BIOINFORMATICS ORIGINAL PAPER Vol. 24 no. 15 2008, pages 1681–1687
doi:10.1093/bioinformatics/btn312

Sequence analysis

Predicting protein function from domain content

Kristoffer Forslund[1,*] and Erik L. L. Sonnhammer[1]

[1]Stockholm Bioinformatics Centre, Stockholm University, 10691 Stockholm, Sweden

Received on April 29, 2008; revised on June 5, 2008; accepted on June 12, 2008

Advance Access publication June 30, 2008

Associate Editor: Burkhard Rost

ABSTRACT

Motivation: Computational assignment of protein function may be the single most vital application of bioinformatics in the post-genome era. These assignments are made based on various protein features, where one is the presence of identifiable domains. The relationship between protein domain content and function is important to investigate, to understand how domain combinations encode complex functions.

Results: Two different models are presented on how protein domain combinations yield specific functions: one rule-based and one probabilistic. We demonstrate how these are useful for Gene Ontology annotation transfer. The first is an intuitive generalization of the Pfam2GO mapping, and detects cases of strict functional implications of sets of domains. The second uses a probabilistic model to represent the relationship between domain content and annotation terms, and was found to be better suited for incomplete training sets. We implemented these models as predictors of Gene Ontology functional annotation terms. Both predictors were more accurate than conventional best BLAST-hit annotation transfer and more sensitive than a single-domain model on a large-scale dataset. We present a number of cases where combinations of Pfam-A protein domains predict functional terms that do not follow from the individual domains.

Availability: Scripts and documentation are available for download at http://sonnhammer.sbc.su.se/multipfam2go_source_docs.tar

Contact: Kristoffer.Forslund@sbc.su.se

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

As the rate of genome sequencing increases, a wealth of new proteins await functional annotation. While fully automated protein annotation may be utopic, the development of tools that narrow the hypothesis space is crucial and possible (Friedberg, 2006). Most forms of semi-automatic protein annotation are in some form transfer methods, which build on finding already annotated proteins related to the query and transferring their annotations to it. Some methods exist that rely on mRNA or protein co-expression data for such annotation transfer (Massjouni *et al.*, 2006; Zhu *et al.*, 2007), but the majority of annotation transfer tools rely ultimately on protein sequence (Friedberg, 2006). These range from methods using high-level sequence features such as localization signals or motifs to simple similarity searches to find annotated homologs. A third class of methods extend existing annotations to maximize internal consistency across the set of annotated proteins. Likewise, methods vary in the way in which these features are integrated and evaluated for inferring transfer. A standard approach is simply to transfer annotations from some fashion of best-annotated BLAST hit (Altschul *et al.*, 1990; see Jones *et al.*, 2005) or from more complex analysis of BLAST or BLAST-like results (Hawkins *et al.*, 2006; Verspoor *et al.*, 2006). From there on, various machine learning approaches have been applied, including but not limited to Bayesian networks (Engelhardt *et al.*, 2005; Nariai *et al.*, 2007), Support Vector Machines (Vinayagam *et al.*, 2004) and several flavors of rule-based classification (Hayete and Bienkowska, 2005; Kretschmann *et al.*, 2001; Schug *et al.*, 2002; Syed and Yona, 2003). Moreover, the approaches for evaluating annotation transfer tools vary considerably, both with regard to benchmarks, annotation systems, and evaluation metrics (Friedberg, 2006; Jones *et al.*, 2007). As a result, there is no solution to this problem that as yet has been universally shown to be effective.

Among relevant sequence features of a protein, domains occupy a key position. They are sequential and structural motifs found independently in different proteins, in different combinations, and as such seem to be the building blocks of proteins above the raw amino acid sequence level (Richardson, 1981). Several approaches to define and delimit different domains have been developed, some based on observed distinct structure classes (Murzin *et al.*, 1995), others on clustering conserved subsequences (Mulder *et al.*, 2007; Sonnhammer *et al.*, 1998). One of the most widely used domain schemata is the Pfam database (Finn *et al.*, 2006; Sonnhammer *et al.*, 1998). At the core of this database are sets of distinct representative sequences, manually selected for each domain family, for which then Hidden Markov Models are generated and are used for domain annotation of the rest of the protein sequence space. Thus, Pfam domain assignment is generally straightforward and achieves good coverage.

Under the assumption that domains are structural protein architecture modules, it makes sense that protein function should follow largely from domain architecture. This in turn would imply that many aspects of such function could be inferred without recourse to more detailed information about the raw sequence (Bashton and Chothia, 2007). Whether this is indeed the case, and if so, what the inner workings of the relationship between domain architecture and function are, is unknown. However, we are beginning to accumulate enough data from diverse sources to be able to test the extent to which this hypothesis is true.

*To whom correspondence should be addressed.

It is also known that certain sets of domains are frequently found together, which may indicate functional cooperation. This conservation of domain context has been the basis for work attempting to improve domain detection by integrating such context information (Beaussart *et al.*, 2007; Coin *et al.*, 2003) and for detecting homologs using domain content similarity (Song *et al.*, 2007). On the whole, it seems clear that domain context is important for protein function.

Attempts have been made to link the terms of the Gene Ontology, which is a widely-used controlled protein annotation vocabulary (Ashburner *et al.*, 2000), to InterPro domains. The InterPro database is a metadatabase of several constituent domain schemata, each with some automated method for determining whether a protein sequence belongs to a given domain family or not (Mulder *et al.*, 2007). This interpro2go map (Mulder *et al.*, 2007) is based on the premise that if annotated proteins possessing a given domain are never found in a trusted training set without a given GO term, that term is implied by the presence of the domain. By subsequently mapping INTERPRO domains to Pfam domains, the Pfam2GO approach thus enables a study of the relationship between Pfam domains in isolation and Gene Ontology functional assignments (Hayete and Bienkowska, 2005; Mulder *et al.*, 2007). Hayete and Bienkowska (2005) built a model using decision trees to predict GO terms from combinations of Pfam domains and other sequence features, and Schug *et al.* (2002) built a similar rule-based predictor using ProDom and CDD domains, both with some measure of success. It is reasonable to assume that a significant number of protein functions arise from the interplay between domains, where a combination of domains implies a certain function with greater specificity than the individual domains. This project presents a model for such interplay, and attempts to chart which such cases exist.

In evaluating this hypothesis on the interaction of domains to form specific functions, we implement it as a simple, standalone function annotation tool, enabling us to test how well it holds up in practice. Clearly, for maximum effect, protein function prediction should make use of all available data, including non-sequence based information such as interaction data. Our purpose here is not primarily to present a full-fledged prediction tool, but rather to present an approach by which such tools may make use of domain architecture information to transfer annotation between distantly related proteins. In contrast to previous work, our present approach makes use only of Pfam-A domain architecture. Moreover, it is more intuitive, computationally faster, and better scalable than previous decision tree-based approaches, lending itself not merely to prediction but also to drawing specific conclusions concerning domain functional interplay.

2 METHODS

This work makes a number of simplifying assumptions. Protein function in this context specifically refers to any Gene Ontology term assigned to a protein. A protein assigned a specific term is also by definition considered to have any ancestors of that term in the GO graph as well. The predictors may assign any term individually regardless of GO level, but if a term is assigned, all its ancestors are also assigned automatically. By the domain architecture content of a protein is meant the set of Pfam-A domains it contains. Thus, both the sequential order of domains and the number of times each domain occurs in a protein is ignored with regard to functional interplay. A domain is simply regarded as absent or present. This is done primarily on grounds of the assumption that the presence or absence of the domain at all is more vital for its functional contribution to the protein than its sequential position, but it also helps avoid building a model too complex. Thus, the domain architecture A*B*B*A*C contains the domain subsets {A}, {B}, {C}, {AB}, {AC}, {BC} and {ABC}. We consider two possible forms of the relationship between the domain set of a protein and its function.

2.1 Strict implication

In analogy with the Pfam2GO approach, we state that a domain (sub-)set with a GO term strictly implies that term if and only if all instances of annotated proteins displaying that set of domains also display the term. This may or may not mean that the properties of those domains in combination causes a protein to have that function. In case a domain set and one of its subsets (such as {AB} and {A}) both imply a functional term, only the smaller subset is considered to imply it, in keeping with Occam's razor and seeking the simplest possible explanations. Thus, a multi-domain set will only predict a GO term if all its domain subsets are supported by at least one training example lacking the GO term. Finding the set of such non-redundant strictly implying domain set-GO term relationships from a set of annotated proteins is fairly straightforward.

The use of strict implication has one important drawback: it can be foiled by false negatives in the training set. A single protein missing a valid GO term annotation is enough to disqualify the domain sets present in its architecture from predicting that term. While this can be avoided partly by using only manually annotated proteins as a training set, this approach still cannot protect from cases where only part of the function of a training set protein is known. Relying only on manually annotated proteins, while prudent at this stage, makes large-scale analysis difficult. Furthermore, domain sets occurring once only will imply any functional terms associated with that single protein. All of these problems will decrease as databases improve, and eventually, strict implication predictions should become stable. Our goal at this stage is to extend the Pfam2GO framework to multi-domain sets, and the resulting predictions could then be manually evaluated.

2.2 Probabilistic approach

The second form of relationship between domain content and function is probabilistic, similar to a Naïve Bayesian network (Friedman *et al.*, 1997). Consider a functional annotation term F (in this case a Gene Ontology term, but the approach would be analogous using another annotation vocabulary) and a domain set D. The probability that a protein exhibiting D would possess F is modeled as

$$P(F|D) = P(D|F)P(F)/P(D) \tag{1}$$

and for the complement $\neg F$, that is, the case that the protein does not possess F

$$P(\neg F|D) = P(D|\neg F)P(\neg F)/P(D), \tag{2}$$

where $P(F|D)+P(\neg F|D)=1$. Thus, we have the odds ratio α as

$$\alpha = P(F|D)/P(\neg F|D) = P(D|F)P(F)/P(D|\neg F)P(\neg F) \tag{3}$$

and

$$P(F|D) = \alpha/(1+\alpha), \tag{4}$$

which is the posterior probability of annotation F given D. From a sufficiently large training set, the prior probabilities of F and $\neg F$ can be estimated. The same may or may not be true for $P(D|F)$ and $P(D|\neg F)$, particularly if the domain set is uncommon. This is where the Naïve Bayesian-like assumption comes in – the distinct sets for which $P(D|F)$ and $P(D|\neg F)$ significantly differ are assumed to occur independently. This is clearly not the case, as any multi-domain protein will contain some domain sets that are subsets to others, but the simplification may nevertheless be sufficiently reasonable that predictions remain possible. Following this, we have

$$P(D|F)/P(D|\neg F) = \Pi P(D_i|F)/P(D_i|\neg F) \tag{5}$$

and the odds ratio

$$\alpha = \Pi(P(D_i|F)/P(D_i|\neg F)) \times P(F)/P(\neg F), \tag{6}$$

where the product is taken over the $i=0..K$ subsets of D. There are $K=2N-1$ such subsets for N unique domains in D.

Table 1. Cross-validation results

Dataset	All				Pfam-A domains only				Curated annotations only			
Dataset size	654 180 proteins				506 315 proteins				31 861 proteins			
	Sens. (%)	Spec. (%)	Prec. (%)	MCC	Sens. (%)	Spec. (%)	Prec. (%)	MCC	Sens. (%)	Spec. (%)	Prec. (%)	MCC
Best BLAST	87.8	>99.9	82.1	0.85	89.6	>99.9	82.5	0.86	38.0	>99.9	42.4	0.40
Pfam2GO	53.3	>99.9	99.6	0.73	65.5	>99.9	99.7	0.81	5.5	>99.9	55.2	0.17
MultiPfam2GO	56.7	>99.9	99.4	0.75	69.7	>99.9	99.4	0.83	7.5	>99.9	52.3	0.20
Probabilistic	69.1	>99.9	93.9	0.81	85.0	>99.9	93.9	0.89	25.9	>99.9	59.3	0.39

The probability model is thus taken as follows: let $f(D_i|F), N(D_i|F)$ be the frequency and number of proteins, respectively, of (sub-)set D_i among proteins with annotation F, and $f(D_i|\neg F), N(D_i|\neg F)$ the corresponding frequency and number of proteins without the annotation. Sampling is done using pseudocounts, that is, all tallies of proteins in different categories from the dataset are incremented by one. Then:

$$P(D_i|F)/P(D_i|\neg F) = f(D_i|F)/f(D_i|\neg F). \quad (7)$$

As a result, what needs to be sampled are frequencies of all domain sets with and without each annotation, as well as the prior distribution $P(F)/P(\neg F)$, which can be taken as the frequency $f(F)/f(\neg F)$, for each annotation F.

2.3 Reducing Bayesian naïvete

While the above model works well, the assumption of independent occurrence of domain subsets is problematic, not least because it will tend to cause false annotation transfers to proteins with many domains. For the reasons presented, explicit handling of dependencies between domain subsets is difficult. We experimented with several variants of the model where this effect would be reduced, including using only the highest-scoring subset for each combination of protein and annotation term, and concluded that using an averaged contribution from each subset improves accuracy with only a small loss of sensitivity (data not shown). This is effectively a form of normalization with the respect to the size of the domain set. Hence, the model is adjusted so that Equation 6 becomes

$$\alpha = (\Pi(P(D_i|F)/P(D_i|\neg F)))(1/K) \times P(F)/P(\neg F). \quad (8)$$

2.4 From models to predictors

To test the usefulness of the above models, we applied them as predictors of Gene Ontology terms (drawn from all three sub-ontologies) under 10-fold cross-validation. Prediction is fairly straightforward: under the strict implication model, all annotations implied by its (sub-)sets are transferred to it, under the probabilistic model, all annotations with posterior probability over a given threshold are also transferred to it. Next, predictions are completed under the Gene Ontology True Path Rule (see ontology description at the Gene Ontology website, http://www.geneontology.org), that is, if a term has been transferred, all its ancestor terms are automatically transferred also.

2.5 Datasets

Primarily, we are interested in the case of transfer between proteins that are evolutionarily well separated. If annotated homologs exist with which the query has nearly full-length sequence identity, there is no point in going beyond a simple BLAST search. Because of this, we chose the UniRef50 nonredundant dataset, which was downloaded on September 3, 2007. It is generated by choosing a reference sequence for all clusters of proteins (mainly taken from UniProt) sharing more than 50% sequence identity (Suzek *et al.*, 2007). Those UniRef50 proteins which had Gene Ontology annotation according to the Gene Ontology Annotation (GOA) database (Camon *et al.*, 2004), and whose representative proteins are present in UniProt (Wu *et al.*, 2006), were used as our test set for a 10-fold cross-validation procedure. The Gene Ontology annotation flatfile used for the predictor performance evaluation was likewise downloaded from the Gene Ontology Consortium on September 3, 2007. Most, but not all, of these sequences have at least one Pfam-A domain and are thus amenable to our analyses. While other studies have shown that functional transfer from sequences that are annotated electronically is error-prone (Jones *et al.*, 2007), the decision was made to include such training examples. This makes the dataset less biased towards extensively studied proteins, as well as large enough for cross-validation to make sense. Domain architectures from version 22.0 of Pfam were used.

Three versions of the dataset were used. The 'All annotations' version is the raw data from Gene Ontology and UniRef50. The 'Curated annotations only' version is the subset which results when excluding Gene Ontology annotations with evidence code IEA (Inferred by Electronic Annotation). Last, the 'All annotations, only proteins with Pfam-A domains' version is the subset where proteins without Pfam-A domains have been excluded. The sizes of the relative datasets are shown in Table 1. For each dataset, all three sub-ontologies (Biological Process, Molecular Function, Cellular Localization) were used.

2.6 Reliability of functional implications

To compute a confidence score (P-value) for the functional implication of an annotation term F by a domain combination D, we utilize the following procedure. It is assumed that D is always found to co-occur with F in the training data. Let D_j be the individual domains (single-domain subsets) making up D.

We sample the frequencies $f(F|D_j\neg D)$ and take

$$P(F|D_j\neg D) = f(F|D_j\neg D). \quad (9)$$

Then, under the null hypothesis that there is no domain functional interplay on D,

$$P(F|D_j) = P(F|D_j\neg D) \quad (10)$$

and so

$$P(F|D) = 1 - \Pi(1 - P(F|D_j)). \quad (11)$$

The number of proteins with set D found with or without F will then be binomially distributed, and we may compute the P-value for each domain combination-annotation term combination as the probability that the observed number would follow from the null hypothesis.

2.7 Mapping of domain combinations to gene ontology terms

The models presented above for the relationship between protein domains and protein functions were used to construct predictors. From the second approach, the extension of Pfam2GO to multidomain combinations,

we generated a collection of domain sets—Gene Ontology term correspondences from all proteins with Pfam-A domains in Pfam 22.0 (which includes 73% of UniProt proteins) that were annotated with at least one Gene Ontology term. The GO annotations flatfile used for the final mapping was downloaded on February 4, 2008, as was the corresponding Pfam2GO flatfile used in the analysis.

A correspondence between a domain set and an annotation term was listed if $P < 0.001$, according to the previous section. These correspondences or predictions were then made non-redundant in several ways. If an implication follows from another because one annotation term is a parent or ancestor of the other, only the more specific term is listed. We computed confidence values for every functional prediction as described previously. If one domain set was a subset of another predicting the same term, only the one with the lower *P*-value was retained. We also excluded any correspondence represented by less than 10 proteins in UniProt, or represented by only one unique Pfam-A domain architecture, as we judged there to be insufficient data in such cases. Last, we excluded any prediction which could be replicated solely from Pfam2GO, as our purpose specifically was to find predictions that can be made from domain combinations but not from single domains.

2.8 Best BLAST annotation transfer

As a full length sequence comparison annotation transfer method, we implemented a simple BLAST-based transfer tool. A protein was assigned the annotations of its best BLAST hit in a training set of GO-annotated proteins, as well as any ancestor terms of these annotations. If there were no hits at the E-value cutoff 1e-6, no annotations were transferred. We used BLASTP in the NCBI BLAST package, version 2.2.16, and left all other parameters at their default settings.

2.9 Performance evaluation statistics

We used the following definitions with regard to the gold standard set used for the testing: TP—true positives, predicted assignments of annotation terms to a protein which are correct. FP—false positives, predicted assignments of annotation terms to a protein which are not correct. TN—true negatives, not predicted assignments of annotation terms which are correct. FN—false negatives, not predicted assignments of annotation terms which are not correct.

Sensitivity is defined as $TP/(TP+FN)$, i.e. the fraction of positive cases that are detected. Specificity is defined as $TN/(TN+FP)$, i.e. the fraction of negative cases that are detected. Precision is defined as $TP/(TP+FP)$, i.e. the fraction of positive predictions that are true. Matthew's Correlation Coefficient (MCC) is a composite score combining the separate criteria tested using the other metrics. It is defined as $(TP \cdot TN - FP \cdot FN)/\sqrt{((TP+FP)\cdot(TP+FN)\cdot(TN+FP)\cdot(TN+FN))}$. It scales between -1 and 1. An MCC score of 1 would mean a perfect predictor, whereas an MCC score of -1 would mean a predictor which is always wrong. A score of 0 means a random predictor.

The analysis was performed using 10-fold cross-validation. The dataset was divided into 10 parts, and predictions were made for each part using the remaining nine as a training set for the domain-based predictors and as a reference database for the BLAST analysis. The final performance evaluation statistics were averaged across the 10 partitions of the data, with very little variation in the results observed between the partitions.

3 RESULTS

We developed two new methods for predicting protein function from domain content. To evaluate the performance of these methods and to compare them to existing methods, we derived three datasets based on the UniProt50 database. This is a non-redundant subset of UniProt at the 50% level, i.e. no protein is more than 50% identical to another one. We also used the subset of UniProt50 with only curated function annotation (which is a small fraction of GO-annotated data), and the subset with assigned Pfam-A domains.

By evaluating the respective predictive capacity of the methods, we assessed the usefulness of their underlying models for the relationship between domain architecture and protein function.

3.1 Evaluation

Four approaches were evaluated, two existing methods and two new ones. First, best-BLAST annotation transfer, where annotations were transferred to a query from its highest-scoring GO-annotated BLAST hit. Second, Pfam2GO, implementing the Pfam2GO approach under the current cross-validation scheme. It should be noted that the publicly available Pfam2GO mapping also undergoes additional manual curation, which is not performed in our implementation at this stage. The first novel method is MultiPfam2GO, which extends Pfam2GO to multiple-domains MultiPfam2GO under a strict implication model. The second novel method is a probabilistic Naïve Bayesian model. Table 1 shows the results in terms of average sensitivity, specificity, precision (or positive predictive value), and the Matthews Correlation Coefficient (MCC) composite score.

All methods had very high specificity (above 99.9%), a consequence of our definition of true negatives. However, the other statistics revealed large differences between the methods. On the complete dataset, BLAST recovered a high fraction of true annotations (87.8%), but at the cost of the highest false positive rate, leading to a low precision (82.1%). Both strict implication methods performed in the extreme opposite fashion, yielding almost perfect precision scores (above 99%), but with very low sensitivity. The probabilistic method performed in the middle of these extremes both in terms of sensitivity and precision.

Limiting the analysis to only proteins with Pfam-A domains, which is where the domain-based methods are at all applicable, rendered these methods more sensitive. The probabilistic method increased its sensitivity to 85%, only 4.6% below BLAST, yet maintained a high precision (93.9%, 11.4% above BLAST). Its utility as a balanced tradeoff between sensitivity and precision (coverage and accuracy) is further shown in its high MCC score, the highest for all datasets and methods. While BLAST remained very sensitive, its precision stayed low even at lower E-value cutoffs (data not shown). There was a significant gap between the sensitivity of the strict implication methods and the probabilistic approach, indicative of the high frequency of missing data in the training set, i.e. proteins that should have a certain annotation but do not yet have it. The probabilistic approach was as expected much better at handling incomplete training data, which is a reality for the foreseeable future. From a perspective of whole-genome annotation using domain-based methods, the results would suggest first the application of the probabilistic method, and flagging those annotations as relatively more reliable that also are reproduced using the strict implication methods.

Comparing the single-domain Pfam2GO method with its multiple domain extension, the gain in sensitivity was smaller than we had expected, typically a few percent. While clearly a fraction of annotation terms can only be inferred from the presence of multiple domains, in most cases there is some domain that always co-occurs with the function. If the domains individually are not found elsewhere, considering the combination would not improve

Fig. 1. Protein domains may encode different functions in different combinations. This example shows several proteins that share many domains, and the proteins' functional annotations. Each protein is marked with its name in bold in the style of a FASTA sequence header, followed by a text description of its putative function. The line below shows the domains as their Pfam-A accession numbers separated by dashes, in the order they occur in the protein. The particular domains making up the predictive combination are highlighted in color. On the line below, the GO annotation terms assigned to the protein are listed (the leaf nodes in the GO Biological Process and Molecular Function categories, taken from UniProt). The box at the top contains proteins with the functional annotations GO:0006094 (biological process gluconeogenesis, orange) and GO:0004736 (molecular function pyruvate carboxylase activity, blue). The domains that predict these functions are PF02786 (Carbamoyl-phosphate synthase L chain, ATP binding domain, turquoise), PF02785 (Biotin carboxylase C-terminal domain, purple), PF00682 (HMGL-like, green), and PF00364 (Biotin-requiring enzyme, red). Below the box we list a number of other domain architectures that the domains occur in, which are associated with other functions. Only in the specific domain combination in the box are the domains associated with the two highlighted GO terms. A total of 175 proteins in the dataset used exhibit this domain combination.

prediction further, even if the function depends on properties of all the domains in the set.

For the dataset with only curated annotations, all methods performed considerably worse. This set is apparently too small for any method to achieve high sensitivity under cross-validation. At the same time, it is known that current electronically annotated data can be error-prone, so the figures on a sufficiently large experimental set may be more realistic.

3.2 Protein function predicted by domain combinations

The performance evaluation indicates that protein function in many cases is produced by a specific domain combination in a more complex way than simply by adding the functions of the individual domains. Exploring these cases would be interesting not least for the purpose of learning how individual domain functions combine to yield more specific functions.

As an example of such an informative combination, the set of domains PF00364 (Biotin-requiring enzyme), PF00682 (HMGL-like), PF02785 (Biotin carboxylase C-terminal domain), and PF02786 (Carbamoyl-phosphate synthase L chain, ATP binding domain) can be mentioned. These four domains in combination are highly specific for the process of gluconeogenesis (GO:0006094) and the associated molecular function of pyruvate carboxylase activity (GO:0004736). While the domains are known to be connected to proteins with this and similar roles, they are also found independently in proteins with other roles, hence the presence of any of these domains alone cannot be used to infer participation in gluconeogenesis. When appearing together, however, we may conclude this functional role reliably. See Figure 1 for some sample proteins where these domains appear, together or in isolation, along with their respective Gene Ontology term annotations. A larger domain set, also including PF00289 (Carbamoyl-phosphate synthase L chain, N-terminal domain) and PF02436 (Conserved carboxylase domain) was also found to predict a gluconeogenetic function, but was pruned from the prediction set as the four domains above formed a more statistically significant subset. As shown in Figure 1, gluconeogenetic proteins of this type display multiple distinct domain architectures; however, the relative sequential order of the domains appears to be conserved.

3.3 Mapping of domain combinations to annotation terms

From our set of 2181143 UniProt proteins with Pfam-A domains and Gene Ontology assignments, we selected a set of 805 statistically significant mappings between 457 combinations of Pfam-A domains and 186 Gene Ontology terms. Note that if a domain combination predicts several annotation terms which are related as ancestor-descendant in the Gene Ontology DAG, only the leafmost one will be included. We present this MultiPfam2GO mapping in an online format similar to Pfam2GO (with the addition of the *P*-value of the inference), and will successively maintain and update it. For certain predictions, the *P*-value is listed as 0, which means that it is smaller than the smallest number the software can handle, which is on the order of 10e-320. The datafile can be found at http://sonnhammer.sbc.su.se/MultiPfam2GO, but is also included here as Table S1. The distribution of domain combination sizes in this prediction set is shown in Table 2.

Table 2. Distribution of domain combination size in mapping set

Number of domains	Number of predicting combinations
2	582
3	161
4	59
5	2
6	1

The distribution of numbers of domains across domain combinations predicting annotation terms in the MultiPFfam2GO mapping set.

4 DISCUSSION

We have presented two simple models, one conservative, one more permissive, for how domains in a protein interplay to produce its function. While neither method can capture all patterns that exist, the approaches are nevertheless useful in integrating and extending the knowledge we already have, and we demonstrate that, at least for more distantly related proteins, our approaches are superior to simple sequence similarity annotation transfer and to single-domain strict implication.

The only previously published work on using Pfam domain combinations for predicting Gene Ontology functions is, to our knowledge, that of Hayete and Bienkowska (2005). It is difficult to make a fair comparison to that method because of very different benchmarking approaches. Still, our method achieved significantly higher sensitivity and precision than reported in that work, in a fraction of the computer time.

We demonstrate that domain functional interplay may not follow directly from the properties of the domains in isolation. We have thus started to unravel the language by which protein function is encoded in a set of protein domains.

As databases grow in coverage and quality, an approach like this, which scales well numerically, is likely to reveal even more mechanisms and relationships, and has the potential of functioning as an important component in automated genome annotation pipeline. To avoid compound errors, a production implementation of this method should ideally be trained on the latest set of curated proteins available at the time.

Future work will likely include integrating predictors based on these models in the form of a web service, queryable interactively or in bulk. Another potential area of exploration is to extend this framework from merely the use of domain sets to also take into account conservation of sequential order of the domains, and to investigate to which degree such conservation is important for predicting protein function.

Funding: A grant from Pharmacia.

REFERENCES

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Ashburner,M. *et al.* (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Bashton,M. and Chothia,C. (2007) The generation of new protein functions by the combination of domains. *Structure*, **15**, 85–99.

Beaussart,F. *et al.* (2007) Automated Improvement of Domain ANnotations using context analysis of domain arrangements (AIDAN). *Bioinformatics*, **23**, 1834–1836.

Camon,E. *et al*. (2004) The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with gene ontology. *Nucleic Acids Res*., **32**, D262–D266.

Coin,L. *et al*. (2003) Enhanced protein domain discovery by using language modeling techniques from speech recognition. *PNAS*, **100**, 4516–4520.

Engelhardt,B.E. *et al*. (2005) Protein molecular function prediction by Bayesian phylogenomics. *PLoS Comput. Biol*., **1**, e45.

Finn,R.D. *et al*. (2006) Pfam: clans, web tools and services. *Nucleic Acids Res*., **34**, D247–D251.

Friedberg,I. (2006) Automated protein function prediction–the genomic challenge. *Brief Bioinform*., **7**, 225–242.

Friedman,N. *et al*. (1997) Bayesian Network Classifiers. *Machine Learning*, **29**, 131–163.

Hawkins,T. *et al*. (2006) Enhanced automated function prediction using distantly related sequences and contextual association by PFP. *Protein Sci*., **15**, 1550–1556.

Hayete,B. and Bienkowska,J.R. (2005) GOTrees: Predicting GO associations from protein domain composition using decision trees. *Pacific Symp. Biocomp.*, **2005**, 140–151.

Jones,C.E. *et al*. (2005) Automated methods of predicting the function of biological sequences using GO and BLAST. *BMC Bioinformatics*, **6**, 272.

Jones,C.E. *et al*. (2007) Estimating the annotation error rate of curated GO database sequence annotations. *BMC Bioinformatics*, **8**, 170.

Kretschmann,E. *et al*. (2001) Automatic rule generation for protein annotation with the C4.5 data mining algorithm applied on SWISS-PROT. *Bioinformatics*, **17**, 920–926.

Massjouni,N. *et al*. (2006) VIRGO: computational prediction of gene functions. *Nucleic Acids Res*., **34**, W340–W344.

Mulder,N.J. *et al*. (2007) New developments in the InterPro database. *Nucleic Acids Res*., **35**, D224–D228.

Murzin,A.G. *et al*. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol*., **247**, 536–540.

Nariai,N. *et al*. (2007) Probabilistic protein function prediction from heterogeneous genome-wide data. *PLoS ONE*, **2**, e337.

Richardson,J.S. (1981) The anatomy and taxonomy of protein structure. *Advances Protein Chem.*, **34,** 246.

Schug,J. *et al*. (2002) Predicting gene ontology functions from ProDom and CDD protein domains. *Genome Res*., **12**, 648–655.

Song,N. *et al*. (2007) Domain architecture comparison for multidomain homology identification. *J. Comput. Biol.*, **14**, 496–516.

Sonnhammer,E.L. *et al*. (1998) Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res*., **26**, 320–322.

Suzek,B.E. *et al*. (2007) UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics*, **23**, 1282–1288.

Syed,U. and Yona,G. (2003) Using a mixture of probabilistic decision trees for direct prediction of protein function. In the proceedings of *RECOMB*, pp. 224–234.

Verspoor,K. *et al*. (2006) A categorization approach to automated ontological function annotation. *Protein Sci*., **15**, 1544–1549.

Vinayagam,A. *et al*. (2004) Applying support vector machines for gene ontology based gene function prediction. *BMC Bioinformatics*, **5**, 116.

Wu,C.H. *et al*. (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res*., **34**, D187–D191.

Zhu,M. *et al*. (2007) Globally predicting protein functions based on co-expressed protein-protein interaction networks and ontology taxonomy similarities. *Gene*, **391**, 113–119.