al. 2000). An evaluation of all sequences assigned to a function is needed to ascertain the essential sequence features of the function (Henikoff and Henikoff 1994; Bateman et al. 1999; Corpet et al. 1999; Hofmann et al. 1999).

The GO function ontology consists of function classifications arranged in a directed acyclic graph, that is, a parent-child hierarchy where components may belong to more than one component but no component can be its own descendant. The top level has general terms such as *enzyme* and *nucleic acid binding*; more specific terms are located deeper in the hierarchy. The length of a path to a term is termed its *depth*. The terms in a path to a term are called the *ancestors* of the term. A term is a *leaf* if it is at the end of a path. Our method takes advantage of the structure of the GO ontology to make only as specific a prediction as is consistent with the data.

We chose the ProDom (Corpet et al. 1999) and CDD (Bryant and NCBI Structure Group 2001) databases as our initial sources of domain models. ProDom contains protein signatures determined from sequences organized into classes of functional domains by sequence clustering. The Conserved Domain Database (CDD) is comprised of a mirror of two major domain databases, the Simple Modular Architecture Research Tool (SMART) (Schultz et al. 2000) and Pfam (Bateman et al. 2000); it also includes entries from the Library Of Ancient Domains (LOAD), maintained by I. Aravind, E. Koonin, and colleagues at the National Center for Biotechnology Information.

RESULTS

We applied our algorithm (Fig. 2) using version 2.61 of the GO function ontology and 15,241 GO function associations for 11,679 gene products in *M. musculus*, *D. melanogaster*, and

[1]Corresponding author.
E-MAIL jschug@pcbi.upenn.edu; FAX (215) 573-3111.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.222902.

12:648–655 ©2002 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/01 $5.00; www.genome.org

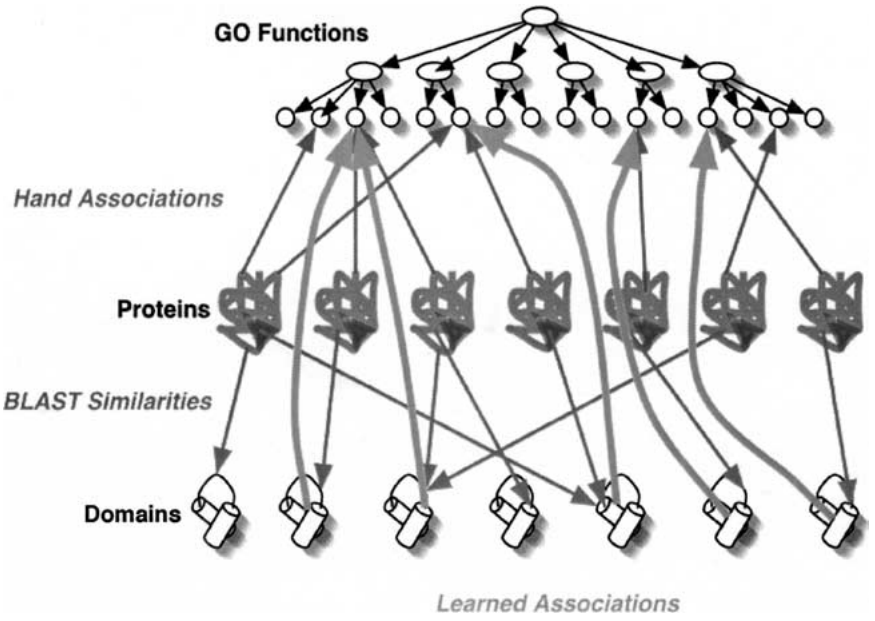

Figure 1 Illustration of the approach used to assign Gene Ontology (GO) functions to domains. Gene products (proteins) are assigned a molecular function by members of the Gene Ontology Consortium (hand associations). BLAST similarities are used to associate the GO-assigned proteins to domains. A heuristic algorithm is then used to assign GO functions to domains (learned associations).

S. cerevisiae (http://www.geneontology.org), to generate domain-function associations or rules. Manual review of approximately 400 of the rules indicates that the error rate, in terms of completely wrong associations, is about 10%–15%. The majority of incorrect assignments can be attributed to domains that have repetitive elements or are found in various classes of proteins that result in significant similarities to GO-associated proteins (GOAPs), but are not definite predictors of their function. Two examples are the WD40 motif and the LIM domain. WD40 repeats are often found in proteins having a variety of functions (Smith et al. 1999). The LIM domain is involved in mediating protein–protein interactions (Bach 2000) and is found in proteins of varied function. Some of these suspect domain-function assignments can be identified as having large numbers of HSPs (high-scoring pairs) in BLAST (Altschul et al. 1997) comparisons. We selected a cutoff of five HSPs to avoid these repeats when assembling a list of similarities to a given GO-associated protein. Similarly, when predicting GO functions, we ignore such similarities.

Rule Generation

Four groups of rules were generated. To ascertain the impact of each domain database, we initially generated rules and made predictions using ProDom and CDD separately. Subsequent analysis was performed to determine the impact of joining the two databases. Furthermore, two separate rule groups were generated for each motif database used; one group included the use of all GO associations (ProDom-All or CDD-All), and the other group excludes those with an "Inferred from Electronic Annotation" (IEA) evidence code (ProDom-NoIEA or CDD-NoIEA). Associations to GO:005554, *molecular function unknown*, were excluded from rule generation in all cases; these associations do not help us learn meaningful domain-function associations.

We categorize rules by types (Fig. 3) that describe the amount and consistency of the data used to generate the rule. Table 1 reveals the total number and distribution of the rule types for each rule group. The majority of domains in both of the ProDom rule groups have only one GOAP. Consequently, there is no opportunity to intersect functions, and we simply associate all of its functions with the domain. This is also the most frequent rule type generated in the CDD rule groups; however, it does not represent the majority of the domains. We will examine the accuracy of this type of rule below.

ProDom Rules

The two ProDom rule groups exhibit a similar rule-type distribution, as shown in Table 1, although the percentage of "one protein" rules is higher in the group that does not make use of the IEA-based associations. The ordering of rule types from most to least common does not change with the inclusion of GO associations with the IEA evidence code. In both ProDom rule groups, the most common rule type in cases where the domain has multiple GOAPs is "single function," where all GOAPs have the same associated GO functions. This is the simplest possible nontrivial intersection. It is possible, in this case, for the GOAPs to have multiple common functions, only one of which should actually be associated with the domain under consideration. This was not seen to be case in practice. Rather, this rule type primarily associates GOAP members of the same gene family to a ProDom domain describing that family. For example, in the ProDom-NoIEA group, the rule for ProDom domain PD001109 HEXOKINASE TRANSFERASE GLYCOLYSIS ATP-BINDING TYPE HK ALLOSTERIC ENZYME GLUCOKINASE was based on six GOAPs each with the function GO:004396 *hexokinase*. (We note that the ProDom domain names typically include key words taken from the proteins that contain the domain, but may be unrelated to the function of the domain.) Almost as common as "single function" rules are the "consensus ancestor" rules that make explicit use of the hierarchical structure of the GO function ontology. By considering the ancestor functions of GOAPs, a more general description of function shared by all GOAPs may be found. In practice, neither of the ProDom groups generated a rule set based on a "near consensus ancestor" function; all were based on a "consensus ancestor" function. An example of this rule type is the association of the GO function GO008236 *serine-type peptidase* with the domain PD000068 PROTEASE SERINE HYDROLASE PRECURSOR SIGNAL ZYMOGEN GLYCOPROTEIN 3.4.21.-FACTOR FAMILY. Fourteen GOAPs were used to create the rule; three were annotated as *serine-type peptidase* but the rest were annotated as more specific terms *serine-type endopeptidase*, *trypsin*, *tissue kallikrein*, or *monophenol monooxygenase activator* that were generalized to *serine-type peptidase*. The third most common rule type is that of "consensus leaf" (leaf refers to GOAPs being assigned functions at the same hierarchy

```
1. BLAST GO-annotated proteins against domains. Only keep results
   with p-values <= 10e-5.
2. For each domain:
   a. Generate a list of proteins and their p-values from the
      BLAST runs. Sort the list according to p-value. If there
      are no proteins on the list, then generate a "no protein"
      rule and go on to the next domain.
   b. Go through the list to generate a rule for the domain.
      i.   Assign a function(s) to the domain based on the best p-
           value. This is a "one protein" rule. If there are no
           more proteins, go on to the next domain.
      ii.  Consider the next protein on list with those above it.
           For these proteins, go through the rule generators in
           the order "single function", "consensus leaf", "near
           consensus leaf", and the deeper of "consensus ancestor"
           and "near consensus ancestor", until the rule
           conditions are met. Assign that rule to the domain at
           the p-value for the lowest protein on the list
           considered. Repeat this step until there are no more
           proteins on the list and go on to the next domain.
      iii. If no rule conditions are met, then assign a "no call"
           rule.
```

Figure 2 Algorithm for assigning a GO function to a domain.

level). The "consensus leaf" rules apply to multiple GOAPs with a common GO function, when at least one of the proteins has at least one other GO function. For example, the rule for ProDom PD156480: PHOSPHATASE TYROSINE HYDROLASE PROTEIN-TYROSINE STRUCTURAL TYPE CYTOSKELETON NON-RECEPTOR ISOFORM SPLICING was assigned GO:0008092 *ligand binding or carrier:protein binding:cytoskeletal protein binding protein*. One GOAP containing this domain was annotated with an additional function, *protein tyrosine phosphatase*, that is associated with a kinase domain that only that GOAP contained. This function was not corroborated by the other GOAPs, and so it was not associated with the ProDom domain PD156480.

In cases where most, but not all, GOAPs share a function, it is possible that the function in question is indeed shared, but is simply not annotated for all GOAPs. The "near consensus leaf" rule takes this into account by requiring at least 80% of at least five GOAPs to have a common GO function. This turned out to be the least common rule, mainly because the number of GOAPs per domain tends to be fewer than five. For domains similar to 10 or more GOAPs, the "near consensus leaf" was more common than "single function" and "consensus leaf" rules combined. Indeed, the rule with the most GOAPs per domain (76) was a "near consensus leaf." An example of this rule type is the rule associating GO:0003723 *RNA binding* with ProDom domain PD259036 NUCLEAR RNA-BINDING REPEAT POLY-U-BINDING-SPLICING-FACTOR RO POLYU-BINDING CONSENSUS C18A3.5 NSAP1 GRY-RBP. The rule uses 22 GOAPS and was generated by ignoring two leaf functions, *translation initiation factor* and *helicase* that were associated with one GOAP each.

The domain coverage (i.e., the percentage of ProDom domains containing more than one sequence for which we can associate one or more GO functions) is higher in the ProDom-All group. Without using the IEA associations, 11% of the 95,518 ProDom domains containing more than one sequence are assigned one or more GO functions. Incorporating the IEA associations increases the number of domains for which we can assign function by 62%, yielding a domain coverage of 18%. These results are expected and reflect one of the impacts of including IEA associations; however, a concern is the potential negative impact on rule accuracy, as discussed below.

CDD Rules

The CDD rule groups show a similar rule-type distribution (Table 1); the ordering of rule types from most to least common does not change with the inclusion of GO associations with the IEA evidence code. Again, the percentage of "one protein" rules is higher in the group that does not make use of the IEA-based associations. The two most common rule types for CDD domains with multiple GOAPs are the reverse of those for ProDom rule groups. The most common rule type in cases where the domain has multiple GOAPs is "consensus ancestor". An example of this rule type is given by the CDD entry: pfam00783: IPPC, Inositol polyphosphate phosphatase family catalytic domain that is assigned GO:0004437 *inositol/phospha tidylinositol phosphatase* based on seven GOAPs. Additional, more specific functions included *inositol-1,4,5-triphosphate phosphatase* and *phosphatidylinositol-bisphosphatase*. The second most common rule type for both CDD rule groups is "single function". An example of a "single function" rule for CDD is illustrated by the assignment of GO:0004169 *dolichyl-phosphate-mannose-protein mannosyltransferase* to CDD entry pfam02366: PMT: Dolichyl-phosphate-mannose-protein mannosyltransferase

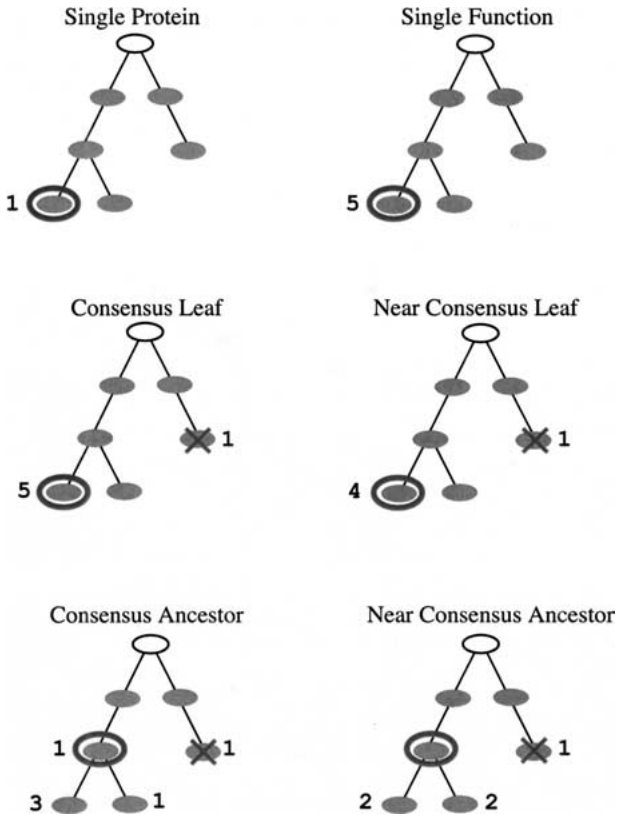

Figure 3 Examples of the six rule types. The drawings represent a portion of the GO hierarchy where each gray oval is a GO term. The top open oval is the GO term *molecular function* that serves to distinguish functions from *cellular component* or *biological process*. In every example, except "single protein", there are five GOAPs having similarity to a particular domain. The numbers next to the term ovals indicate the number of times the term is the deepest function associated with a GOAP. In "consensus leaf" and "consensus ancestor", one of the proteins has additional GO function terms that our method does not associate with the domain. In "near consensus leaf" and "near consensus ancestor", a spurious protein has similarity to the domain, but the GOAP and its GO function terms were not used in generating the rule.

Table 1. Distribution of Rule Types

Rule group[a]	Total number of rules[b]	One similar GO protein[c]	Single function[c]	Consensus leaf[c]	Near consensus leaf[c]	Consensus ancestor[c]
ProDom-All	17504	11498 (66%)	2466 (14%)	1251 (7%)	472 (3%)	1817 (10%)
ProDom-NoIEA	10724	7579 (71%)	1361 (13%)	540 (5%)	176 (2%)	1068 (10%)
CDD-All	1792	651 (36%)	337 (19%)	179 (10%)	161 (9%)	464 (26%)
CDD-NoIEA	1396	639 (46%)	249 (18%)	105 (7%)	84 (6%)	319 (23%)

The number and percentage of rules and rule-types for each combination of domain database (ProDom or CDD) and set of protein GO associations (All or No IEA) are shown. No "near consensus ancestor" rules were ever created. [a]Protein domain database and types of GO associations considered when building rules. [b]Total number of rules built. [c]Number and percent of rules as a function of rule types illustrated in Figure 3.

based on seven GOAPs. The two least common rule types were that of "consensus leaf" and "near consensus leaf".

As with the ProDom rule groups, an increase in coverage can be seen in the CDD-All group over the CDD-NoIEA group, as shown in Table 2. Without using the IEA associations, 38% of the CDD entries are assigned one or more GO functions. Incorporating the IEA associations increases the coverage to 52%, without significantly changing the distribution of rule types.

P-Value Thresholds

As part of determining the association between a domain and GO functions, our method determines a BLAST p-value threshold such that only similarities between the domain and novel proteins that have a lower p-value are used to determine function. We observe that 70%–80% of the domains are found to be similar only to proteins with completely consistent GO functions. In these cases, the threshold is simply the highest p-value between the domain and the GOAPs. Only a small fraction (1% for CDD and 3% for ProDom) of the GOAPs were used in a "near consensus" rule and were not consistent with the rule. Our method will predict an incorrect function for such proteins. On average our method rejected 40% of the GOAPs per domain as being inconsistent with a rule when there were at least five GOAPs for a domain. Most (95%) of the p-value thresholds for these rules were between 10^{-5} and 10^{-40} with an average of 10^{-14} and a mode of 10^{-6}.

Table 2. Coverage of Motif Databases as a Function of Usage of IEA Protein-Function Associations

Rule set[a]	Number of domains with a rule[b]	Coverage[c]
ProDom-NoIEA	10724	11%
CDD-NoIEA	1396	38%
LOAD	27	49%
Smart	353	59%
Pfam	1016	34%
ProDom-All	17504	18%
CDD-All	1792	52%
LOAD	31	59%
Smart	417	75%
Pfam	1344	47%

LOAD, Smart, and Pfam are component databases in CDD. We only considered those ProDom domains derived from at least two sequences. [a]Domain (sub)database and GO associations used to build rules. [b]Total number of domains with at least one rule. [c]Percent of (sub)database for which a rule was created.

This is consistent with the data of Hegyi and Gerstein (2001), and thus our method is able to identify reasonable similarity cutoffs that should prevent erroneous functional assignments.

Impact of Using Electronic Annotation During Rule Generation

Including GO associations based on electronic annotation in our training set allows us to learn more domain-function associations, but this may come at the expense of accuracy if the IEA annotation is incorrect frequently enough. We measured the effect of IEA annotation by comparing rules made with and without IEA annotation on a domain-by-domain basis. We restricted the analysis to the 3627 ProDom and CDD domains that associated a single function in both UNION-noIEA and UNION-IEA rule sets where the UNION-noIEA rule had a rule type other than "one protein". For each domain, we determined the set of functions that were common to both rule sets and compared each rule set to the set of common functions. We found that 4% of the UNION-noIEA rules had nothing in common with the UNION-IEA rules, 87% of rules agreed completely with the common set, and 9% of the rules were at least one level deeper. The corresponding statistics for the UNION-IEA rules were 91% identical and 5% more specific. To measure the amount of support that IEA GOAPs add to rules, we compared the change in the number of GOAPs used for each domain. Of the rules that had some functions in common, 4% used fewer GOAPS, 57% had the same amount, and the remaining 39% added one or more GOAPs. While there was some evidence of inconsistencies, in practice, it appears that the majority of IEA data, when used with our algorithm, yields rules that are consistent with rules produced without it.

Agreement with InterPro to GO Mapping

We performed a comparison of our learned domain-function associations with the InterPro to GO mappings available at www.geneontology.org. The results are summarized in Figure 4. Of the ProDom domains considered, 29% are mapped to GO functions in both our learned associations and the InterPro to GO mapping. Within this set of domains, 89% agree on at least one term; the average depth of these agreements is 3.26 terms. Domains for which neither approach produced an association accounted for 36% of the ProDom entries considered. The remaining 35% is divided between two sets, those for which Interpro provides a mapping and we do not (21%), and those for which we provide an association but there is no mapping available from InterPro (14%). Pfam comparisons

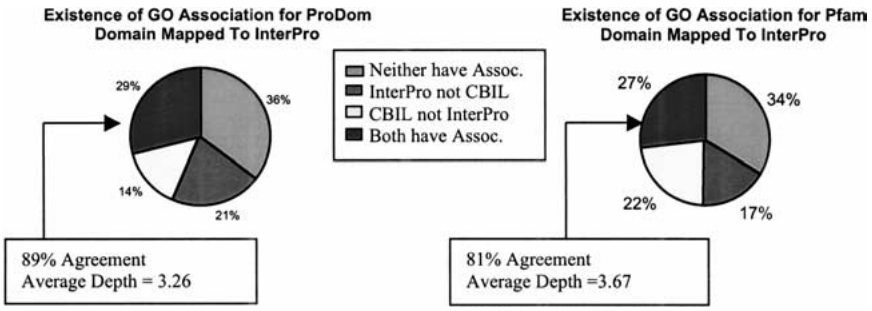

Figure 4 Comparison with InterPro to GO Mapping. The two charts show the percentages of domains in ProDom and Pfam that have been associated with GO functions by InterPro and/or by our method (Computational Biology and Informatics Laboratory [CBIL]). In cases where both methods have yielded assignments, the figure indicates the percentage of assignments that have some agreement and the average depth of agreement. We consider only ProDom domains derived from at least two sequences.

show a similar result, with 27% having GO associations in InterPro mapping as well as in our learned associations. Within this set, 81% agree on at least one term, and these agreements have an average depth of 3.67 terms. Entries for which neither approach produces an association account for 34% of Pfam. The remaining 39% is again divided between entries for which InterPro provides a mapping and we do not (17%), and entries for which we provide an association and InterPro does not (22%).

Agreement and Coverage Assessments

To assess the performance of our rules when used to annotate proteins, we made predictions for 4357 manually curated human proteins from EBI (ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/HUMAN) that we were able map to SWISS-PROT. None of these sequences were used in creating the rules. To increase coverage and to examine the effects of the p-value thresholds, we relaxed the BLAST p-value (pv) requirements to allow domain-protein similarities as long as the similarity p-value (sim pv) $\leq 10^{-30}$ or $-\log(\text{sim pv}) / -\log(\text{rule pv}) \geq 0.8$; for example, we would consider similarities as high as 10^{-16} for a rule with a threshold of 10^{-20}. The results are shown in Table 3. Using all rules and all similarities as described, we achieved a protein coverage of 81%. Of the proteins with predictions, 74% had a prediction that agreed with the curated function, and 18% had no agreeing prediction. Our algorithm automatically assigns a confidence (high, medium, or low) to each rule that depends on its type and p-value threshold. The confidences can be adjusted when a rule is curated. Considering only the rules with high and medium confidence decreases coverage but increases agreement to 81% of covered proteins. As expected, low-confidence and 'one protein' rules generate predictions with lower agreement, as does the use of domain-to-protein similarities higher than the rule threshold. If we use just the most reliable rules and similarities, we cover 51% of the proteins considered with only 11% nonagreement. Similar trends are observed when IEA associations are included. We find that including IEA annotation yields significantly greater coverage (67%) with essentially the same reliability.

Table 3. Coverage and Agreement of Predictions with EBI-Curated Annotation of 4357 Human Genes Found in SWISS-PROT Using Both ProDom and CDD Rules

IEA[a]	Rule subset[b]	P-value ratio[c]	Coverage[d]	Agreement[e]		
				yes	some	none
No	All	All	3517 (81%)	74%	9%	18%
	High+Medium Confidence	All	2631 (74%)	81%	6%	12%
		>=1	2463 (70%)	82%	6%	12%
		< 1	400 (11%)	67%	16%	17%
	Low confidence	All	2821 (80%)	67%	11%	23%
		>=1	2685 (76%)	69%	10%	20%
		< 1	601 (17%)	39%	15%	46%
	'One Protein'	All	2275 (65%)	61%	15%	25%
		>=1	2052 (58%)	64%	14%	23%
		< 1	486 (14%)	37%	22%	41%
	H+M - 'One Protein'	>=1	2201 (63%)	84%	5%	11%
Yes	All	All	3909 (90%)	76%	7%	17%
	H+M - 'One Protein'	>=1	2902 (74%)	83%	5%	12%

[a]Indicates whether IEA GO Function associations were used to build rules. [b]Indicates the subsets of the rules that were considered when making predictions. We considered difference rule confidences and rule type. Though 'one protein' rules are less desirable a priori, some have high confidence if the p-value threshold is low or they have been manually reviewed. "H+M-'One protein' " means the high and medium confidence rules that are not 'one protein' rules. [c]Indicates what subset of similarities between human proteins and domains were considered when making predictions. The p-value ratio is defined as -log(sim pv)/ -log(rule pv), so a value less than one indicates a similarity that is less stringent than the threshold associated with the rule [d]Number and percent coverage of proteins. Bold entries indicate coverage based on 4357 proteins, others are based on number of proteins covered with all rules and all p-values with or without use of IEA annotation. Note that multiple functions are often predicted for proteins and so coverage percents may sum to more than 100%. [e]Best agreement per protein between predictions and curated annotation is categorized as agreeing (exactly or with more or less specificity), showing some agreement (paths to terms overlap but differ at leaf terms), or showing no agreement.

GO function(s) were predicted for other data sets, proteins in SWISS-PROT and musDoTS (EST assemblies from *M. musculus*) (http://www.allgenes.org) as shown in Tables 4 and 5. The average depth of the predictions is about 3.5 terms. Coverage is increased to 31% by the inclusion of multiple domain databases. The inclusion of IEA associations yields an increase from 48% to 56% in the number of SWISS-PROT proteins having a GO function association when considering the combined domain sources. The corresponding increase for *M. musculus* genes was 11%. For other species, the coverage of UNION-NoIEA rules was *A. thaliana*, 37% of 25009; *C. elegans*, 40% of 19774; *D. melanogaster*, 48% of 13228; and *S. cerevisiae*, 47% of 6358.

DISCUSSION

Association of molecular functions to protein domains provides a means to predict functions for novel sequences and may also lead to insights regarding the physical basis of the functions. We have used the physical basis for molecular function, as encoded in amino acid sequence, to make these associations. For these associations to be meaningful, multiple instances of proteins must be involved; otherwise the risk of coincidental assignments is high. Only ProDom domains represented by two or more proteins were included in our analysis, to reduce the risk that GO-assigned gene products would match ProDom domains based on similarity to regions not involved in eliciting the GO function. This is an improvement over direct BLAST searches of unassigned proteins against those with assigned GO functions. The limited amount of curated data currently available meant that we could not always realize the full benefits of our approach. We found that the application of p-value thresholds was an effective means of limiting errors and that we can identify rules that are more likely to produce incorrect assignments. This information can be considered during curation or, perhaps more importantly, before curation by any user of the predictions.

Our approach performed well in terms of assigning functions that were appropriate based on manual review of a large number of the rules. As mentioned above, domains containing repeats were the major cause of incorrect assignments (false positives). Another source of error arises from incomplete sets of GOAPs matched to domains as a result of a p-value threshold. These issues can be addressed by including steps to remove or reduce the influence of repeats and to include an evaluation for allowing GOAPs that just miss the p-value threshold. Another concern is errors of nonassignments (false negatives). We covered 40% of *C. elegans* and 37% of *A. thaliana*, two species that we did not use in our training set. This amount of coverage is certainly useful, but we recognize the need for broader coverage. Coverage is limited by both available domains and annotation of protein function. Combining multiple domain databases offers some means for improvement in this area. Interpro (Apweiler et al. 2001) has integrated several such databases, providing the opportunity to generate rules that may take advantage of the same domain described in multiple fashions for assigning GOAPs. Occasionally, for those domains that are associated with functions, we found that the functions could be more precise (deeper in the GO hierarchy) or that additional functions could be added. These were the result of missing annotation for the GOAP. The "near consensus" rules are able to accommodate these cases to a certain degree and may be optimized further. We found that the use of IEA annotation increased coverage with only a small increase in the rate of bad predictions and a slight decrease in the level of detail of the prediction. In the present study, we analyzed our method using the more conservative UNION-noIEA rules so that we can make more reliable predictions, but for the purposes of a first pass of automatic annotation prior to manual curation, we would recommend using the IEA annotation since it increases the coverage.

Other methods have been described for predicting gene function. A decision tree method was used to generate rules assigning functions to SWISS-PROT features such as keywords and species as well as based on homology (King et al. 2000). At the time we began the present study, the GO functional hierarchy was not populated sufficiently for such an approach to work because of concerns of over-fitting due to sparse data. Several computational strategies were recently used to assign GO terms to a collection of *M. musculus* cDNAs (Kawai et al. 2001). In one strategy (Fleischmann et al. 1999) that also used a notion of the consensus of annotation, the *M. musculus* sequences, with similarity to protein sequences from SWISS-PROT/TrEMBL (SPTR), were assigned GO terms using a translation table mapping the SPTR keywords to GO terms. In another strategy, (Ashbumer 2000. http://www.geneontology.org/egad2go) the sequences were assigned EGAD cellular roles and a translation table was used to map the EGAD cellular role(s) to GO terms. Our approach differs since we have attempted to relate function to a particular subsequence of protein (as defined by ProDom or CDD) rather than to a pro-

Table 4. GO Function Prediction Coverage Summary for musDoTS (*M. musculus* EST assemblies) with Similarity to NRDB as a Function of the Domain Database and Use of IEA Annotation

Rule group[a]	Number of assemblies with prediction(s)[b]	Number of predictions[c]	Average number of predictions per assembly[d]	Average depth of predictions[e]
ProDom-NoIEA	9633 (28%)	44219	4.6	3.3
ProDom-All	11721 (35%)	53386	4.6	3.3
CDD-NoIEA	10074 (30%)	44423	4.4	3.1
CDD-All	10771 (32%)	47996	4.5	3.2
UNION-NoIEA	12711 (37%)	64818	5.1	3.3
UNION-All	14013 (41%)	73276	5.2	3.4

[a]Domain database and type of GO associations used to build rules. [b]Number and percent of EST assemblies for which a prediction was made. [c]Total number of GO function terms predicted. [d]Average number of GO function terms predicted per assembly. [e]Average number of depth of predicted GO functions terms.

Table 5. GO Function Prediction Coverage Summary for SWISS-PROT as a Function of the Domain Database and Use of IEA Annotation

Rule group[a]	Number of entries with prediction(s)[b]	Number of predictions[c]	Average number of predictions per entry[d]	Average depth of predictions[e]
ProDom-NoIEA	38520 (39%)	197068	5.1	3.6
ProDom-All	48960 (50%)	264566	5.4	3.8
CDD-NoIEA	40530 (41%)	178252	4.4	3.3
CDD-All	47076 (48%)	216976	4.6	3.5
UNION-NoIEA	47105 (48%)	248943	5.3	3.6
UNION-All	55287 (56%)	310180	5.6	3.8

[a]Domain database and type of GO associations used to build rules. [b]Number and percent of entries for which a prediction was made. [c]Total number of GO function terms predicted. [d]Average number of GO function terms predicted per entry. [e]Average number of depth of predicted GO functions terms.

tein with particular descriptive features. A benefit of our method is that it may provide a basis for further annotation of GOAPs assigned to a common domain through investigation of additional and lower-level functions associated with some members. In general, our approach is complementary to the others in providing predictions where others do not and providing the correct prediction in some cases where approaches disagree.

Our future plans include the incorporation of more domain databases, consideration of combinations of domains, and manual curation of predictions made on musDoTS EST assemblies. This will serve to carefully check the predictions we are making, and subsequently to provide more data to use in building new rules.

METHODS

In a noise-free situation, the learning process would be simple because of our assumptions about the independent and consistent actions of domains in determining GO functions. Every protein p has a set of associated GO functions F_p. Given a domain d there exists a set P_d of proteins that each contain the domain. Any protein that contains a domain should have the domain's function(s) associated with it, and therefore that function will be a member of the intersection of the sets of functions for each protein in P_d. We therefore define the set of functions associated with the domain d as $F_d = \cap F_p$ for all $p \in P_d$. However, since errors in annotation and the identification of protein are expected, we define a nonideal intersection described below.

We implemented our algorithm as follows. We use `WU-BlastP` (http://blast.wustl.edu) to compare GOAPs with ProDom domains and `RPS-Blast` (Bryant and NCBI Structure Group 2001) to compare GOAPs with CDD domains. These results are stored in a data warehouse, GUS (Genomic Unified Schema) (Davidson et al. 2001). Then, for each domain, we perform the nonideal intersection process, illustrated in Figures 2 and 3, on sets of functions associated with proteins containing the domain. We do not consider ProDom domains that contain only one known protein, to avoid searching with multidomain proteins rather than individual domains. Recognizing that a `BLAST` similarity is not a perfect reflection of domain membership, we record similarities with p-values (or e-values in the case of `RPS-BLAST`) as high as 10^{-5} knowing that a fraction of them are incorrect. We assume that incorrect assignments tend to occur when the p-value is too high, and vary a p-value threshold to select the largest set of proteins that yields a (nearly) complete intersection of GO functions. A simple method of doing this is to take the highest p-value among the similarities that yields a non-null intersection. However, often there is not a p-value threshold that cleanly separates the GO functions, and rigid adherence to this method could degenerate to taking the lowest p-value only. We take two approaches to this situation that can be used separately or together. First, using the hierarchical nature of GO, we can perform the intersection process at lower resolution if need be. The protein associations are made to nodes deep in the GO hierarchy that are located on one or more paths back to the root. We form domain-function associations to nodes that are in the intersection of sets of these paths rather than on sets of the leaves at the ends of the paths (leaves). If the paths intersect deeply enough in the hierarchy, then the computed F_d is a useful prediction. By using the hierarchy of the categories, we can often regain coverage by making a less detailed prediction. This is analogous to expanding a confidence interval when making predictions in real-valued variables. Rules of this type are called "consensus ancestor". The second approach is to relax the constraint that a predicted function must be associated with every member of P_d. We adopt the heuristic that a node is a member of the intersection if it is associated with at least 80% of the proteins in P_d and P_d contains at least five proteins. We accept the highest p-value that generates an acceptable approximate intersection and is consistent with the associations of the protein with that p-value. Rules of this type are called 'near consensus'. When the two heuristics are combined, the rule type is "near consensus ancestor". We look for a near ancestor that is either a consensus or a near consensus and take the deeper of the two (possibly) distinct nodes. It is possible for a directed acyclic graph (DAC) to have a consensus node that is deeper than a nonconsensus node. The domain-GO function associations are stored in GUS with the threshold p-value, the rule type used, and links to supporting evidence for the rule, including the similarities and the GO protein associations. A concise version of the algorithm used to generate rules for assigning a GO function to a ProDom or CDD domain is given in Figure 2.

The effects on our method of some nonideal conditions are as follows. Consistently co-occurring domains in GOAPs are unresolved by this method, since we cannot separate their functions. The sliding p-value threshold effectively deals with the incorrect identification of proteins and missing or incorrect annotation. Nonindependent or inconsistent domain behavior is expected to cause our algorithm to select the lowest candidate p-value. When P_d is large enough and the set of GOAPS unused due to high p-values is large enough, we can identify this situation and flag the domain as not meeting our independence and constancy assumptions. Multiple independent domains in a single model should cause similar behavior. In both these cases, we assume that there is no segregation of function by p-value. If this segregation does occur, then our

method would choose the function associated with the set of lower p-values.

To predict GO functions for a novel protein sequence, we perform a BLAST search against ProDom and CDD and then relate any GO functions associated with a domain that has a BLAST p-value/e-value that meets the threshold recorded with the domain-function association. For example, if novel protein *p* has a p-value match to protein domain *d* of 10^{-45}, and domain *d* has a rule based on a p-value of 10^{-10}, then the rule can be applied to protein *p*. However, if the rule was based on a p-value of 10^{-50}, then the rule would not be applied. For cases such as the latter, where the protein-domain similarity p-value is only slightly higher than the domain-rule threshold, we have the option of allowing such assignments to increase coverage. In all cases, we record as evidence the rule(s) used to generate the prediction as well as the corresponding BLAST similarities of the novel sequence against the protein domain database(s).

Materials

Required data include the GO function ontology, GOAPs and their associations, and the chosen protein domain databases. Version 2.61 of the GO function ontology and the GO function associations used in training were obtained from www.geneontology.org. The latest ProDom release (2001.1) was downloaded from ftp://ftp.toulouse.inra.fr/pub/prodom/current_release/. The most recent CDD release (dated June 27, 2001) was obtained from ftp://ncbi.nlm.nih.gov/pub/mmdb/cdd/.

ACKNOWLEDGMENTS

This work was funded in part by grants from the Department of Energy (DE-FG02-00ER62893) and National Institute of Health (RO1-HG01539).

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M. D., et al. 2001. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.* **29:** 37–40.

Bach, I. 2000. The LIM domain: Regulation by association. *Mech. Dev.* **91:** 5–17.

Bateman, A., Birney, E., Durbin, R., Eddy, S.R., Howe, K.L., Sonnhammer, E.L. 2000. The Pfam protein families database. *Nucleic Acids Res.* **28:** 263–266

Bryant, S.H and NCBI Structure Group 2001. NCBI conserved domain database. http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml.

Corpet, F., Gouzy, J., and Kahn, D. 1999. Recent improvements of the ProDom database of protein domain families. *Nucleic Acids Res.* **27:** 263–267.

Davidson, S., Crabtree, J., Brunk, B., Schug, J., Tannen, V., Overton, G.C., and Stoeckert, C., 2001. Data integration and warehousing in genomics: Two case studies. *IBM Systems Journal* **40:** 512–531.

Fleischmann, W., Moller, S., Gateau, A., and Apweiler, R. 1999. A novel method for automatic functional annotation of proteins. *Bioinformatics* **15:** 228–233.

Gene Ontology Consortium, 2001. Creating the gene ontology resource: Design and implementation. *Genome Res.* **11:** 1425–1433.

Hegyi, H. and Gerstein, M. 2001. Annotation transfer for genomics: Measuring functional divergence in multi-domain proteins. *Genome Res.* **11:** 1632–1640.

Henikoff, S. and Henikoff, J.G. 1994. Protein family classification based on searching a database of blocks. *Genomics* **19:** 97–107.

Henikoff, J.G., Greene, E.A., Pietrokovski, S., and Henikoff, S. 2000. Increased coverage of protein families with the blocks database servers. *Nucleic Acids Res.* **28:** 228–230.

Hofmann, K., Bucher, P., Falquet, L., and Bairoch, A. 1999. The PROSITE database, its status in 1999. *Nucleic Acids Res.* **27:** 215–219.

Kawai, J., Shinagawa, A., Shibata, K., Yoshino, M., Itoh, M., Ishii, Y., Arakawa, T., Hara, A., Fukunishi, Y., Konno, H., et al. 2001. Functional annotation of a full-length mouse cDNA collection. *Nature* **409:** 685–690.

King, R.D., Karwath, A., Clare, A., and Dehaspe, L. 2000. Accurate prediction of protein functional class from sequence in the *Mycobacterium tuberculosis* and *Escherichia coli* genomes using data mining. *Yeast* **17:** 283–293.

Myers, E.W., Sutton, G.G., Delcher A.L., Dew, I.M., Fasulo, D.P., Flanigan, M.J., Kravitz, S.A., Mobarry, C.M., Reinert, K.H., Remington, K.A., et al. 2000. A whole genome assembly of *Drosophila*. *Science* **287:** 2196–2204.

Schultz, J., Copley, R.R., Doerks, T., Ponting, C.P., and Bork, P. 2000. SMART: A web-based tool for the study of genetically mobile domains. *Nucleic Acids Res.* **28:** 231–234.

Smith, T.F., Gaitatzes, C., Saxena, K., and Neer, E.J. 1999. The WD repeat: A common architecture for diverse functions. *Trends Biochem. Sci.* **24:** 181–185.

WEB SITE REFERENCES

http://www.allgenes.org; All Genes homepage.

http://blast.wustl.edu; WU-BLAST archives.

http://www.cbil.upenn.edu/GO; Predicting GO functions using protein domains.

http://www.geneontology.org; Gene ontology home page.

http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml; NCBI conserved domain database.

Received November 5, 2001; accepted in revised form February 13, 2002.