J Struct Funct Genomics (2011) 12:9–20
DOI 10.1007/s10969-011-9105-3

A new approach to assess and predict the functional roles of proteins across all known structures

Elchin S. Julfayev · Ryan J. McLaughlin · Yi-Ping Tao · William A. McLaughlin

Received: 17 December 2010 / Accepted: 14 March 2011 / Published online: 29 March 2011

Abstract The three dimensional atomic structures of proteins provide information regarding their function; and codified relationships between structure and function enable the assessment of function *from* structure. In the current study, a new data mining tool was implemented that checks current gene ontology (GO) annotations and predicts new ones across all the protein structures available in the Protein Data Bank (PDB). The tool overcomes some of the challenges of utilizing large amounts of protein annotation and measurement information to form correspondences between protein structure and function. Protein attributes were extracted from the Structural Biology Knowledgebase and open source biological databases. Based on the presence or absence of a given set of attributes, a given protein's functional annotations were inferred. The results show that attributes derived from the three dimensional structures of proteins enhanced predictions over that using attributes only derived from primary amino acid sequence. Some predictions reflected known but not completely documented GO annotations. For example, predictions for the GO term for copper ion binding reflected used information a copper ion was known to interact with the protein based on information in a ligand interaction database. Other predictions were novel and require further experimental validation. These include predictions for proteins labeled as unknown function in the PDB. Two examples are a role in the regulation of transcription for the protein AF1396 from *Archaeoglobus fulgidus* and a role in RNA metabolism for the protein psuG from *Thermotoga maritima.*

Keywords Protein function prediction · Gene ontology · Three-dimensional structure

Introduction

The number of available three dimensional protein structures has increased rapidly over the past decade, due in part to the work of the Protein Structure Initiative (PSI) [1]. Structural variety has also increased as there has been a systematic effort by the PSI to cover the various types of protein structures found in nature [2–4]. Representatives from protein sequence families are selected that are likely to have structures different from those already available in the Protein Data Bank (PDB) [5]. A counterpart to the increase in structural variety is an increase the number of different functions associated with the structures, and the breadth of different functional categories represented has expanded [6, 7].

Protein structure can dictate function [8], and the correspondences between protein structure and biological function provide a means to automatically assess function *from* structure [9–13]. Given an under-characterized protein, structural similarity can first be detected with that of known function. If the similarity is high enough then functional equivalence can be inferred; and the functional annotation from the characterized protein can be

Electronic supplementary material The online version of this article (doi:10.1007/s10969-011-9105-3) contains supplementary material, which is available to authorized users.

E. S. Julfayev · R. J. McLaughlin · W. A. McLaughlin (✉)
Department of Basic Science, The Commonwealth Medical College, 525 Pine Street, Scranton, PA 18509, USA
e-mail: wmclaughlin@tcmedc.org

Y.-P. Tao
Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, 610 Taylor Road, Piscataway, NJ 08854-8087, USA

transferred. Computational methods are available to detect structural similarity and to aid with the transfer of functional annotation. Example online servers that provide predicted annotations using a host of structural comparison and annotation methods include MarkUS [14], ProKnow [15], and ProFunc [16]. These servers identify specific structure/function correspondences. ProKnow and ProFunc consider these correspondences and additionally predict protein function based on the Gene Ontology (GO) classification system [17].

GO provides a systematic means to partition functional space, and it has the advantage of being both machine and human readable. Applications used for functional annotation based on GO terms using three-dimensional structural information have been the subject of many reviews [11, 18, 19]. Comparable methods are available to predict protein function according to GO terms using the information that can be retrieved based only a protein's primary amino acid sequence [20, 21]. Information from primary sequence and three dimensional structure provide complementary contributions to prediction models of protein function [22]. Overall prediction accuracy is expected to be highest when both forms of information are used together.

Methods for the assessment of a protein's function based on three dimensional structure and primary amino acid sequence can be improved in different ways. The accuracy of the correspondences between structure and function can be improved, and more structure/function correspondences can be identified. Fundamental questions remain as to how and where knowledge of a protein's three dimensional structures can extend that which can gained through sequence alone for the application of functional assessment and prediction. For what functions is either information sufficient? Where can the combination of structural and sequence information be used to improve functional prediction accuracy over that obtained using the information that can be retrieved based on the primary sequence alone?

In the current study, a systematic approach was made to evaluate functional annotations across the entire set of protein structures in the PDB. Protein attributes were assembled from the PSI Structural Biology Knowledgebase (SBKB) [23] and other open biological databases. These attributes were examined with regard to their presence in a group of structures that had a given GO term annotation versus structures those that were not assigned with that GO term. The Adaboost classification algorithm, as implemented within the icsiBoost program, was used to identify those protein attributes that differed between the two groups and created classification models [24, 25]. The classification models that were created for all of the GO terms were evaluated against all of the structures in the PDB. As examples of the utility of the predictions made, a review was done for predictions made for protein structures labeled in the PDB as unknown function.

Materials and methods

Assembly of protein attributes

The assembly of protein attributes for each protein structural chain in the PDB was divided into three automated steps: data download, parsing, and integration. The computer programs for the analysis were written in Python, and there were supplementary scripts written in BASH shell and SQL. A list of the programs and their utilities are given in the supplemental data and are available upon request. Downloaded data file formats included csv, tsv, XML, owl, sdf and txt files. The time stamp for the current releases from each data source was 1 Sep 2010.

The following protein attributes were assembled. Cellular and biochemical pathway assignments were extracted from BioCyc [26], CellMap [27], HumanCyc [27], INOH [28], and the NCI Pathway Interaction Database (PID) [29]. Small molecule associations were obtained from BioCyc, BindingDB [30], HumanCyc, DrugBank [31], ChEBI [32], ChEMBL [33], and SMPDB [34]. A common nomenclature for the small molecules called the InChIKey was used, where available, across the small molecules resources of ChEBI, DrugBank, ChEMBL, and SMPDB. The resources SNPs3D and MIM provided disease associations of the protein structures [35–37]. Molecular functions, biological processes, cellular locations were based on the Gene Ontology classification system as assigned in SIFTS [38]. Enzyme classifications were as assigned as in the EC2PDB database [11, 39]. Structural domains were identified according to the databases CATH [40] and SCOP [41]. Sequences domain assignments were identified through the Pfam resource [42].

Groups of structurally related proteins were found based on the jFatCat alignment algorithm [43, 44]. At the time of download of precomputed jFatCat structural comparisons as available from the PDB, there were 18,590 groups of structures that corresponded to the number sequence clusters that had more than 40% identity in sequence. The precomputed structural comparisons had been done in an all-versus-all fashion across the representatives from the sequence clusters. To estimate the probability of structural similarity, a Bonferroni correction was made that divided a normally acceptable threshold P value of 0.01 for a single comparison by the total number of comparisons. That gave a P value threshold to detect significant structural similarity of 5.37×10^{-7}. All structures below that threshold were kept in the structural comparison group, and the name of each group was based on the name of the structural

representative of that group. The FEATURE resource provided predictions of functional sites within the protein structures [45]. The prediction models of FEATURE were run against all the structures in PDB. A given functional site was assigned to a given structure if the structural similarity of the model to the structure was higher than a threshold value of similarity at which the model had more than 99% specificity.

Dataset generation

GO term classification was done on the protein structure chain level for all protein structural chains in the PDB. Representative chains were selected from groups of amino acid sequences that were exact matches and 100% identical. A study by Devos et al. [46] demonstrated that proteins with higher than 95% sequence identity can vary with regard to their annotations. Representatives at 100% sequence identity were used here to limit the loss of different strings of annotations associated with the structures when classification models were created. Of the total of 155,269 sequences chains in the PDB that had counterparts available as three dimensional structures, 45,803 nonidentical representative chains were selected. Gene ontology term assignments for these representative chains were found in an automated manner. The assignments were based on correspondences between Pfam domains and GO terms as available from the Pfam resource. If a protein structural chain had a given Pfam domain, it was associated with the corresponding GO term.

All chains assigned to a given GO term were collected and referred to as the positive set. All attributes were retrieved for these chains. Values for each of the attributes were found. Pfam domains were excluded as attributes if they had a direct association with the given GO term according to the information from the Pfam database. A negative set contained ten times the number of chains as the positive set. Chains in the negative set were randomly selected from all those that were not known to be associated with the given GO term. In the selection process, all chains that did not have the given GO term were assigned as the initial set of candidates. Each of these candidates was then checked against the following two exclusion criteria. Candidates that had a known GO term assignment that fell below the given GO term with regards to an ancestral lineage of the GO term hierarchy were excluded. Candidates that had Pfam assignments that had a direct correspondence with the given GO term were also excluded. A random number generator was used to choose instances for the negative set from the remaining candidates. Candidates were added to the negative set until their number set reached ten times that of the positive set.

The process of creating positive and negative sets of protein chains with attribute values was done for all of the GO terms. Terms that had ten or greater members in the positive set were kept for further analyses. Of the total number of 1,105 terms in the GO hierarchy, 655 GO terms had greater than or equal protein chain members that were non-identical in sequence as available from the PDB.

Prediction model generation

For each of the GO term datasets assembled, the Adaboost classification algorithm, as implemented within the icsiBoost program [24], was applied. Java code was implemented to run each GO term dataset so as to pass the information about the attribute values of the protein chains from the positive and negative sets. In each of the iterations of boosting in the learning cycle of the Adaboost algorithm, the icsiBoost program used a decision stump learner to assess each attribute value with regard to its presence or absence in the positive and negative sets. The percentages in the positive and negative sets for each attribute value were translated into rules for deciding to which set a given protein belongs given the attribute value. Classifications models consisted of the accumulated set of rules based on the attribute values. A ten-fold cross-validation procedure was implemented in Java to test the accuracy of each classification model. Each fold utilized 90% of the positive and negative sets for training and 10% for testing. The ratio of positive and negative instances in the training sets and test sets were kept in same proportion, 1:10. The statistical parameters of sensitivity, specificity, positive predictive value, overall accuracy, and the area under the receiver operator curve were calculated to estimate each classification model's accuracy.

The following provides further details regarding the classification algorithm. A decision stump learner was applied to modify the weights of the rules found upon each iteration of boosting. The coefficients (a_1, a_2, …, a_M) were calculated to find the contribution that each iteration had to the final classification model. The decision formula of the final model was:

$$G(x) = \text{sign}\left[\sum_{m=1}^{M} a_m \cdot G_m(x)\right]$$

The number M is the total number of iterations, and x is the set of attributes or classifiers x_1, x_2, …, x_N.

The following formula was used to calculate the probability of assigning to the class C given that model had set of attributes.

$$P(C|x_i) = \frac{1}{1 + e^{-2\sigma}}$$

The value σ is the final score estimating the assignment of a structure to the class C. On each iteration step, icsiBoost reduced the value calculated for a stronger attribute in order to get an increase in the contribution for the next attribute. All significant attributes with their values contributing to the final score were extracted from the program output. The total probability for a prediction to belong to each GO term was found for each protein chain.

Predictions of GO term associations across the entire PDB

All of the attributes present in the structures that had a given GO term formed the attribute complement for that term. Values of the attribute complement of the given GO term were found for all the chains in the PDB. The classification model created for each GO term was used to classify each chain as associated or not associated with that term based on the values of the chain's attribute complement. Potentially new predictions were evaluated in two steps. The first step was to remove chains already known to have the given GO annotation. These known associations were either through direct assignment or inferred through the GO term hierarchy: if there was assignment lower in the hierarchy then the chain had the given term. The second step was to include only predictions above a threshold probability of 0.95. A schematic of the method is presented in Fig. 1. Standard data mining techniques were used. These include preprocessing, e.g. extract, clean, refresh; integration which entailed mapping of each attribute to each protein chain; database creation; and analysis which entailed classification model generation and application [47].

Results

Classification model statistics and summary of predictions

Each of the classification models created for the GO terms was evaluated by a ten-fold cross-validation. Accuracy measurements were averaged across all the 655 models that were created for GO terms with greater than ten non-identical protein chain members from the PDB. The average measurements of sensitivity, specificity, positive predictive value (PPV), and area under the receiver operator curve are presented in Table 1. The classification models created with information from structure plus sequence were compared to those created with primary sequence information alone. Structure plus sequence included information from the jFatCat, FEATURE, SCOP, and CATH resources in addition to the attributes from the resources based on the protein sequence information only. For the area under the receiver operator curve measurement, the overall P value for the significance of the difference between the models generated with structure plus sequence versus sequence alone information, based on a student's t-test, was 1.9×10^{-11}. The sample size for the comparison was 655, which corresponded to the number of GO terms that had greater than or equal to 10 representative members. The result indicated that the addition of structural information improved the accuracy of the classification models.

The accuracy measurements were broken down according to the level within the GO hierarchy at the first thirteen levels. Results for the positive predictive values for those levels are shown in Fig. 2. The positive predictive value measures the percentage of correct positive predictions, that is the percentage of structures correctly predicted to have a given GO term annotation. The plot indicates that the information from structure complements that from sequence across all the different levels, and the ability to accurately predict functions at the different levels of functional granularity is improved using structural information. Results for the values of the area under the receiver operator curves (AUC) for the different GO term levels are shown in Fig. 3. The AUC provides a measure of the accuracy of positive and negative predictions, i.e. predictions to have or to not have a given GO term annotation. Across the different levels of the hierarchy, the classification models created with the structure plus sequence information outperform those created with sequence alone with regard to the ability to discern whether *or not* a structure has a given GO annotation.

Summaries of the number of predicted GO annotations at or above 95% probability are presented in Fig. 4 and in Table 2. With the information provided by structure and sequence, the classification models produced 48,829 GO annotation predictions that were made across the entire PDB, and 454 predictions were made for structures labeled in the PDB as unknown function. Using information derived only from protein sequence, the number of predictions across the entire PDB was 53,748; and there were 251 for the subset of structures with unknown function. A comparison of the predictions was made for each protein chain as to whether it was predicted by a model derived from information from structure and sequence or sequence alone. The number of predictions made with structure and sequence but not with sequence alone was 11,854 for the entire PDB and 254 for proteins with unknown function. The percentage of the predictions for structures with unknown function was 2%. In contrast, the number of predictions made with sequence alone and not made using structure and sequence information was 16,782 for the entire PDB and 51 for structures of unknown function. The percentage of predictions for proteins with unknown

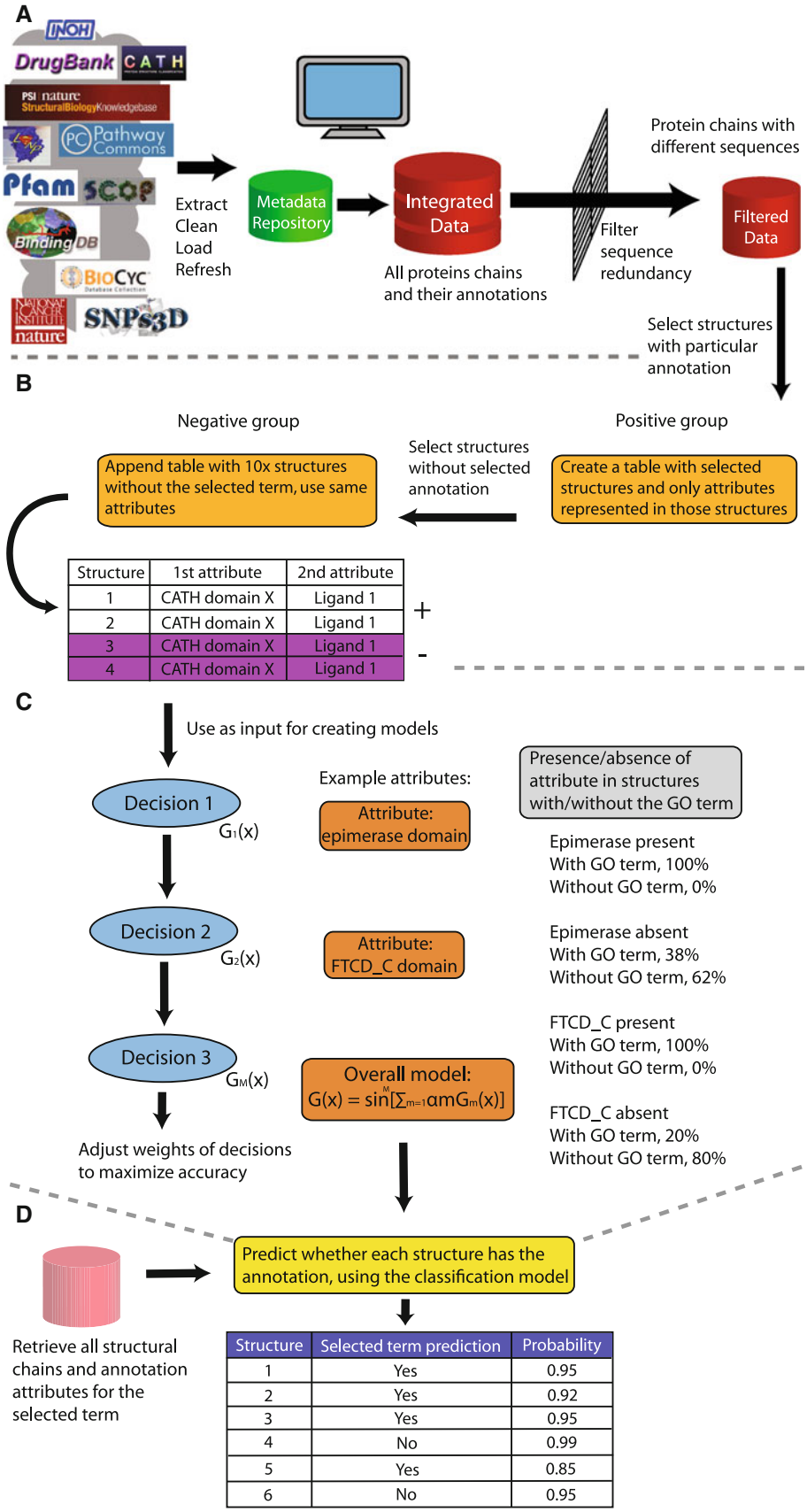

Fig. 1 Schematic of the method used to generate the classification models. The method can be divided into the following steps: **a** data assembly, which includes preprocessing and integration of protein attributes; **b** selection of data sets for classification model generation based on known Gene Ontology term associations; **c** generation of classification models for each GO term; and **d** application of the classification models to predict new GO associations

Table 1 Statistics for analysis of the classification models created for 655 GO terms that had greater than ten non-identical structural protein chain members

	Sensitivity		Specificity		PPV		Overall accuracy		Area under ROC	
	A	B	A	B	A	B	A	B	A	B
Average	0.8843	0.8495	0.9977	0.9947	0.9790	0.9562	0.9886	0.9830	0.9397	0.9198
SD	0.1279	0.1634	0.0042	0.0085	0.0319	0.0623	0.0115	0.0173	0.0653	0.0847

The results are averages for ten-fold cross-validations for the classification models created with information from structure plus sequence (A) and sequence alone (B). *ROC* Receiver operator curve, *SD* standard deviation

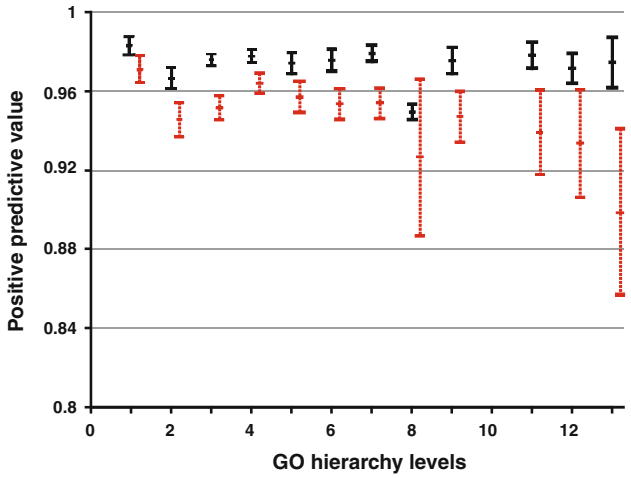

Fig. 2 Plot of the average positively predictive values of the classification models versus the levels of the GO hierarchy. The *dotted* series (*red*) are the results for using only attributes derived from the proteins' primary amino acid sequences. The *solid* series (*black*) are the results for using attributes derived from information from the protein three dimensional structure and primary sequence. The *dotted* series are moved slightly to the right in order to better show the overlay. Results are for the averages for the 10-fold cross-validation tests

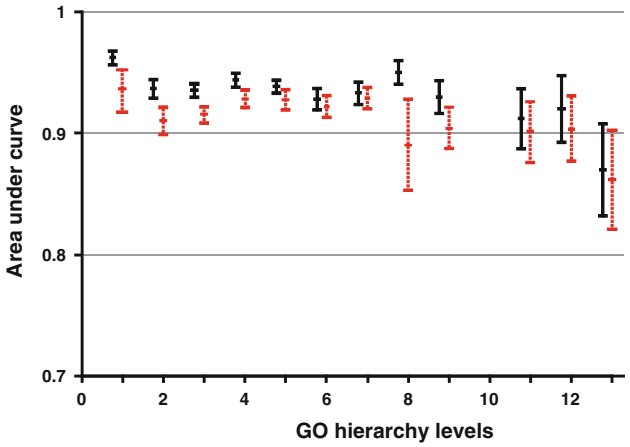

Fig. 3 Plot of the average area under the receiver operator curve measurements at each level of the GO term hierarchy. The *dotted* series (*red*) are the results of using only attributes derived from the proteins primary amino acid sequence. The *solid* series (*black*) are the results when using attributes derived from the primary sequences and three dimensional structures. The *dotted* series are moved slightly to the right in order to better show the overlay. Results are for the averages for the 10-fold cross-validation tests

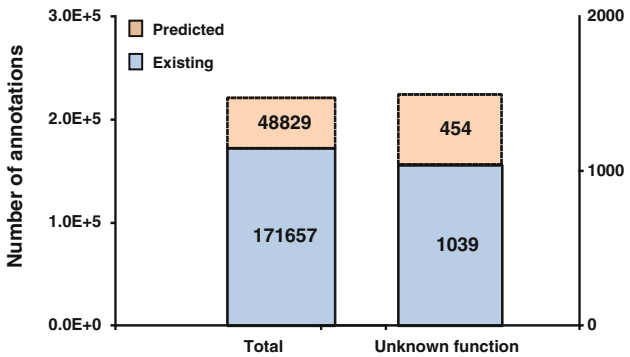

Fig. 4 The number of new high confidence annotation predictions for the entire set of structures in the PDB and for the subset of structures with unknown function. The results are for the predictions of GO term annotations with a probability greater than 95% using information derived from information from both three dimensional structure and primary amino acid sequence

function proteins was 0.3%. A protein structure of unknown function was approximately seven times more likely to be characterized using the models generated with structure plus sequence information as compared to those using information only from knowledge of the protein's sequence, 2% divided by 0.3%. The result indicates that structural information makes a larger relative contribution to the characterization of protein structures of unknown function.

A utility of using information form structure and sequence rather than sequence alone is apparent in the data presented Table 2. The group "B not A" are predictions using information from sequence only that were not found among those predictions made using information both structure and sequence. The observation of "B not A" predictions indicates that these predictions were negated after structural information was added to that from sequence to create the classification models. The result is consistent with the observed higher positive predictive value (PPV) that is associated with classification models created with structure and sequence information, as compared to those created with information from sequence alone. Overall, the

Table 2 Total number of predicted GO annotations across the entire set of structures in the PDB and for structures labeled as unknown function

Statistics	Previously known	A- Structure and sequence	B- Sequence only	A not B	B not A
Total	171,657	48,829	53,748	11,856	16,775
Unknown function	1,039	454	251	254	51

The totals for the classifications models created with information from the protein three dimensional structure and primary sequence attributes given in A. Predictions made using information based on the sequence only are listed in column B. The numbers predicted with information from structure plus sequence but not with sequence alone are listed in the A not B column. Predictions made using sequence alone and not with structure plus sequence are listed in the B not A column

"B not A" predictions are not as viable as those made using information from both structure and sequence.

The classification models created with information from sequence and structure had a higher sensitivity than those created with information from sequence alone. Compare columns A and B for the sensitivity measures in Table 1. Further as shown in Fig. 5, there was a significantly higher sensitivity for models created with sequence and structural information, based on paired *t*-tests, at similarity levels of 100, 95, 90, 80, 70, 60, 50, and 40% identity. At 30% sequence identity and below, the average performance of models created with sequence and structural information were not significantly different than the average performance of the models created with sequence information alone.

The difference in the sensitivity measures was also found to *increase* when sequence similarity between the target protein and the known instances in the training set of the given classification model *decreased* from 100% sequence identity to 50% identity. See Fig. 5. At the level 50% was observed the maximum of the differences with the mean difference of 0.13482 ± 0.196. At the level 100% the difference was 0.02805 ± 0.0604. A paired *t*-test between the differences between the methods at the levels 50 and 100% was significant, and the *P* value calculated for this difference was 0.002070. The observation that structural information makes a relatively larger contribution to the prediction sensitivity as sequence similarity is decreased indicates that structural information extends that from sequence to a greater extent at 50% identity as compared to 100% identity. A reason is that different sequences can adopt the same three-dimensional structure [48–50].

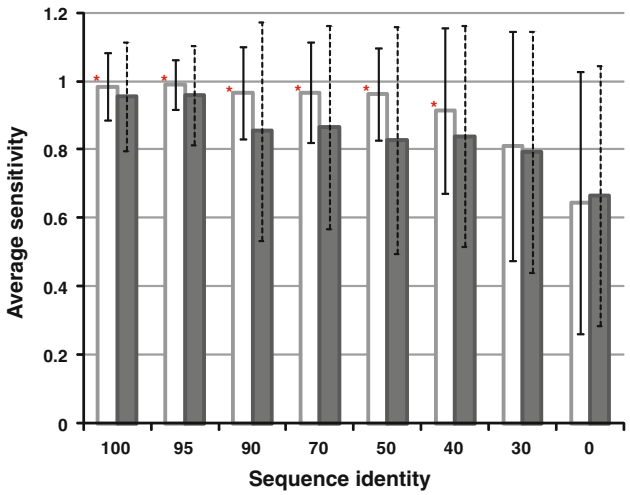

Fig. 5 Plot of the average of sensitivity of the classification models at different levels of protein sequence identity. The series in *gray* are for the classification models created using information from protein sequence alone. The series in *white* are for the classification models created using structure plus sequence attributes. *Asterisks* indicate where there was a significant difference between the average sensitivity for a models created with information from sequence and structure versus that of the models created with sequence alone at the given sequence similarity level. One test set for each GO term was selected for the analysis

For the analysis presented in Fig. 5, sequence identity was estimated by extracting pairs of protein chains in clusters of sequences at a given level of sequence identity [51], as provided at the URL <ftp://resources.rcsb.org/> [44, 52]. The level of sequence identity was determined by the closest match between the given protein and any protein in the training set that was used to create the classification model. Each classification model for each GO term was generated using 90% of the positive and negative examples. The models were applied to the remaining 10% side aside as the test sets. One test set was used for each GO term. Calculations were done across all GO terms with greater than 10 nonidentical members.

Example predictions

The following are examples of predictions of GO term annotations for the protein structures that are labeled in the PDB as unknown function. These protein structures have been solved through the PSI and are listed in the functional sleuth section of the SBKB at the URL <http://http://sbkb.org/KB/unkstrucs.txt>. Predictions may provide leads or clues to facilitate further experimental characterization of the structures of unknown function, as outlined by an NIH notice (NOT-GM-08-123).

The structure of the protein APE2225 from *Aeropyrum pernix K1*, PDB+2ns9, is predicted to participate in the

biological process called response to biotic stimulus, GO+0009607. As evidence that the prediction is correct, the protein was reported to have structural similarity with proteins that have the Bet v onefold that are involved in stress response and defense [53]. The protein lp_2219 from *Lactobacillus plantarum*, PDB+3hfq, is predicted to have the molecular function of amine dehydrogenase activity, GO+0030058, and to participate in the process of methylamine metabolism, GO+0030416. Evidence to support the prediction is that the protein was picked up in a screen for genes that contribute to nitric oxide generation in that organism [54].

A third prediction is for the crystal structure of the protein AF1396 from *Archaeoglobus fulgidus*, PDB+2nwi. It is predicted to participate in the process of regulation of DNA-dependent transcription, GO+0006355. The protein has been demonstrated to be remotely related in sequence, through a reciprocal PSI-BLAST search, to the UbiC transcription regulator-associated (UTRA) domain that resides within GntR transcription regulators and other proteins [55]. The UTRA domain binds to different cellular ligands, e.g. histidine, sugars and fatty acids, to activate or repress transcription in response these ligands [55]. A comparison between the UTRA binding domain from the GntR transcription regulator from *Bacillus anthracis*, PDB+3lhe, with the structure of AF1396 is shown in Fig. 6. The overall topology of the proteins and the placement of the ligand binding pockets are structurally similar.

The solution crystal structure of the protein psuG (TM1464) from *Thermotoga maritima* [56], PDB+1vkm, was predicted to participate in the metabolism of RNA, GO+0016070. The protein functional role was initially described as a possibly being involved in the biosynthesis of the blue pigment indigoidine [56, 57], but that was subsequently stated to be incorrect [58]. Evidence that the prediction that psuG has a role in RNA metabolism comes from a study that demonstrated the protein is a pseudouridine-5′-phosphate glycosidase [57]. Also, a protein YeiN, which is homologous of the psuG protein, is also involved in the hydrolysis of pseudouridine as part of the breakdown of RNA [59]. The finding that the homolog of YeiN is involved the catabolism of RNA further corroborates the prediction that the psuG is involved in RNA metabolism.

The protein structure shikimate 5-dehydrogenase orthologue YdiB [60], PDB+1npd, is predicted to have the molecular function of 3-dehydroquinate dehydratase activity, GO+0003855. Evidence from the literature that support the prediction includes the following. An investigation of the crystal structure and biochemical characterization was done for a novel shikimate dehydrogenase, protein HI0607 [61]. HI0607 protein's crystal structure is similar to that of AroE and YdiB, and it catalyzes the oxidation of shikimate with $NADP^+$. The YdiB protein was found to be a bifunctional enzyme that catalyzes the reversible reductions of dehydroquinate to quinate *and* dehydroshikimate to shikimate in the presence of either NADH or NADPH [61]. An attribute that was used for the prediction of the 3-dehydroquinate dehydratase activity of YdiB was its assignment to the binding of 3-dehydroquinate, as given in ChEBI [32]. ChEBI derived that information from KEGG enzyme database [62], based on the EC number for quinate/shikimate dehydrogenase of 1.1.1.282 [63]. The example highlights the complexity of functional assignment, and the importance of integration of information from different data sources. YbiB is a bifunctional enzyme with different but related enzymatic activities.

A full list of all the predictions is presented in the supplement and available for download that the URL < http://204.139.53.100/KRole-downloads/>.

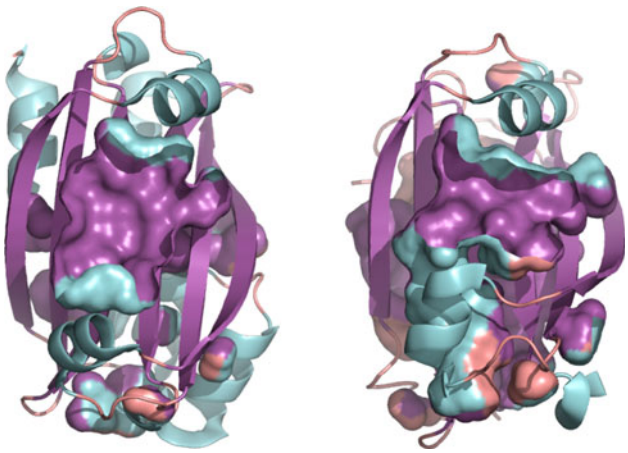

Fig. 6 Comparison of the structure of the protein GntR (*left*), a protein with known transcription regulatory activities in response to cellular ligands, and the structure of AF1396, which was predicted to have transcription regulatory activity. The overall fold as depicted in the coloring based on secondary structure elements is similar between the two proteins. Further, the cavities of the ligand pockets, which are shown as surfaces on the front of the models, are similarly placed

Discussion

Computational challenges and bottlenecks addressed

The methods and technologies utilized here overcame some of the challenges of prediction of function from structure and sequence. The enabling technologies included a large scale integration of protein attributes from varied biological resources as available in the SBKB and other databases, the use of state of the art data mining algorithms, and new computer code implementations that allowed for large datasets of proteins and attributes to be analyzed. A result of

the large scale integration effort was that connections between varied resources allowed for annotations that are synonymous with GO term annotations to be identified. The integration of information from varied sources provides a means to find equivalence in meaning between resources and a way to infer the synonymous annotation counterparts based on the GO term nomenclature. For example, for the GO term copper ion binding, GO+0005507, there were 956 known annotations. An additional 1,536 annotations were inferred based on copper ion associations with protein chains as available from the ligand interaction resources ChEBI, ChEMBL, SMPDB, and DrugBank. Integration thereby provides a means to identify synonymous annotations. The observation highlights the need to expand the annotations of GO terms based on information in primary biological databases.

The following are two examples of new code implementations and their utility. The number of iterations used for the Adaboost classification algorithm have been shown empirically to minimize classification error as the number of iterations is made large [64], e.g. setting the number to 1,000 or higher. For some classification models of the GO terms, a run of 1,000 iterations was more than that required to minimize errors. A Java class was created to monitor the training and test errors as the iterations progressed. When test set error remained unchanged or increased over a range of four hundred iterations, the iterations were discontinued. The total computation time was reduced to 3 weeks of 3 weeks as compared to 2 months without the early stopping procedure implemented. Calculations were done on a local high performance computer cluster (HPCC) that had 136 processors and an average RAM of 1.6 GB. The icsiBoost program itself also provided advances that enabled the study. A parallel version of the program allows iterations of the boosting analyses to be run in parallel.

Ranking of the attributes used in the classification models

The following describes the contribution of the different biological resources to the generation of the classification models. The conditional probabilities of either having a given GO term or not was evaluated for each attribute in the attribute complement for that GO term. See "Materials and methods" for details. The probability to assign a given GO term annotation, referred to as class C, to a give protein structure that has some attribute A_i is the following.

$$P_p(C|A_i) = 1/(1 + e^{-2S_p})$$

S_p is a contribution score given the presence of attribute A_i. P_p is a probability by presence of attribute A_i to assign GO term annotation, class C. The probability P_a of not assigning a GO term annotation given that structure does not have the attribute A_i was calculated by the similar way. The ranking score was calculated as a product of P_p and P_a, and the result was multiplied to 10,000 in order to present it in a convenient form. For each data source the highest ranking score found across all of the attributes across all of the GO terms was selected. In the Table 3, statistics regarding rank for each data source are presented. In addition to the ranking scores, the average numbers of attributes from each data source per structure are given. The frequencies of use of the attributes from each data source, as used for the classification models, are also shown. Sources such as FEATURE have relatively low number of attribute assignments per structure but contribute a relatively larger degree to the classification models, as manifested by their relatively high ranking scores. The result indicates that when they are available these resources contribute relatively more to the classification models used for the prediction of function.

Comparison to other functional prediction methods

Example methods that make predictions of GO term annotations of proteins based in part on structural information are ProFunc and ProKnow. Some differences with these methods with the current method, which is referred to as *k*nowledge prediction of a functional *role* or Krole, are outlined. Overall the Krole method provides an estimate of the prediction probability which varies from 0 to 1. Such estimates are relevant to experimentalists when a protein is checked for possible functional assignments [65]. ProKnow and ProFunc provide a likelihood score that is not normalized to a probability estimate for each GO term prediction. In addition, a conceptual difference between Krole and ProKnow is regarding which predictions are presented. ProKnow presents all predictions where there was no direct assignments of a protein with a given GO term annotation. A prediction was not considered here if the predicted term was an ancestor within a lineage of a known term for the protein based on the GO term hierarchy. Such annotations are implied by the hierarchy.

A comparison of the performance Krole versus Profunc and Proknow was also done. The ProKnow analysis the accuracy measurement of positive predictive value decreases as the depth of the hierarchy is increased [15, 66]. Here we report values that the positive predictive values remain consistently high as the depth of the hierarchy is increased, as shown in Fig. 2. For ProFunc, each prediction requires a scan to be executed across the slate of analysis programs which is computationally demanding [16]; and scans are not available for all the protein structural chains in the PDB. The current method uses

Table 3 The ranked scores for each used data source

Data type and source	Average number of known attributes per structure	Total number of times the attributes were used in the classification models	Ranking score (R)
Structural cluster identifier from jFatCat	16.2082	5,337	7,438.72
Ligand identifiers from ChEBI, DrugBank, ChEMBL, SMPDB	8.7118	5,387	7,247.65
Term identifiers from GO	4.9796	3,547	7,289.45
Ligand identifiers from BindingDB	1.6200	62	5,265.60
Disease names from SNPs3D	1.4215	1,312	6,049.10
Domain identifiers from CATH	0.7003	1,178	7,218.35
Domain identifiers from SCOP	0.6253	1,267	6,943.01
Ligand identifiers from BioCyc	0.5579	113	5,041.68
Accession codes from Pfam	0.5068	4,416	4,980.89
Pathway identifiers from BioCyc	0.4467	52	4,895.42
Pathway identifiers from NCI PID	0.4359	281	5,821.42
Enzyme codes from EC2PDB	0.4221	778	5,158.60
Disease identifiers from MIM	0.2588	341	4,899.80
Pathway identifiers from INOH	0.1175	15	4,856.66
Functional attributes from FEATURE	0.0649	599	5,716.62
Pathway identifiers from CellMap	0.0386	47	5,081.34
Ligand identifiers from HumanCyc	0.0176	3	4,260.35
Pathway identifiers from HumanCyc	0.0116	1	376.37
Biochemical reaction identifiers from HumanCyc	0.0056	0	0.00

The data sources are ordered by the average number of attributes per structure

precomputed values to build the prediction models, and predictions are available across all the chains.

Acknowledgments We thank Grace Tang in Russ Altman's group at Stanford for updating the models for the characterization of structural features related to protein function, as available in the FEATURE resource. Andreas Prlic kindly provided links to the lists of structures used as representatives for the pre-calculated structural comparisons by the jFATCAT computer program. Margeret Gabanyi, John Westbrook, and Helen Berman at Rutgers University provided assistance with access to data collections of the SBKB and direction on the analyses of structures in its functional sleuth section. Support was provided in part by an NIH grant [grant number 3U54GM074958-04S1].

References

1. Terwilliger TC, Stuart D, Yokoyama S (2009) Lessons from structural genomics. Annu Rev Biophys 38:371–383
2. Nair R et al (2009) Structural genomics is the largest contributor of novel structural leverage. J Struct Funct Genomics 10(2): 181–191
3. Marsden RL, Orengo CA (2008) Target selection for structural genomics: an overview. Methods Mol Biol 426:3–25
4. Dessailly BH et al (2009) PSI-2: structural genomics to cover protein domain family space. Structure 17(6):869–881
5. Berman HM et al (2000) The Protein Data Bank. Nucleic Acids Res 28(1):235–242
6. Kouranov A et al (2006) The RCSB PDB information portal for structural genomics. Nucleic Acids Res 34(Database issue): D302–D305
7. Xie L, Bourne PE (2005) Functional coverage of the human genome by existing structures, structural genomics targets, and homology models. PLoS Comput Biol 1(3):e31
8. Burley SK et al (2008) Contributions to the NIH-NIGMS Protein Structure Initiative from the PSI Production Centers. Structure 16(1):5–11
9. Pazos F, Sternberg MJ (2004) Automated prediction of protein function and detection of functional sites from structure. Proc Natl Acad Sci USA 101(41):14754–14759
10. Rost B et al (2003) Automatic prediction of protein function. Cell Mol Life Sci 60(12):2637–2650
11. Laskowski RA, Watson JD, Thornton JM (2005) Protein function prediction using local 3D templates. J Mol Biol 351(3): 614–626
12. Jaroszewski L et al (2009) Exploration of uncharted regions of the protein universe. PLoS Biol 7(9):e1000205
13. Ward RM et al (2009) Evolutionary Trace Annotation Server: automated enzyme function prediction in protein structures using 3D templates. Bioinformatics 25(11):1426–1427
14. Petrey D, Fischer M, Honig B (2009) Structural relationships among proteins with different global topologies and their implications for function annotation strategies. Proc Natl Acad Sci USA 106(41):17377–17382
15. Pal D, Eisenberg D (2005) Inference of protein function from protein structure. Structure 13(1):121–130

16. Laskowski RA, Watson JD, Thornton JM (2005) ProFunc: a server for predicting protein function from 3D structure. Nucleic Acids Res 33(Web Server issue):W89–W93
17. Ashburner M et al (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25(1): 25–29
18. Friedberg I (2006) Automated protein function prediction–the genomic challenge. Brief Bioinform 7(3):225–242
19. Lee D, Redfern O, Orengo C (2007) Predicting protein function from sequence and structure. Nat Rev Mol Cell Biol 8(12): 995–1005
20. Juncker AS et al (2009) Sequence-based feature prediction and annotation of proteins. Genome Biol 10(2):206
21. Marcotte EM et al (1999) A combined algorithm for genome-wide prediction of protein function. Nature 402(6757):83–86
22. Rentzsch R, Orengo CA (2009) Protein function prediction–the power of multiplicity. Trends Biotechnol 27(4):210–219
23. Berman HM et al (2008) The protein structure initiative structural genomics knowledgebase. Nucleic Acids Res 37(Database issue): D365–368
24. Schapire RE et al (1998) Boosting the margin: a new explanation for the effectiveness of voting methods. Ann Stat 26(5): 1651–1686
25. Favre B, Hakkani D Icsiboost. http://code.google.com/p/icsiboost/
26. Karp PD, Ouzounis CA, Moore-Kochlacs C, Goldovsky L, Kaipa P, Ahren D, Tsoka S, Darzentas N, Kunin V, Lopez-Bigas N (2005) Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. Nucleic Acids Res 33(19):6083–6089
27. Cerami EG, Gross BE, Demir E, Rodchenkov I, Babur O, Anwar N, Schultz N, Bader GD, Sander C Pathway Commons, a web resource for biological pathway data. Nucleic Acids Res 39(Database issue):D685–690
28. Fukuda K (2008) INOH pathway database: curation, annotation, integration. InterOntology08 1(1):47–50
29. Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, Buetow KH (2009) The pathway interaction database. Nucleic Acids Res 37:D674–D679
30. Liu T et al (2007) BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. Nucleic Acids Res 35(Database issue):D198–D201
31. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M (2008) DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res 36(Database issue):D901–D906
32. Degtyarenko K et al (2008) ChEBI: a database and ontology for chemical entities of biological interest. Nucleic Acids Res 36(Database issue):D344–D350
33. ChEMBL. Available from: http://www.ebi.ac.uk/chembldb/
34. Frolkis A et al (2009) SMPDB: the small molecule pathway database. Nucleic Acids Res **38**(Database issue):D480–D487
35. Yue P, Melamud E, Moult J (2006) SNPs3D: candidate gene and SNP selection for association studies. BMC Bioinform 7(1):166
36. Online Mendelian Inheritance in Man, OMIM (TM), McKusick-Nathans Institute of Genomic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine. Bethesda, MD
37. McKusick VA (2007) Mendelian Inheritance in Man and its online version, OMIM. Am J Hum Genet 80(4):588–604
38. Velankar S et al (2005) E-MSD: an integrated data resource for bioinformatics. Nucleic Acids Res 33(Database issue):D262-D265
39. Bairoch A (2000) The ENZYME database in 2000. Nucleic Acids Res 28(1):304–305
40. Orengo CA, Pearl FM, Bray JE, Todd AE, Martin AC, Lo Conte L, Thornton JM (1999) The CATH Database provides insights into protein structure/function relationships. Nucleic Acids Res 27(1):275–279
41. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247:536–540
42. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunesekaran P, Ceric G, Forslund K, Holm L, Sonnhammer EL, Eddy SR, Bateman A (2010) The Pfam protein families database. Nucleic Acids Res 38(1):D211–D222
43. Godzik YYA (2003) Flexible structure alignment by chaining aligned fragment pairs allowing twists. Bioinformatics 19(2): ii246–ii255
44. Prlic A et al (2010) Pre-calculated protein structure alignments at the RCSB PDB website. Bioinformatics 26(23):2983–2985
45. Halperin I, Glazer DS, Wu S, Altman RB (2008) The FEATURE framework for protein function annotation: modeling new functions, improving performance, and extending to novel applications. BMC Genomics 9(Suppl 2:S2)
46. Devos D, Valencia A (2000) Practical limits of function prediction. Proteins 41(1):98–107
47. Han J, Kamber M (2006) Data mining: concepts and techniques, 2nd ed. Morgan Kaufmann, Boston, xxviii
48. Levitt M, Chothia C (1976) Structural patterns in globular proteins. Nature 261(5561):552–558
49. Richardson JS (1981) The anatomy and taxonomy of protein structure. Adv Protein Chem 34:167–339
50. Bowie JU, Luthy R, Eisenberg D (1991) A method to identify protein sequences that fold into a known three-dimensional structure. Science 253(5016):164–170
51. Altschul SF et al (1990) Basic local alignment search tool. J Mol Biol 215(3):403–410
52. Rose PW et al (2010) The RCSB Protein Data Bank: redesigned web site and web services. Nucleic Acids Res 39(Database issue):D392–D401
53. Radauer C, Lackner P, Breiteneder H (2008) The Bet v 1 fold: an ancient, versatile scaffold for binding of large, hydrophobic ligands. BMC Evol Biol 8:286
54. Yarullina D, Ilinskaya O (2007) Genomic determinants of nitric oxide biosynthesis in <i>*Lactobacillus plantarum*</i>: potential opportunities and reality. Mol Biol 41(5):820–826
55. Aravind L, Anantharaman V (2003) HutC/FarR-like bacterial transcription factors of the GntR family contain a small molecule-binding domain of the chorismate lyase fold. FEMS Microbiol Lett 222(1):17–23
56. Levin I et al (2005) Crystal structure of an indigoidine synthase A (IndA)-like protein (TM1464) from *Thermotoga maritima* at 1.90 A resolution reveals a new fold. Proteins 59(4):864–868
57. Preumont A et al (2008) Molecular identification of pseudouridine-metabolizing enzymes. J Biol Chem 283(37):25238–25246
58. Takahashi H et al (2007) Cloning and characterization of a *Streptomyces* single module type non-ribosomal peptide synthetase catalyzing a blue pigment synthesis. J Biol Chem 282(12): 9073–9081
59. Preumont A et al (2010) HDHD1, which is often deleted in X-linked ichthyosis, encodes a pseudouridine-5′-phosphatase. Biochem J 431(2):237–244
60. Benach J et al (2003) The 2.3-A crystal structure of the shikimate 5-dehydrogenase orthologue YdiB from *Escherichia coli* suggests a novel catalytic environment for an NAD-dependent dehydrogenase. J Biol Chem 278(21):19176–19182
61. Singh S, Korolev S, Koroleva O, Zarembinski T, Collart F, Joachimiak A, Christendat D (2005) Crystal structure of a novel shikimate dehydrogenase from *Haemophilus influenzae*. J Biol Chem 17(280):17101–17108
62. Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28(1):27–30

63. Chang A et al (2009) BRENDA, AMENDA and FRENDA the enzyme information system: new content and tools in 2009. Nucleic Acids Res 37(Database issue):D588–D592
64. Denison DD (2003) Nonlinear estimation and classification. Lecture notes in statistics. Springer, New York, vii
65. Chruszcz M et al (2010) Unmet challenges of structural genomics. Curr Opin Struct Biol 20(5):587–597
66. Medrano-Soto A, Pal D, Eisenberg D (2008) Inferring molecular function: contributions from functional linkages. Trends Genet 24(12):587–590