PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Analysis of protein function and its prediction from amino acid sequence

Wyatt T. Clark and Predrag Radivojac*

School of Informatics and Computing, Indiana University, Bloomington, Indiana

ABSTRACT

Understanding protein function is one of the keys to understanding life at the molecular level. It is also important in the context of human disease because many conditions arise as a consequence of alterations of protein function. The recent availability of relatively inexpensive sequencing technology has resulted in thousands of complete or partially sequenced genomes with millions of functionally uncharacterized proteins. Such a large volume of data, combined with the lack of high-throughput experimental assays to functionally annotate proteins, attributes to the growing importance of automated function prediction. Here, we study proteins annotated by Gene Ontology (GO) terms and estimate the accuracy of functional transfer from protein sequence only. We find that the transfer of GO terms by pairwise sequence alignments is only moderately accurate, showing a surprisingly small influence of sequence identity (SID) in a broad range (30–100%). We developed and evaluated a new predictor of protein function, functional annotator (FANN), from amino acid sequence. The predictor exploits a multioutput neural network framework which is well suited to simultaneously modeling dependencies between functional terms. Experiments provide evidence that FANN-GO (predictor of GO terms; available from http://www.informatics.indiana.edu/predrag) outperforms standard methods such as transfer by global or local SID as well as GOtcha, a method that incorporates the structure of GO.

Proteins 2011; 79:2086–2096.

Key words: protein function; prediction; protein function transfer; gene ontology; neural network.

INTRODUCTION

Protein function is a multifaceted concept that reflects "everything that happens to or through a protein,"[1] and is typically considered from the biochemical, biological, and phenotypic perspectives.[2] From the biochemical or molecular perspective, a protein may be a kinase, whereas in terms of its biological function this kinase can be involved in numerous processes, such as cell cycle regulation or cell–cell signaling. Two proteins with the same molecular function may be involved in drastically different biological processes, and conversely, the set of proteins associated with a particular biological process will generally be drawn from a wide range of molecular functions. From the phenotypic viewpoint, a protein is generally associated with variation in observable physical or behavioral traits. For example, kinase variants or mutants may be responsible for disease. Adding another level of complexity to the study of function is the fact that a protein's molecular function, while generally considered to be a static notion, is modulated by a particular cellular context, for example, the presence of other molecules, or properties of the physical environment, for example, temperature.[3]

Several classification systems have been proposed to standardize functional annotation and to facilitate computation. With few exceptions, these classification systems usually take on the structure of hierarchical ontologies. Enzyme Commission (EC) numbers[4] and the Munich Information Center for Protein Sequences (MIPS) functional catalog[5] are two well-accepted schemes; however, the most commonly used functional classification is the Gene Ontology (GO). GO provides three hierarchical classifications as directed acyclic graphs: molecular function ontology (MFO), biological process ontology (BPO), and cellular component ontology.[6] With respect to defining a particular gene's phenotype, the existing classifications are species-specific (e.g., the Human Phenotype Ontology[7]) and predominantly constructed to address human disease. The Unified Medical Language System,[8] for example, incorporates a number of vocabularies together with semantic relationships between terms, mainly for the purpose of defining associations between genes and medical disorders.

Experimentally determined protein function annotations have been steadily accumulating over the years. Currently, there are about 50,000 proteins with at least one experimentally annotated GO term from MFO or BPO in the GO database and Swiss-Prot[9] combined. However, owing

Additional Supporting Informstion may be found in the sonline version of this article.
Grant sponsor: National Science Foundation; Grant number: DBI-0644017
*Correspondence to: Predrag Radivojac, School of Informatics and Computing, Indiana University, 150 South Woodlawn Avenue, Bloomington, IN 47405. E-mail: predrag@indiana.edu.
Received 19 October 2010; Revised 15 February 2011; Accepted 3 March 2011
Published online 21 March 2011 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/prot.23029

to the numerous sequencing projects,[10] the gap between annotated and nonannotated proteins has exceeded two orders of magnitude and will only get wider. In addition, annotated proteins may not be sufficiently covered, both with respect to the resolution of functional annotations and the inclusion of other yet unknown functions. Thus, it is important to develop algorithms capable of accurately predicting protein function.

Historically, sequence-based inference was the first strategy used to predict protein function, even if most studies at the time avoided explicitly relating homology and function.[11] Global and local sequence alignments were used to query sequence databases for similarities with a target protein. With the accumulation of experimentally determined protein functions, the most similar annotated sequences have traditionally been used to infer function.[1] More advanced methods exploited predicted physicochemical properties,[12,13] evolutionary relationships,[14–16] or the structure of functional ontologies to achieve different confidence levels at different ontological terms.[17–19] Microarrays, protein–protein interaction networks, protein structures, and predicted ligands have also been exploited.[20–29] However, most of these methods are limited to a few organisms where such data are available. One way or another, sequence alignment-based inference is the cornerstone of functional inference and is the focal point of this study.

Sequence alignment-based transfer of function has been thoroughly studied in the last decade, predominantly for enzymes.[1,30–34] The results of these studies indicate that at least 60% sequence identity (SID), and more likely closer to 80%, is required for the accurate transfer of the third level of EC classification. More sophisticated approaches were proposed as well: the GOtcha method was developed to take sequence alignment scores between a query protein and a functionally annotated database and overlay them on the functional ontology, cumulatively propagating such scores toward the root of the ontology.[17] PFP refined this technique by incorporating PSI-BLAST (Position-Specific Iterated–Basic Local Alignment Search Tool) alignments at very low significance levels and conditional probabilities that a protein is associated with pairs of functional terms.[18] Other methods such as ProtFun,[13] ConFunc,[35] GOsling,[36] and GOstruct[37] were developed for high-throughput prediction tasks. Finally, phylogenetic methods attempt to exploit particular evolutionary relationships within a gene family.[38,39] Methods such as SIFTER[15] or ortholog identification methods[40] belong to this category. Several recent reviews provide good perspectives on protein function prediction at all scales.[1,41–47]

This study addresses protein function at the molecular function and biological process levels, as defined by the GO consortium, and functional inference from protein sequence alone. We first analyze the distribution of GO terms associated with experimentally annotated proteins, and the relationship between their number and the estimated number of domains in a protein. We then consider the power of functional transfer of GO terms based on sequence similarity and finally present our method for the probabilistic inference of GO terms using supervised learning. Our algorithm, functional annotator (FANN), uses multioutput artificial neural networks. We show that in the GO annotation task, FANN-GO outperforms standard sequence alignment methods and GOtcha, especially for sequences with low similarity to any currently annotated sequence.

RESULTS

Multifunctional proteins

It is well-known that some proteins have the ability to carry out more than one function, or "moonlight."[48,49] Such abilities can be facilitated by either having multiple functional domains, a single domain which binds multiple partners, or by different behavior upon post-translational modifications or change in physiological conditions.[49,50] Here, we systematically analyze multifunctionality with respect to MFO and BPO in GO. As an approximation of distinct functions we only consider the number of experimentally determined leaf GO terms associated with each protein. A GO term *g* was included in the count if no other term associated with the protein had *g* as its more general term in the ontology. For example, if a protein is associated with the term "protein binding," the term "binding" is not counted as distinct function because it is a generalization of "protein binding;" however, the term "nucleic acid binding" would be counted because neither of the two terms is a generalization of the other.

Figure 1 shows the distribution of the number of leaf terms associated with each functionally annotated sequence for both ontologies. In total, 26,707 sequences were included in the MFO analysis, and 29,118 sequences were included in the BPO analysis (see Datasets section). Only experimental evidence, traceable author's statement, or curator's inference were considered (exclusion of traceable author's statement and curator's inference resulted in very similar distributions; data not shown). The plots show greater diversity in a protein's participation in a biological process than its ability to carry out distinct molecular functions. About 34% of the proteins experimentally annotated by MFO terms have more than one leaf term, with some proteins having as many as 14. Similarly, 56% of proteins have more than 1 BPO leaf term, with six proteins having 50 terms or more. A higher percentage of proteins with multiple BPO leaf terms is consistent with the expectation that biological processes are governed more by the context in which a protein is used and less by the physicochemical abilities of the protein. Interestingly, both the molecular function and biological process ontologies show a scale free-like decrease in the probabil-

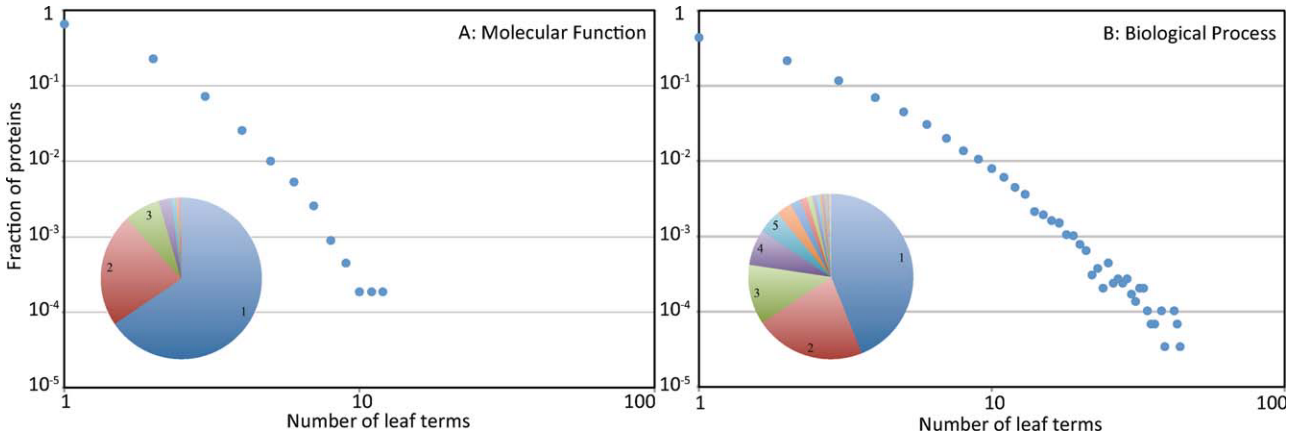

Figure 1

The distribution of the number of leaf terms in (**A**) molecular function and (**B**) biological process ontologies. The *x*-axis represents the number of leaf terms associated with a protein; the *y*-axis represents the fraction of proteins in the dataset with the given number of leaf terms. Both axes are in $\log_{10}$ scale. The inset in each panel provides a pie chart that corresponds to each plot.

ity that a protein is associated with an increasing number of functional terms. In the context of automated function annotation, these scale free-like distributions emphasize the need to develop prediction algorithms that can accommodate complex output patterns.

We also analyzed the relationship between the number of molecular functions and biological processes a protein is associated with. It seems intuitive to postulate that proteins with the ability to carry out multiple molecular functions should be more easily used in multiple different contexts, giving rise to an association with more biological processes. Out of 19,240 proteins in the intersection of datasets for molecular function and biological process, we found a Pearson's correlation coefficient of 0.261 between the numbers of associated leaf terms in the two ontologies. While this correlation may seem weak, we determined that this value is statistically significant by using a permutation test where the numbers of MFO and BPO terms were permuted in the dataset of proteins. We carried out 100,000 such permutations and did not find any cases in which the correlation coefficient was 0.261 or greater (the mean correlation coefficient for permuted data was 6.1×10^{-6} and standard deviation was 7.2×10^{-3}).

While, in general, a protein performing multiple molecular functions is associated with multiple biological processes, we conducted further analysis of proteins associated with a single term from one ontology and multiple terms from the other. For example, among the proteins that have only one MFO leaf term, but multiple BPO terms, we found that receptor binding terms such as "chemokine receptor binding" and "cytokine receptor binding" are the most enriched ($P < 1.0 \times 10^{-7}$; binomial test). On the other hand, there are also cases in which proteins that are associated with only one leaf term in BPO are related to multiple MFO terms. Such BPO terms are almost all related to metabolic processes, for example, "cellular metabolic process," "primary metabolic process" ($P < 1.0 \times 10^{-7}$; binomial test). When analyzing this class of proteins, we also found that some terms did not occur as often as expected. When looking at the class of proteins with one MFO term and multiple BPO terms, we found that "catalytic activity" was depleted. While sequences are annotated with this term 28% of the time in the whole dataset (29% in the set of proteins with a single MFO leaf term), it only occurs 17% of the time when we only consider sequences with exactly one MFO term but three or more BPO terms associated with them. Similarly, proteins having one BPO leaf term, but multiple MFO terms are usually depleted in terms involved in reproduction. We note that these data need to be interpreted with caution, because MFO and BPO terms are incomplete for most proteins and also because there may exist biases in ways current functions are acquired.

Finally, we looked at the relationship between the number of Pfam domains[51] found in a protein and the number of BPO and MFO terms it was associated with. We found that the Pearson's correlation between the number of MFO leaf terms and Pfam domains (detected using a threshold of 0.001) was larger than that of BPO leaf terms and Pfam domains (0.126 and 0.090 respectively). While, again, these correlation values might seem small, we found both were significant using the permutation test.

Transfer of function by sequence similarity

We evaluated the performance of simple function transfer between similar sequences. The following steps were taken: for each range of pairwise global SIDs, a target protein received all functional terms from each experimentally annotated protein within the given SID range. For each

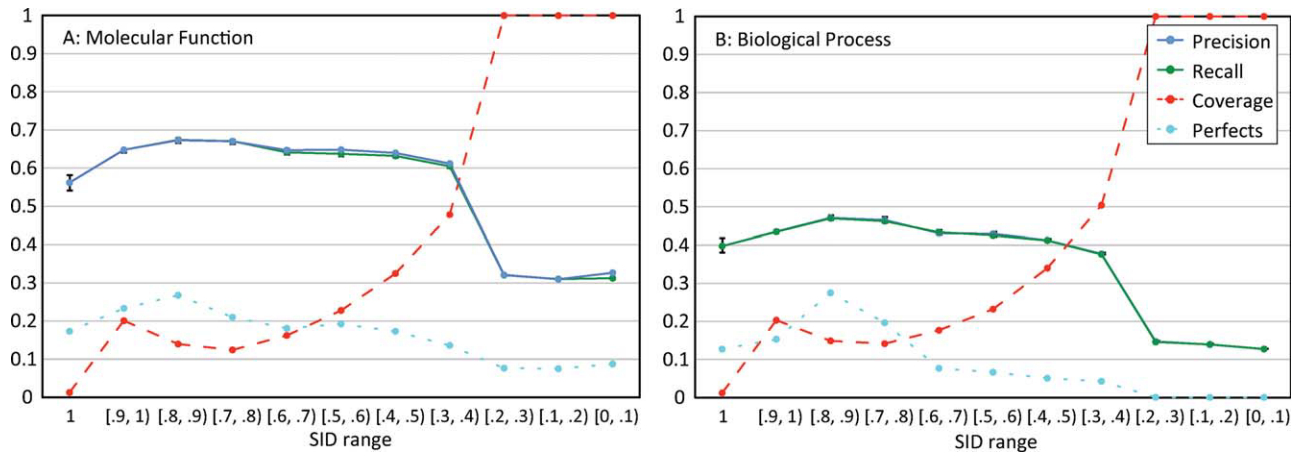

Figure 2

Accuracy of function transfer using global pairwise sequence identities (**A**: molecular function; **B**: biological process). For each sequence identity range (*x*-axis), the average precision (blue solid line) and recall (green solid line) of function transfer by pairwise similarity are shown. The teal dotted line represents the percentage of pairs with perfect annotations (e.g., both precision and recall equal to 1). The red dashed curve represents the percent of proteins that have pairwise matches (annotated with GO terms; experimental evidence code) in a given range. The error bars represent 95% confidence intervals. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

target sequence with more than one match in the given identity range, the precision and recall (see Model Selection and Evaluation Section) were calculated as averages over all pairs. Then, the precision and recall for the entire dataset are reported (Fig. 2) as averages of the averaged precisions and recalls calculated for each sequence covered (a leave-one-out estimation). We also report the coverage for each range of SID as the fraction of proteins with at least one other annotated sequence whose pairwise SID falls within the defined range. Finally, for all covered sequences we report the percentage of perfect annotations, that is, the percentage of all pairwise annotations in a given bin where both precision and recall were one.

The results shown in Figure 2 suggest that using global SID for the transfer of functional annotations is only moderately accurate (pairwise local alignments performed similarly; data not shown). Surprisingly, even for 100% identity transfer of function does not achieve either precision or recall of one (only 17% of identical sequence pairs had perfect transfer of MFO and 12% of BPO terms). This relatively low rate of perfect annotations among perfect matches, and similarly in the remaining identity bins, in our dataset is caused by three factors: (i) sparsity of database annotations, where proteins are incompletely annotated with respect to functionality and also specificity of annotation; (ii) database errors, caused by incorrect interpretation of experiments or by curation errors,[52,53] and (iii) organismal context, where the difference between two organisms influences a particular functional role of individual proteins, even at 100% SID.

In general, transferring MFO annotations is more accurate than transferring BPO annotations, which has previously also been observed by Rogers and Ben-Hur[54] in a different prediction scenario. This is probably a result of the fact that MFO terms are less dependent on cellular, tissue, or organismal context, but also that the topological properties, including the average branching factor, the number of terms, and the average depth of a leaf node, between the two ontologies differ (data not shown). For both ontologies, the precision of transferring predictions rapidly decreases once the 20–30% identity is reached. This range of SID has been termed the "twilight zone" for the inference of protein structure from sequence.[55] In the context of function transfer such a twilight zone cannot be clearly defined (or rather should be extended to the entire range 30–100%), with the range below 30% being one where function transfer breaks down completely, "midnight zone". We also evaluated an alternative approach where functional terms from all matches in a given identity range were transferred to the target protein and found a very similar level of precision but an increased recall (Supporting Information Fig. S1).

The quality of function transfer is highly dependent on the particular class of protein. We show this by splitting the MFO/BPO datasets into enzymes (proteins annotated with term catalytic activity) and non-enzymes. As seen in Figure 3, the ability to transfer molecular functions to these two classes of proteins was noticeably different and most likely points to a higher quality of functional annotations for enzymes.

Another interesting trend in the data is the fact that the quality of annotations transferred between sequences from the same species is higher than that obtained between different species. This can be seen by comparing the precision/recall curves in Figure 4 obtained by only considering pairs of sequences from the same species

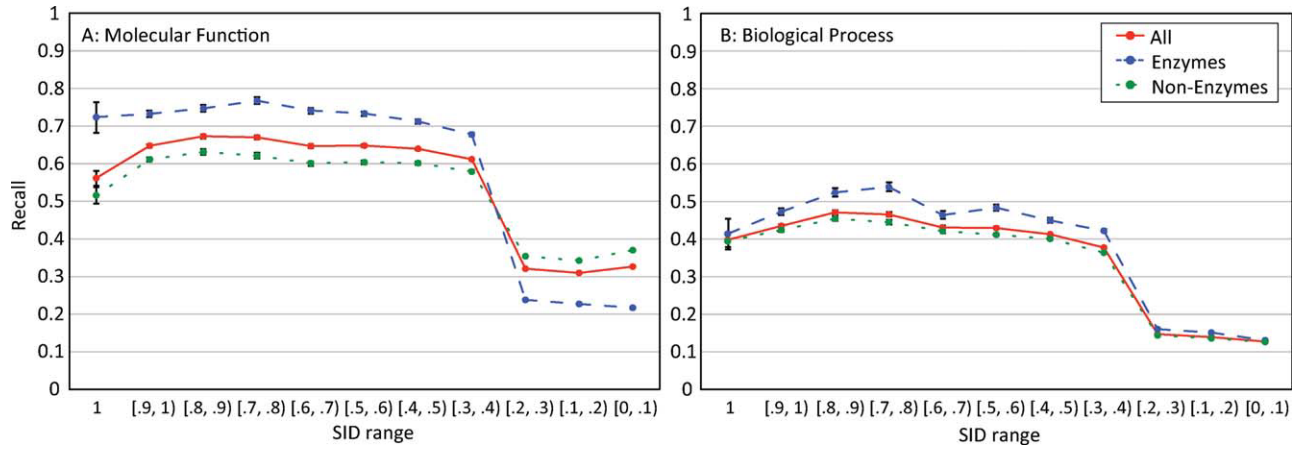

Figure 3

Accuracy of function transfer using global pairwise sequence identities for enzymes versus non-enzymes (**A**: molecular function; **B**: biological process). For each sequence identity range (*x*-axis), the average precision of function transfer by pairwise similarity is shown (blue dashed line: enzymes, green dotted line: non-enzymes; red solid line: combined). The error bars represent 95% confidence intervals. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

("within" curve, blue dashed line), and only considering pairs of sequences from different organisms ("between" curve, green dotted line). A similar trend has been previously observed on protein–protein interaction data by Mika and Rost.[56]

Quality of nonexperimental annotations in Swiss-Prot

Protein databases such as GO or Swiss-Prot contain a number of functional annotations supported by nonexperimental evidence codes. We aimed to assess the quality of such annotations by analyzing nonexperimental annotations for the proteins in Swiss-Prot (v10.0–v15.0) that in a later release (v15.15) accumulated experimental annotations. Figure 5 shows the quality of annotations by nontraceable author statement (NAS) or inferred from electronic annotation (IEA) evidence codes (the remaining nonexperimental codes did not contain enough sequences). Using the current annotations (v15.15) as true function, the precision and recall of each protein's annotation were calculated (Fig. 5).

As shown in Figure 5, the quality of IEA is consistently better than that of NASs for MFO, while the trend is

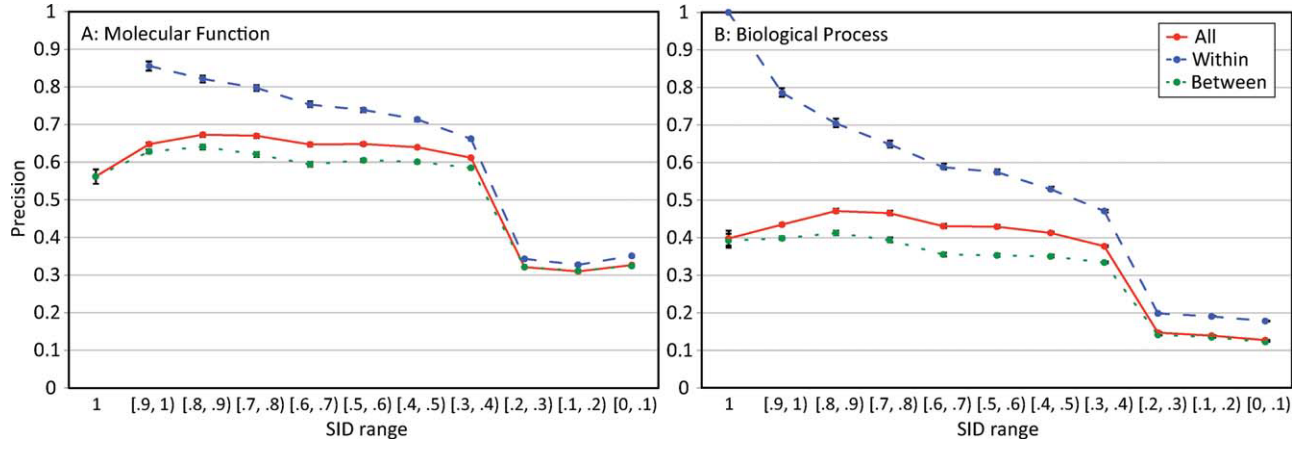

Figure 4

Accuracy of function transfer using global pairwise sequence identities from the same or different organism (**A**: molecular function; **B**: biological process). For each sequence identity range (*x*-axis), the average precision of function transfer by pairwise similarity is shown (blue dashed line: same organism, green dotted line: different organism; red solid line: combined). The error bars represent 95% confidence intervals. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

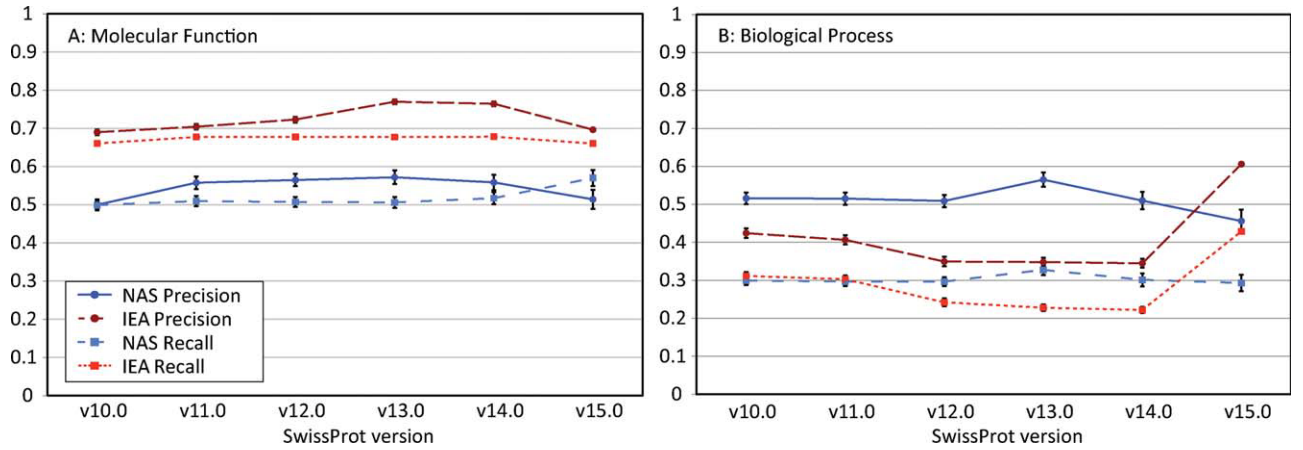

Figure 5

Accuracy of function transfer for the Swiss-Prot database (**A**: molecular function; **B**: biological process). *x*-axis represents a different version of Swiss-Prot. The dark dashed red line represents the precision of IEA annotations, and the dark blue solid line represents the precision of NAS annotations. The light red dotted line represents the recall of IEA annotations, and the light blue dashed line represents the recall of NAS annotations. The error bars represent 95% confidence intervals. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

reversed for BPO. Interestingly, the precision of the IEAs in the Swiss-Prot database for MFO exceeds the level achieved by a simple transfer of annotation (Fig. 2), while the NAS evidence suggest lower confidence levels compared to that of sequence transfer. A historical analysis of BPO annotations suggests that neither IEA nor NASs have been at the level of simple transfer of annotation; however, the latest major release of Swiss-Prot (v15.0) provides more accurate electronic inference than transfer by sequence similarity.

Performance of Functional Annotator

In this work, we present a simple classification method, FANN-GO, for predicting GO terms from MFO and BPO. The model is based on aligning a target sequence to a database of experimentally annotated proteins and calculating the *i*-score[17] that the protein is associated with each functional term. These scores were then used as inputs to an ensemble of multioutput neural networks that were trained to predict the probability that the protein is associated with each function. A variant of FANN-GO, referred to as FANN-GO^{species}, is also presented. This model is based on two groups of inputs, one containing inputs where a target protein is aligned only to proteins from the same species and another where the protein is aligned to proteins from different species.

We compared FANN-GO models with GOtcha as well as transfer by global and local SID. We also implemented a naïve method that uses prior probabilities of functional terms as prediction scores for all target proteins. That is, the score for each function, for all proteins, is simply the relative frequency of that term occurring in the (training) dataset. Performance of all classification models was assessed using 10-fold cross-validation on all functional terms associated with 50 proteins or more (mainly for the purpose of stable accuracy estimation). In doing so, we reduced the number of terms associated with the sequences in the dataset for molecular function to 344 and biological process to 1788. For each ontology, all methods were evaluated using the same datasets.

Figure 6 shows the performance of our two different supervised methods, FANN-GO (blue line) and FANN-GO^{species} (green line) compared to an in-house implementation of the GOtcha method (red line), and two methods using transfer of annotations based on pairwise SID: global-SID (purple line) based on global sequence alignments and local-SID (teal line) based on local BLAST alignments. These methods were benchmarked against a naïve classifier (orange line).

As shown in Figure 6A, FANN-GO outperformed GOtcha in both precision and recall when annotating sequences with MFO terms. Both methods outperformed simple functional transfer based on SID and naïve predictions. Interestingly, annotating sequences with the most probable functional term "binding" causes the naïve model to outperform predictors based on sequence similarity. Only after sequences are annotated with the third most probable function, "catalytic activity," does the naïve curve cross the global-SID curve at the precision and recall obtained at approximately 21% SID. The unusual shape of the naïve curve is indicative of the highly nonuniform nature of the relative frequencies of terms in MFO. With regard to the top three most probable terms in the ontology, 74% of sequences are annotated with the term "binding," 63% are annotated with the term "protein binding," and 28% are annotated with the term "catalytic activity."

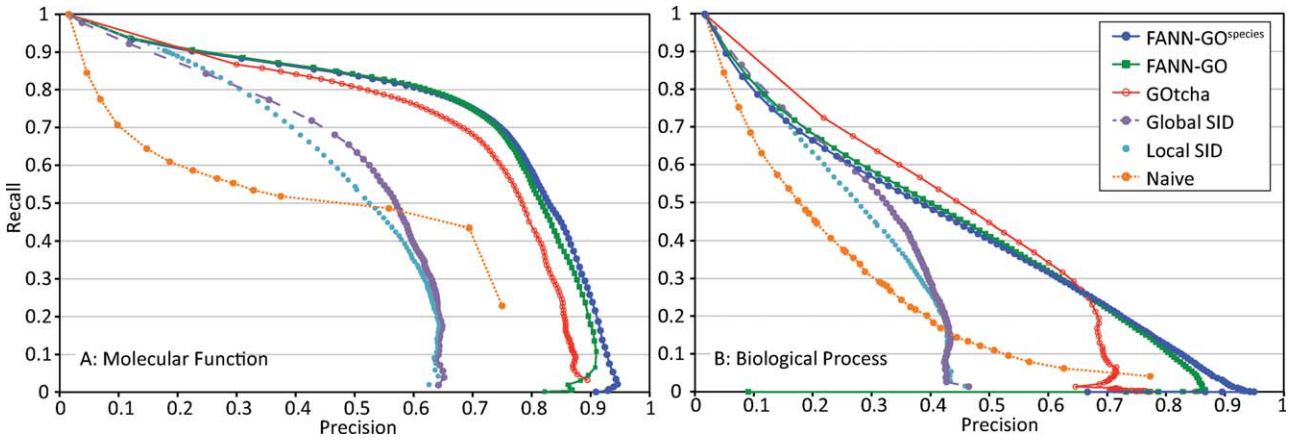

Figure 6

Precision–recall curves for several GO-term prediction algorithms (**A**: molecular function; **B**: biological process). The curves were generated by shifting the decision threshold t from 0 to 1 and considering all terms with scores $>t$ as predicted functions. The decision threshold for the sequence alignment curves (Global-SID, Local-SID) was the pairwise sequence identity (SID). The naïve curve (orange dotted) was generated by assigning all proteins the most common functional term, then the second most common, etc. FANN-GOspecies is represented by the solid blue line. FANN-GO is represented by the solid green line with square markers. GOtcha is represented by the solid red line with red circular markers. Global-SID is represented by the purple dashed line. Finally, Local-SID is represented by small teal markers only.

Figure 6B shows the precision–recall curves for BPO. Here, FANN-GO outperforms GOtcha in the high-precision/low-recall part of the curve, while the trend reverses for FANN-GO thresholds below 0.49 and GOtcha *i*-scores below 0.17 (we note that FANN-GO thresholds of <0.5 do not represent confident predictions). The naïve predictor again outperforms the SID-based transfer of function for the eight most probable terms in BPO. It even outperforms GOtcha when annotating sequences with the single most probable term, "cellular process." We found that the performance advantage of FANN-GO and FANN-GO^{species} over GOtcha was similar regardless of whether a protein was annotated with one or multiple leaf terms (data not shown).

To assess whether prediction was more accurate for proteins more similar to sequences in the experimentally annotated database, we also performed accuracy estimation in which test proteins were split into two sets. The first set contained proteins which shared at least 50% SID with any of the training sequences, and the second contained proteins which did not. When comparing FANN-GO with Gotcha, we found that the difference in performance between these two methods was even greater for the $<50\%$ SID set (Supporting Information Figure S2). Therefore, FANN-GO may be well-suited in annotating proteins that do not share significant SID with sequences that are experimentally annotated.

DISCUSSION

In this work, we investigated the distribution of GO functional annotations for MFO and BPO, the accuracy of their prediction by simple methods such as transfer by sequence similarity and then developed and evaluated a new computational method for predicting protein function, FANN.

With the growing gap between available sequences and experimentally annotated proteins, it is clear that functional annotation of all proteins can only be accomplished by combining experimental and computational methods. Targeted wet lab experiments have been predominantly focused on model organisms with an expectation that this will provide a detailed understanding of these organisms and that the gap between species can be accurately filled by computational methods. Indeed, model organisms provide a large fraction of the genes with experimentally verified functional annotations. In Swiss-Prot v15.15, we found that approximately 90% of annotated proteins in MFO and BPO belong to nine model organisms only (*H. sapiens*, *S. cerevisiae*, *M. musculus*, *R. norvegicus*, *A. thaliana*, *D. melanogaster*, *S. pombe*, *E. coli K-12*, and *C. elegans*). However, nearly 60% of the proteins from these model organisms still do not have any experimentally determined MFO or BPO terms. Thus, the development and assessment of computational methods is critical not only for filling the gap between model and non-model organisms but also for completing the annotation of model organisms and driving experimental analyses.

A frequent interpretation of the sequence–structure–function paradigm is that a protein must adopt a single structure (minimum energy state, kinetically reachable) to be functional, with such conformation usually called the native state. However, such an understanding has been challenged from both structural and functional perspectives. Many proteins have been characterized as

intrinsically disordered. In such proteins, no single structure is seen as being dominant (i.e., high probability conformation with deep energy minimum) and a presence of conformational ensembles (i.e., macrostates[57]) is probably even required for function.[58–60] At the same time, it is now recognized that multifunctional proteins are also common.[49] We find that at least 34% of functionally characterized proteins (by experimental studies) are already assigned more than one distinct molecular function term and that at least 56% of proteins participate in more than one distinct biological process. We believe that the ability of a protein to be multifunctional in terms of its biochemical function needs to be achieved by developing new structural conformations and physicochemical interfaces (including the addition of new domains), whereas its involvement in multiple biological processes does not.

We also analyzed the quality of MFO and BPO term "transfer" by simple sequence similarity and found that inference by similarity shows flat accuracy in the entire range from 30% to 100% of pairwise SID (unless within the same organism). This leads to the conclusion that more sophisticated computational methods are necessary. To date, much attention has specifically been paid to the quality of function transfer for enzymes.[1,34] Here, we extended such analyses to nonenzymatic proteins and observed that transfer of function to nonenzymatic proteins is less accurate than that achieved for proteins annotated with any function from the catalytic activity portion of the ontology. While the underlying reasons may simply lie in the sparseness of these parts of the ontology, annotating a protein with functions without knowledge of the associated mechanics (information that is often known for enzymes) can result in less accurate assignment of proteins with such terms. This may also be true for other classes of terms in the ontology. For example, terms which group together proteins that carry out similar tasks in the cell, but do so through different molecular mechanisms, will be less likely to be defined in terms of sequence similarity among member proteins. Conversely, terms which define a function carried out by a specific mechanism (e.g., zinc finger binding) will be more likely to be inferred by sequence similarity.

Our new algorithm, FANN was developed as a supervised method for the prediction of protein function based on sequence alignments only. While diverse types of data may be available in some model organisms, function transfer between model and non-model organisms is ultimately and critically dependent on sequence-based predictions. Our attempt is a simple algorithm that encodes every protein as a vector of similarities in a space of protein functions, with neural networks used to learn posterior probabilities of sequence–function relationships. Neural networks are well suited for this classification task due to their natural ability to learn multiple outputs that are dependent and mutually nonexclusive. What methodology is best suited for such prediction tasks is an open problem, with examples including one-versus-all training of binary classifiers or their combination with postprocessing methods for multilabel classification. Compared to previous attempts to incorporate the structure of the ontology in a supervised manner, our method does not require postprocessing[19] or constraints on the kernel functions between input and output spaces.[37] On the other hand, neural networks, as implemented here, do not necessarily produce a consistent set of predictions, that is, a parent term may have a lower score than any of its children's terms. Although from a machine learning perspective this is undesired, we believe this feature can be useful for detecting problems with the ontology itself, as it should not be assumed that either MFO or BPO are error free and complete. It should be pointed out that this inconsistency is also a drawback of one-versus-all methods.

Currently, it is believed that the most accurate computational models for sequence-based function prediction are GOtcha-based unsupervised algorithms.[47,61,62] While these models are very good in their performance, we show that supervised methods outperform them on the same data. One drawback, however, is that supervised methods are limited to only functional terms that are associated with a sufficient number of sequences. Therefore, a function prediction task may be well suited for the methods combining supervised (for terms where enough proteins are available) and unsupervised (when only a small number of sequences are available) approaches. Finally, it should be kept in mind that a separate category of *de novo* prediction methods is needed for proteins with novel or organism-specific functions. Such algorithms, however, are beyond the scope of this work.

METHODS

Problem formulation

We consider a set of protein sequences $S = \{s_1, s_2, \ldots, s_m\}$, a set of protein molecular function terms $F = \{f_1, f_2, \ldots, f_{|F|}\}$ and a set of biological process terms $P = \{p_1, p_2, \ldots, p_{|P|}\}$, where each sequence in S is annotated by at least one element from F or P. Additionally, each function $f \in F$ and process $p \in P$ is associated with at least one sequence from S. Our objective is to construct a classification model which, given an amino acid sequence $s \in S$, assigns posterior probability that the sequence has the ability to carry out each particular function from F or is involved in each particular biological process in P. Similarly, given a functional term f, or a biological process p, our objective is to find the most likely sequences associated with that function or process.

Datasets

We used the Swiss-Prot database[9] from May 2010 (v.15.15). A dataset D_{MFO} of 26,707 protein sequences

Table I
The Number of Proteins in Each Dataset with the Breakdown According to Species

	MFO	BPO
H. sapiens	7253	6515
S. cerevisiae	4062	4071
M. musculus	3945	4672
R. norvegicus	2624	2696
A. thaliana	2042	2813
D. melanogaster	1627	1591
E. coli K-12	1498	1009
S. pombe	1079	1935
C. elegans	569	1599
All other	2008	2217
Total	26,707	29,118

was generated by selecting sequences with molecular functions that were supported by at least one of the following experimental evidence codes: EXP, IDA, IPI, IMP, IGI, IEP, TAS, IC. The 26,707 sequences in D_{MFO} consisted of a total of 4276 molecular function terms associated with them. Similarly, a separate set D_{BPO} consisting of 29,118 sequences with a total of 11,300 associated biological process terms was generated using the same criterion. We note that $|D_{\mathrm{MFO}} \cap D_{\mathrm{BPO}}| = 19{,}240$ and $|D_{\mathrm{MFO}} \cup D_{\mathrm{BPO}}| = 36{,}585$.

To arrive at the final datasets, 1429 proteins with known function in Swiss-Prot v15.15 were removed. Such sequences were either shorter than 50 amino acids or were associated with the same gene name as some other protein in the same organisms. Datasets are summarized in Table I.

Data representation

Sequence alignments were used to represent each protein sequence as a fixed-length vector in a feature space. Each dimension in the feature space was selected to correspond to one term from a set of available functional terms F (or P for the BPO). Although we tested several ways of encoding alignment data into features, we found that using i-scores, as proposed by the GOtcha algorithm,[17] as features worked the best. For the completeness of this work, we briefly summarize this representation.

First, let $e(s, s_i)$ be the E-value obtained by aligning target sequence s to the ith sequence in database of proteins with experimentally determined functions, that is, $s_i \in S$. The r-score for functional term f is then generated as

$$r_f(s) = -\sum_{s_i \in S_f} \log(e(s, s_i)) + c$$

where S_f is a subset of S containing all proteins with functional term f and c is a constant value added to the sum to ensure non-negative r scores (here we used $c = 2$ and e-value threshold of 10). The i-scores for each function were then calculated by normalizing r_f by the r score of the root node in a given ontology (term $f_{\mathrm{root}} \in F$) as

$$i_f = r_f / r_{\mathrm{root}}$$

Finally, a feature vector is obtained by concatenating i-scores for each of the $|F|$ functional terms. We note that the i-score for the root term always equals 1; thus, it was excluded from the vector representation. For the MFO, the i-score vector representation consists of $|F|$ dimensions, while for the BPO, the i-score representation consists of $|P|$ dimensions. These features were used for the basic version of FANN-GO.

To take advantage of the fact that the quality of transferring functions from sequences within the same species was higher than that achieved when transferring functions only from sequences in different species, we made two additional sets of i-score features, one based on the i-scores using the proteins from the same species and another using the i-scores from the proteins from different species only. These features were used for the version of the predictor referred to FANN-GO$^{\mathrm{species}}$.

Classification models

To address the multilabel classification problem of protein function, we used a multioutput feed-forward neural network framework. Multioutput networks have the ability to simultaneously learn multiple dependent target variables, a property that is well suited to the problem of predicting mutually nonexclusive terms of protein function ontologies.

Before training a multioutput neural network, we performed several data preprocessing steps. All features were first normalized using the z-score method. Feature selection filtering was then performed by using the t-test. Finally, principal component analysis was performed to combine highly correlated features (retained variance = 99%).

Owing to the high-memory requirement of a multioutput neural network, it was not practically possible to train a model with 1000 outputs or more, on a dataset of size $|D_{\mathrm{MFO}}|$ or $|D_{\mathrm{BPO}}|$. To overcome this limitation, we created ensembles of 100 networks such that in each network only 100 randomly selected outputs were considered. Prediction values for a test sequence for a particular function were finally calculated as an average over the output scores generated from the subset of networks that included the given term in their output layers. All neural networks had 100 hidden neurons, using the resilient propagation algorithm[63] in training (with at most 1000 epochs). The networks were implemented using Matlab.

Model selection and evaluation

The accuracy of the model was estimated using 10-fold cross-validation. Parameter selection for each neural network was performed on a separate validation set, such that only the best performing parameter set was used on

the test partition. Furthermore, separate sets of features were generated for each fold using only alignments with sequences in the training portion of the data. An individual BLAST database was built for each fold's set of training sequences to ensure that alignment *e*-values were not influenced by sequences in the test portion of the data. All methods were evaluated by plotting precision–recall curves. FANN-GO was evaluated against three different strategies. The global-SID and local-SID strategies represent transfer by sequence similarity in which each prediction was generated by transferring functional terms directly from sequences with SID to the query sequence greater than the threshold. In addition, the performance of FANN was compared to the GOtcha classifier[17] as well as the naïve classifier. The naïve classifier predicts the terms according to their descending prior probabilities in the training data, that is, the term occurring in 75% of training sequences will be predicted with score 0.75 for all target proteins.

Precision–recall curves were generated as follows. For each query sequence, a set of predictions over all $|F|$ functions was generated. A decision threshold t value above which all predictions were taken was incrementally reduced from 1 to 0, in steps of 0.01. Terms with prediction scores above a particular threshold t_i were selected, and each term was propagated towards the root of the ontology. This resulted in a set of predicted terms T_p. The precision (pr) and recall (rc) between the predicted terms T_p and true terms T_t associated with sequence s were then calculated as $\text{pr} = |T_p \cup T_t| \backslash |T_p|$ and $\text{rc} = |T_p \cup T_t| \backslash |T_t|$. The final precision and recall were averaged over all test sequences to create a point in the precision–recall space.

ACKNOWLEDGMENTS

We thank Prof. Matthew W. Hahn and Fuxiao Xin for helpful comments and proofreading this article. We also thank the reviewers for constructive comments that further improved the quality and presentation of this work.

REFERENCES

1. Rost B, Liu J, Nair R, Wrzeszczynski KO, Ofran Y. Automatic prediction of protein function. Cell Mol Life Sci 2003;60:2637–2650.
2. Bartlett G, Todd A, Thornton J. Infering function from structure. In: Bourne PE, Weissig H, editors. Structural bioinformatics: Hoboken, NJ: Wiley-Liss; 2003.
3. Mohan A, Uversky VN, Radivojac P. Influence of sequence changes and environment on intrinsically disordered proteins. PLoS Comput Biol 2009;5:e1000497.
4. (NC-IUBMB) NCotIUoBaMB. Enzyme nomenclature. New York: Academic Press; 1992.
5. Ruepp A, Zollner A, Maier D, Albermann K, Hani J, Mokrejs M, Tetko I, Guldener U, Mannhaupt G, Munsterkotter M, Mewes HW. The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. Nucleic Acids Res 2004;32:5539–5545.
6. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000;25:25–29.
7. Robinson PN, Mundlos S. The human phenotype ontology. Clin Genet 2010;77:525–534.
8. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004;32 (Database issue):D267–D270.
9. Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS. The Universal Protein Resource (UniProt). Nucleic Acids Res 2005;33 (Database Issue):D154–D159.
10. Liolios K, Mavromatis K, Tavernarakis N, Kyrpides NC. The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res 2008;36 (Database issue):D475–D479.
11. Doolittle RF. Of URFS and ORFS: a primer on how to analyze derived amino acid sequences: Mill Valley, CA: University Science Books; 1986.
12. Jensen LJ, Gupta R, Blom N, Devos D, Tamames J, Kesmir C, Nielsen H, Staerfeldt HH, Rapacki K, Workman C, Andersen CA, Knudsen S, Krogh A, Valencia A, Brunak S. Prediction of human protein function from post-translational modifications and localization features. J Mol Biol 2002;319:1257–1265.
13. Jensen LJ, Gupta R, Staerfeldt HH, Brunak S. Prediction of human protein function according to gene ontology categories. Bioinformatics 2003;19:635–642.
14. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci USA 1999;96:4285–4288.
15. Engelhardt BE, Jordan MI, Muratore KE, Brenner SE. Protein molecular function prediction by Bayesian phylogenomics. PLoS Comput Biol 2005;1:e45.
16. Bandyopadhyay D, Huan J, Liu J, Prins J, Snoeyink J, Wang W, Tropsha A. Structure-based function inference using protein family-specific fingerprints. Protein Sci 2006;15:1537–1543.
17. Martin DM, Berriman M, Barton GJ. GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes. BMC Bioinformatics 2004;5:178.
18. Hawkins T, Luban S, Kihara D. Enhanced automated function prediction using distantly related sequences and contextual association by PFP. Protein Sci 2006;15:1550–1556.
19. Barutcuoglu Z, Schapire RE, Troyanskaya OG. Hierarchical multi-label prediction of gene function. Bioinformatics 2006;22:830–836.
20. Vazquez A, Flammini A, Maritan A, Vespignani A. Global protein function prediction from protein–protein interaction networks. Nat Biotechnol 2003;21:697–700.
21. Deng M, Zhang K, Mehta S, Chen T, Sun F. Prediction of protein function using protein–protein interaction data. J Comput Biol 2003;10:947–960.
22. Letovsky S, Kasif S. Predicting protein function from protein/protein interaction data: a probabilistic approach. Bioinformatics 2003;19 (Suppl 1):i197–i204.
23. Troyanskaya OG, Dolinski K, Owen AB, Altman RB, Botstein D. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*). Proc Natl Acad Sci USA 2003;100:8348–8353.
24. Lee I, Date SV, Adai AT, Marcotte EM. A probabilistic functional network of yeast genes. Science 2004;306:1555–1558.
25. Pazos F, Sternberg MJ. Automated prediction of protein function and detection of functional sites from structure. Proc Natl Acad Sci USA 2004;101:14754–14759.
26. Pal D, Eisenberg D. Inference of protein function from protein structure. Structure 2005;13:121–130.
27. Nabieva E, Jim K, Agarwal A, Chazelle B, Singh M. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. Bioinformatics 2005;21 (Suppl 1):i302–i310.

28. Hermann JC, Marti-Arbona R, Fedorov AA, Fedorov E, Almo SC, Shoichet BK, Raushel FM. Structure-based activity prediction for an enzyme of unknown function. Nature 2007;448:775–779.
29. Song L, Kalyanaraman C, Fedorov AA, Fedorov EV, Glasner ME, Brown S, Imker HJ, Babbitt PC, Almo SC, Jacobson MP, Gerlt JA. Prediction and assignment of function for a divergent N-succinyl amino acid racemase. Nat Chem Biol 2007;3:486–491.
30. Devos D, Valencia A. Practical limits of function prediction. Proteins 2000;41:98–107.
31. Wilson CA, Kreychman J, Gerstein M. Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. J Mol Biol 2000;297:233–249.
32. Todd AE, Orengo CA, Thornton JM. Evolution of function in protein superfamilies, from a structural perspective. J Mol Biol 2001; 307:1113–1143.
33. Tian W, Skolnick J. How well is enzyme function conserved as a function of pairwise sequence identity? J Mol Biol 2003;333:863–882.
34. Addou S, Rentzsch R, Lee D, Orengo CA. Domain-based and family-specific sequence identity thresholds increase the levels of reliable protein function transfer. J Mol Biol 2009;387:416–430.
35. Wass MN, Sternberg MJ. ConFunc—functional annotation in the twilight zone. Bioinformatics 2008;24:798–806.
36. Jones CE, Schwerdt J, Bretag TA, Baumann U, Brown AL. GOSLING: a rule-based protein annotator using BLAST and GO. Bioinformatics 2008;24:2628–2629.
37. Sokolov A, Ben-Hur A. Hierarchical classification of gene ontology terms using the GOstruct method. J Bioinform Comput Biol 2010; 8:357–376.
38. Eisen JA. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. Genome Res 1998;8: 163–167.
39. Brown D, Sjolander K. Functional classification using phylogenomic inference. PLoS Comput Biol 2006;2:e77.
40. Remm M, Storm CE, Sonnhammer EL. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J Mol Biol 2001;314:1041–1052.
41. Friedberg I. Automated protein function prediction—the genomic challenge. Brief Bioinform 2006;7:225–242.
42. Lee D, Redfern O, Orengo C. Predicting protein function from sequence and structure. Nat Rev Mol Cell Biol 2007;8:995–1005.
43. Punta M, Ofran Y. The rough guide to in silico function prediction, or how to use sequence and structure information to predict protein function. PLoS Comput Biol 2008;4:e1000160.
44. Laskowski RA, Thornton JM. Understanding the molecular machinery of genetics through 3D structures. Nat Rev Genet 2008;9:141–151.
45. Kann MG. Protein interactions and disease: computational approaches to uncover the etiology of diseases. Brief Bioinform 2007;8:333–346.
46. Dalkilic MM, Costello JC, Clark WT, Radivojac P. From protein–disease associations to disease informatics. Front Biosci 2008;13: 3391–3407.
47. Rentzsch R, Orengo CA. Protein function prediction—the power of multiplicity. Trends Biotechnol 2009;27:210–219.
48. Jeffery CJ. Moonlighting proteins. Trends Biochem Sci 1999;24:8–11.
49. Jeffery CJ. Moonlighting proteins—an update. Mol Biosyst 2009;5: 345–350.
50. Jeffery CJ. Molecular mechanisms for multitasking: recent crystal structures of moonlighting proteins. Curr Opin Struct Biol 2004;14: 663–668.
51. Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz HR, Ceric G, Forslund K, Eddy SR, Sonnhammer EL, Bateman A. The Pfam protein families database. Nucleic Acids Res 2008;36 (Database issue):D281–D288.
52. Brenner SE. Errors in genome annotation. Trends Genet 1999;15: 132–133.
53. Schnoes AM, Brown SD, Dodevski I, Babbitt PC. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. PLoS Comput Biol 2009;5:e1000605.
54. Rogers MF, Ben-Hur A. The use of gene ontology evidence codes in preventing classifier assessment bias. Bioinformatics 2009;25:1173–1177.
55. Rost B. Twilight zone of protein sequence alignments. Protein Eng 1999;12:85–94.
56. Mika S, Rost B. Protein–protein interactions more conserved within species than across species. PLoS Comput Biol 2006;2:e79.
57. Dill KA. Polymer principles and protein folding. Protein Sci 1999;8: 1166–1180.
58. Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradovic Z. Intrinsic disorder and protein function. Biochemistry 2002;41:6573–6582.
59. Dyson HJ, Wright PE. Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol 2005;6:197–208.
60. Radivojac P, Iakoucheva LM, Oldfield CJ, Obradovic Z, Uversky VN, Dunker AK. Intrinsic disorder and functional proteomics. Biophys J 2007;92:1439–1456.
61. Friedberg I, Jambon M, Godzik A. New avenues in protein function prediction. Protein Sci 2006;15:1527–1529.
62. Izarzugaza JM, Grana O, Tress ML, Valencia A, Clarke ND. Assessment of intramolecular contact predictions for CASP7. Proteins 2007;69 (Suppl 8):152–158.
63. Riedmiller M, Braun H. A direct adaptive method for faster backpropagation learning: the RPROP algorithm. Proc IEEE Int Conf Neural Netw 1993;1:586–591.