Journal of Bioinformatics and Computational Biology
Vol. 5, No. 1 (2007) 1–30

Imperial College Press
www.icpress.co.uk

FUNCTION PREDICTION OF UNCHARACTERIZED PROTEINS

TROY HAWKINS
Department of Biological Sciences, Purdue University
West Lafayette, IN, USA
thawkins@purdue.edu

DAISUKE KIHARA
Department of Biological Sciences
Department of Computer Science, Purdue University
West Lafayette, IN, USA
dkihara@purdue.edu

Received 20 June 2006
Revised 23 September 2006
Accepted 10 October 2006

Function prediction of uncharacterized protein sequences generated by genome projects has emerged as an important focus for computational biology. We have categorized several approaches beyond traditional sequence similarity that utilize the overwhelmingly large amounts of available data for computational function prediction, including structure-, association (genomic context)-, interaction (cellular context)-, process (metabolic context)-, and proteomics-experiment-based methods. Because they incorporate structural and experimental data that is not used in sequence-based methods, they can provide additional accuracy and reliability to protein function prediction. Here, first we review the definition of protein function. Then the recent developments of these methods are introduced with special focus on the type of predictions that can be made. The need for further development of comprehensive systems biology techniques that can utilize the ever-increasing data presented by the genomics and proteomics communities is emphasized. For the readers' convenience, tables of useful online resources in each category are included. The role of computational scientists in the near future of biological research and the interplay between computational and experimental biology are also addressed.

Keywords: Protein function prediction; functional genomics; functional motifs; gene ontology; unknown gene; genome annotation; protein–protein interaction.

1. Introduction

Computational function annotation, or computational proteomics, plays a crucial role not only in the annotation process of newly sequenced genomes,[1–3] but also in the interpretation of high-throughput experimental data such as gene expression patterns by microarray[4] or protein–protein interaction data.[5–7] In the analysis of these data, even if a detailed function cannot be predicted, prediction of a broader

category of function or subcellular localization greatly helps to cluster genes and reduce unavoidable errors in the data.[8] This computational prediction relies on all kinds of information — protein sequences, co-occurrence of genes across multiple genomes, protein co-expression patterns, protein interactions and protein structures — to accurately confer function primarily to sequences that are otherwise uncharacterized and also to confer additional function to sequences of partially characterized proteins.[9]

This review will focus on the current state of the field of computational function prediction of protein-coding genes and look toward future methods that unify current resources. The text is organized as follows: first we report the number of uncharacterized sequences in genomes. Next we introduce several functional ontologies, or standardized vocabularies describing protein function. Then we overview computational function prediction methods classified into six categories, namely sequence-based, structure-based, association-based, proteomics-experiment-based, process-based, and multi-context-based methods. The main focus of this review is on the latter five approaches, which attempt to predict function beyond traditional sequence-based methods. Special attention is paid to mention the accuracy and limitations, and also which category of Gene Ontology function definitions can be predicted by each method.

1.1. *The need for function prediction*

As of the writing of this review, over 330 complete genomes have been sequenced.[10,11] The sequencing of these genomes has brought to light the discovery of thousands of possible open reading frames which are all potentially transcribed and translated into gene products, but a great majority of these have yet to be characterized. An analysis of entries in the KEGG Genome collection,[11] which includes incomplete and complete genomes, highlights the need for the field of computational function prediction (Fig. 1). Of the 345 genomes listed, 222 contain at least 50% of gene entries with an ambiguous functional annotation (putative, probable, and unknown), including well-characterized organisms such as *E. coli* (51.17%) and *C. elegans* (87.92%). In annotating a genome sequence, identifying genes is a difficult task, especially for higher eukaryotic genomes.[12,13] Hence these entries certainly contain both incorrect and non-coding open reading frames (ORFs). For an overview of computational gene finding, please refer to recent excellent reviews.[12,14,15] Although misidentified ORFs may contribute somewhat to the number of unannotated genes presented here, the lack of annotation in even widely studied species indicates a compelling need and strong potential for assignment of computationally predicted functions to uncharacterized or functionally ambiguous protein-coding genes.

1.1.1. *Using unified vocabularies to describe protein function*

A major hurdle to annotate function to genomes of different organisms is a lack of coherence in functional annotation.[16] For example, some genomes are annotated

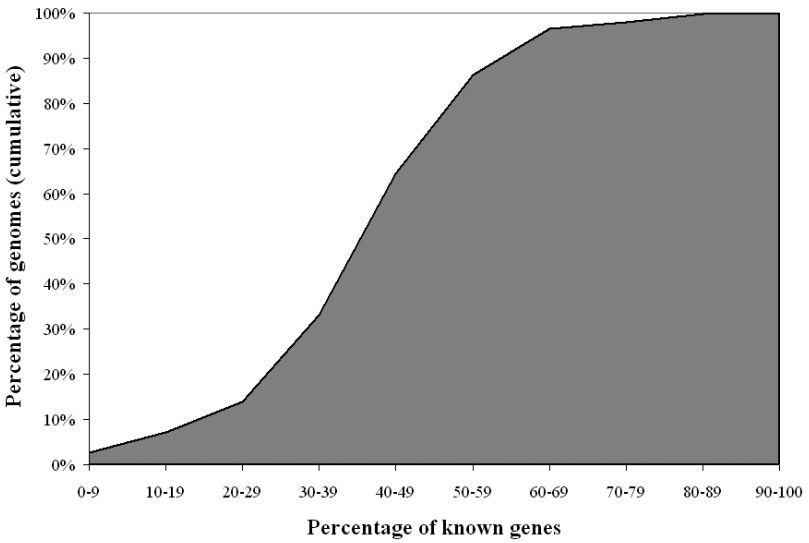


Fig. 1. Analysis of the 345 genomes from KEGG: distribution of unknown function annotations among entries (unknown annotations include the terms "hypothetical," "unknown," "uncharacterized," "predicted," "no hits," "codon recognized," "expressed protein," "conserved protein," and "Vng").

with a full functional description for each entry while others simply use a short name. Terms are used interchangeably, like "putative," "probable," "potential," etc. Computationally, it is difficult even to identify proteins of the same function if a slightly different notation is used. A single function can exist in multiple genomes as completely different descriptions, any of which can be propagated to new entries through current automatic annotation methods.[17,18]

There are several efforts, however, to resolve this issue. The use of ontologies provides support for universal function definitions, which will facilitate uniformity of functional annotations across many databases.[19] An ontology is a hierarchical framework of categorized consensus vocabularies describing function. Among the most widely utilized vocabularies describing protein function are the Gene Ontology (GO),[20] Enzyme Commission (E.C.),[21] and MIPS Functional Catalogue (FunCat).[22] The GO consists of three individual, hierarchical ontologies containing terms which describe molecular function (biochemical activity), biological process (pathway), and cellular component (localization). GO terms annotated to protein sequences carry evidence codes which describe the experimental or computational evidence for the annotation. The E.C. is a four-level hierarchy of enzyme nomenclature describing biochemical activity. MIPS FunCat is a six-level hierarchical scheme used for genome annotation containing over 1300 terms in 28 general categories. GO is the most recent and frequently updated of these ontologies, and there are several individual efforts to improve on its basic format. Several annotation methodologies use a condensed version of GO for functional assignment (e.g. ProtFun[23,24] and

CHUGO[25]), and Myhre *et al.*[26] have devised a Second Layer of connections over the GO that identifies relationships between terms in the three category ontologies.

The organizational structure of the GO lends itself to use in computational biology well, as shown by its widespread use (over 800 PubMed abstracts describing its use since July 2003) and inclusion in the function prediction category of the sixth and seventh biennial CASP (Critical Assessment of Techniques for Protein Structure Prediction).[27] Because GO is so recently popular and because its scheme seems to be very comprehensive, we will use it here to describe the nature of functional annotations that can be derived by each of the predictive methods we describe.

2. Methods of Protein Function Prediction

Because of the complexity and breadth of protein function, the assignment of some function to an experimentally uncharacterized gene product can (and should) be approached from multiple directions. Here, we break down functional cues into six distinct categories — sequence-based, structure-based, association-based, proteomics-experiment-based, process-based, and multi-context-based — and discuss the current available methods in each category, as well as note the types of functional clues in the GO each method can reveal. Sequence-based methods[28–37] are described only briefly because recent review articles summarize them well.[16,38] The available online resources listed in the text are included in the tables accompanying each section.

2.1. *Sequence-based (evolutionary context)*

Probably the most widely utilized computational technique for function assignment is sequence alignment. Direct comparison of an uncharacterized sequence to sequences of known function in a sequence database can reveal evolutionary conservation among species (homology). The most frequently used database search tools include FASTA[28,29] and the BLAST suite (including BLAST[30] and PSI-BLAST[31]). There are also a number of functional motif and alignment databases, including PROSITE,[32] Blocks,[33] SMART,[34,35] PRINTS,[36] and Pfam.[37] Combining different types of sequence-based methods can lead to better function predictions.[38,39] Because there are several excellent reviews on sequence-based function annotation,[16,38,40] here we focus on more recent and interesting ideas, including extensive use of conventional database search methods.

2.1.1. *PFP: Enhanced usage of PSI-BLAST*

Recently, it has been recognized that homology search tools, especially PSI-BLAST, retrieve much more useful information than we can easily use by drawing a uniform *E*-value threshold. Based on this observation, we have designed a novel algorithm, named PFP, which extends a PSI-BLAST search in three ways.[41] First, we extract and score GO annotations based on the frequency of their occurrence in retrieved

sequences. Second, GO terms are also extracted from weak hits in a PSI-BLAST search, which are far below the conventional *E*-value threshold. Weakly similar sequences are not recognized as orthologs to the query sequence but often represent proteins sharing a common functional domain. Third, we additionally consider GO function pair associations which frequently occur in a reference database. Using these associations, GO terms which do not occur in sequence hits but strongly associate to an occurring GO term in the search are also scored. Because the score of each GO term is summed up over the sequence hits, when a detailed biochemical function cannot be predicted, a function of "low-resolution", i.e. one which locates at a higher level on the GO hierarchical tree, can be predicted by PFP. The low-resolution function can be valuable information for omics-type analyses, including clustering of gene expression data from microarrays, where functional clues of genes are essential for interpreting data.

There are several other methods developed recently that build on BLAST or PSI-BLAST searches, using the GO as a prediction standard. Goblet maps GO terms associated to sequences retrieved by a BLAST search on to the GO tree.[42] OntoBlast scores each GO term by multiplying the *E*-values of retrieved sequences.[43] GoFigure and GOtcha consider the hierarchy of the GO tree in scoring GO terms.[44,45] In the GOtcha approach, the score for each GO term is additively propagated onto all of its parents in the GO tree such that the category root always has the highest score.[45] GOPET applies a Support Vector Machine (SVM), a machine learning technique, using BLAST results (including alignment length and *E*-value), GO term frequency, GO term relationships between homologs, annotation quality of the homologs, and the level of annotation within the GO hierarchy as inputs to predict GO terms for a query sequence.[46]

There are also several sequence-based methods that do not rely on comparison by alignment. ProtFun[47] uses post-translational modifications to classify human proteins into functional categories by neural network. Cai *et al.*[48] use an SVM to classify proteins between functionally distinguishable groups. Vries *et al.*[49] use a Bayesian probabilistic model to analyze the distribution of contiguous sequences of four amino acids, and they found that these sequences could be used to distinguish between Pfam families 70% of the time. These methods can be useful for deriving functional clues in the absence of significant annotated homologous sequences in searchable databases.

2.1.2. *Accuracy and limitations*

Sequence similarity searches are generally considered to be simple and accurate methods of inferring function annotation, and the most reliable methods for function prediction (Table 1). SSEARCH[50,51] is the most accurate of the similarity searching algorithms but is slow. Initial searches can accurately and easily be carried out with FASTA or one of the BLAST variants (BLAST, PSI-BLAST). Within these, FASTA is slightly more accurate then BLAST.[52] PSI-BLAST should be used

Table 1. Web resources for sequence-based function prediction.

Website	Description	URL
Tools		
GoFigure	GO annotation by homology search	http://udgenome.ags.udel.edu/frm_go.html
FASTA	Sequence homology search	http://www.ebi.ac.uk/fasta33/
BLAST suite	Sequence homology search	http://www.ncbi.nih.gov/BLAST/
Pfam	Protein family alignment database	http://www.sanger.ac.uk/Software/Pfam/
PROSITE	Functional motif database	http://us.expasy.org/prosite/
Blocks	Database of conserved regions of proteins	http://blocks.fhcrc.org/
SMART	Domain-based annotation resource	http://smart.embl-heidelberg.de/
PRINTS	Protein fingerprint database	http://bioinf.man.ac.uk/dbbrowser/PRINTS/
ProtFun	GO functional category prediction	http://www.cbs.dtu.dk/services/ProtFun/
PFP	GO function prediction server	http://dragon.bio.purdue.edu/pfp
Gotcha	GO function prediction software	http://www.compbio.dundee.ac.uk/Software/Gotcha/gotcha.html
Databases		
KEGG Genes	Genome sequence collection	http://www.genome.jp/kegg/genes.html
COG	Clusters of orthologous groups of proteins	http://www.ncbi.nlm.nih.gov/COG/

to find more distant homologies. Significant sequence similarity is a very strong indicator of homology. Even insignificant hits in a similarity search can offer functional clues,[53] but these scores generally indicate a more distant relationship and are less reliable than strong hits.[52]

Experimentalists relying on sequence alignment for annotation of a query protein must be careful when evaluating the results of sequence similarity searches. Top hit sequences (using BLAST) for open reading frames in *E. coli* fail to represent the closest phylogenetic neighbor 27.3% of the time.[54] This rate is higher for less well characterized organisms but improved upon by using techniques that combine multiple similar sequences to predict annotations.[41,45] The associative data mining used by PFP can add up to 20% coverage to a BLAST search.[41] All sequence similarity-based annotation is limited by the availability of characterized and accurately annotated sequences in public databases. Many BLAST hits are hypothetical or electronically annotated proteins. Erroneous electronic annotations can easily be propagated through a sequence database and contaminate search results with inaccuracies.[55] Sequence similarity can be a powerful method for function annotation, but researchers should be cautious that similarity does not always imply homology, and be aware of inaccurate annotations in public databases. The presence of erroneous annotations in public databases is a concern for any predictive technique utilizing this data for training or evaluation.

Protein family, fingerprint, and motif searching ranges between highly accurate and questionable, depending on the particular motifs recognized in a query protein. This is directly dependent on the strength of conservation and the length of the found motif. Some short motifs such as phosphorylation sites produce lots of false positives so that careful manual inspection is needed.

Depending on the nature of function annotations associated with hits, sequence-based similarity searching algorithms can reveal functions in any or all of the GO categories. That is, if the molecular function, biological process, or cellular component of a homologous or orthologous protein is defined, that definition can be predicted for the query sequence. Motif searches will often reveal signaling peptides that can localize a protein to a certain component of the cell or specific binding or active sites that confer distinct molecular function to a region of the protein. Either of these two predictions can be made without the knowledge of homologous proteins. A biological process, however, can only be matched by sequence-based searching if there is strong global sequence alignment and a defined biological process for the matching homolog to the input sequence.

2.2. *Structure-based*

The function of a protein is inherently linked to its structure. This association is the source of a variety of structure-based function prediction methods,[56,57] which can be grouped into either global fold similarity searching methods or local structure definition (active site characterization).

The advent of structural genomics projects,[58] where structures are solved for novel proteins of unknown function extracted from genomics data, presents strong need for function prediction directly from structural knowledge. For protein sequences lacking both experimentally determined functions and structures, a predicted structure can provide sufficient structural signatures for function prediction. Methods for structure prediction are increasingly more abundant and accurate,[59,60] including homology modeling,[61] *ab initio* modeling[60,62–65] and threading[66–71] methods, which thread a query sequence through a library of known protein folds. Recently, the structures of 60–70% of the proteins from a single genome can be predicted.[66,72]

Proteins that share 30% sequence similarity are generally recognized as having similar folds,[73] and global folds tend to be more conserved than amino acid sequence in the course of evolution.[74,75] Thus in many cases evolutionary relationship, i.e. similarity in function, can be further inferred by comparing protein structures beyond what is found in just the sequence. Protein structure classification databases, which catalogue relationships between protein structure and function, such as SCOP (Structural Classification of Proteins)[76] and CATH,[77,78] have thus become useful resources for predicting protein function. Analysis of the CATH domain database indicates that structural alignments can predict family homology for 80–90% of the entries in PDB[79] that currently lack sequence matches[80] and

that fold is conserved through biological pathways, where a small molecule may be passed successively from protein to protein.[81] Functional similarity based on CATH fold conservation is accurate to 95%.[82] It should be noted, however, that there are several examples of protein families in which global fold similarity does not correlate with functional similarity, including the TIM barrel fold, ferredoxin fold, and Rossmann fold. Function assignment in these cases can be confirmed by conservation of the residues in the active site.[66,83]

Although global fold similarity can be used in many cases to assign a degree of functional similarity, predictions of specific biochemical or enzymatic function can be more accurately obtained from local fold similarity, i.e. in and around the protein active site. An active site of an enzyme can be described by a template which consists of mutual distances and angles of catalytic amino acid residues. A catalytic triad of serine protease can be intuitively represented by this description.[84] A practical advantage of this approach is that it is relatively simple for implementation and matching so that large-scale scanning is possible.[83,85] Catalytic Site Atlas is a hand-annotated collection of catalytic sites of enzymes which can be used for large-scale function prediction.[86]

A different approach to describing binding sites of small chemical compounds, including active sites of enzymes, is to represent the three-dimensional surface shape of the local sites. Most binding sites can be identified geometrically as a local cavity region on a protein surface.[87–89] CASTp is a database of cavities of proteins which allows users to compute cavities of a protein of the user's interest.[90] eF-site is a database of the geometric surface shape and the chemical properties of functional sites of proteins.[91]

There are several other online resources for structure and structural motif classification (Table 2). ProFunc[92] combines various structure matching techniques, including searching against superfamily HMMs, existing PDB global structures, 3D functional templates, surface cleft and nest analysis, and a variety of sequence-based searches to provide probable GO annotations to a query structure.[93] Q-SiteFinder identifies hydrophobic patches on a protein surface which are possible ligand binding sites.[94] WebFEATURE[95] exhaustively searches an input structure for conserved radial microenvironments that represent a distinct set of protein active and binding sites.[96]

2.2.1. *Accuracy and limitations*

Structural similarity is also a very accurate method of predicting function. The global fold of a protein determines the shape and the location of active and binding sites, and the local structural environment determines the catalytic mechanisms of enzymes. There are caveats to using structures to imply homology, however. As mentioned above, there are several examples of global folds that are known to perform varying functions. Identification of a conserved active site is straightforward for predicting function, but methods which do this are relatively new and still in

Table 2. Web resources for structure-based function prediction.

Website	Description	URL
Tools		
ProFunc	3D functional template matching	http://www.ebi.ac.uk/thorntonsrv/databases/ProFunc/
Q-Site Finder	Hydrophobic binding site search	http://www.bioinformatics.leeds.ac.uk/qsitefinder/
WebFEATURE	3D active site matching	http://feature.stanford.edu/ webfeature/
Databases		
CATH	Structure classification database	http://www.biochem.ucl.ac.uk/ bsm/cath/
SCOP	Structure classification database	http://scop.berkeley.edu/
Catalytic Site Atlas	Catalytic site database	http://www.ebi.ac.uk/thorntonsrv/databases/CSA/
CASTp	Protein binding site database	http://cast.engr.uic.edu/cast/
PROCAT	3D active site template database	http://www.biochem.ucl.ac.uk/bsm/PROCAT / PROCAT.html
SURFACE	Protein binding site database	http://cbm.bio.uniroma2.it/surface/
eF-site	Protein binding site database	http://ef-site.protein.osaka-u.ac.jp/eF-site/

development. They rely on not only strong conservation to make a reliable prediction, but also the presence of enough examples to build an accurate model for searching. As more structures become available, active site searching will become a much more reliable means of function prediction. For now, though, structure-based approaches should be followed by close manual inspection of prediction results and coupled with sequence-based or additional methods for reliability.

Specific predictions of active sites and small molecule binding sites can lead directly to predictions of biochemical activity. These kinds of interactions are described by the GO molecular function category. Protein–protein binding sites can give clues toward either specific functions or biochemical pathways. If an interaction can be predicted with a particular protein, it is likely that their functions correlate in some way. Conserved global fold, because of its similarity to sequence conservation, can be an indicator of specific function, but it can additionally predict process relatedness to a group of proteins.

2.3. *Association-based (genomic context)*

The genome as a physical entity provides a set of distinct methods for prediction of protein function. Gene organization is a fundamental source of contextual cues, both within organisms and across species. There are three distinct genomic associations that can be linked to functional associations: (1) conservation of genomic proximity, or conservation of gene order; (2) gene/domain fusion events; and (3) similarity of phylogenetic profile (Fig. 2).

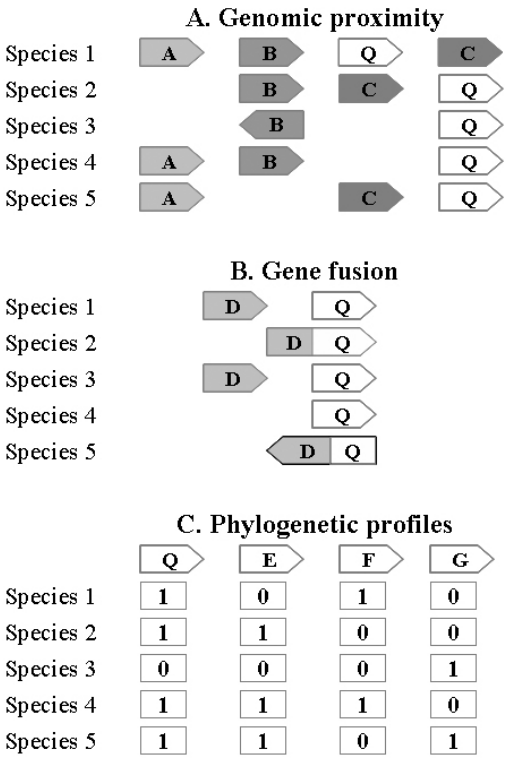


Fig. 2. Association-based methods ("Q" is the query protein, "A" through "G" are other proteins): [A] Genomic proximity — consistent neighbors ("A", "B," and "C") to the query protein indicate functional association. Note that gene orientation and order are not necessarily conserved; [B] gene/domain fusion — fusion of a gene ('D') with the query protein in some species indicates functional and likely physical association; and [C] phylogenetic profiles (1 and 0 bits indicate the presence or absence of the gene in the species, respectively) — similar profiles (similarity to the query protein: "E" > "F" > "G") indicate functional association.

A classic example of genomic context being linked to functional association is the operon: a group of genes at a single locus that operate under a common regulatory mechanism and perform functions leading to a common goal.[97] Operons are evolutionarily conserved in terms of organization, occurrence, and regulation. Although all genes are not associated to others this strictly, conserved gene orders among genomes can serve as important clues for assigning functions to uncharacterized gene products. If two genes retain close proximity even across large phylogenetic distances, the likelihood of a functional association is high because that indicates the presence of a selective evolutionary force maintaining the gene organization. But of course, conversely, a lack of conserved proximity does not indicate a lack of functional association.

Practically, comparative genomics analysis has two steps: It starts with identifying corresponding genes, i.e. orthologous genes, or more permissive, highly

homologous genes in a set of input genomes. Then a cluster of genes is defined as a set of genes locating on a same strand and within a certain distance (e.g. within 300 base pairs[98]). The COG database,[99,100] Pfam database,[37] KEGG,[11] or MBGD[101] will provide a pre-calculated set of orthologous genes. Note here that the conservation of the gene order basically suggests functional association of included genes in the cluster and not necessarily physical association, although in some cases it does.[102] The function of some classes of membrane proteins, which can account for 30% of some genomes, can be inferred from the number of transmembrane helices[103] and the neighboring genes,[104,105] i.e. by a combination of structural and genomic context information.

Another clue for functional association is the presence of domain fusion events, wherein two proteins that are expressed as independent proteins in one organism are expressed as two domains fused into a single protein in another organism.[106,107] This kind of permanent evolutionary linkage is a strong indicator of some functional association. In this case, it is most likely that physical association of the gene products in addition to functional association could be predicted.

The third genomic approach is to examine the co-occurrence of genes. Those genes that show correlated presence and absence across genomes are assumed to be functionally linked.[108,109] The co-occurrence and the co-absence of a gene in genomes are expressed as a vector of one or zero bits, which is called a "phylogenetic profile." Genes with a similar phylogenetic profile are predicted to have functional association.

STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) is an interactive database of known and predicted functional interactions between genes.[110] A primary sequence can be searched either as a protein or as a Cluster of Orthologous Groups (COG) against over 440,000 genes from 110 organisms (release 5.1). Functional associations are predicted via genomic neighborhood, domain fusions, co-occurrence, co-expression pattern in microarray analyses, and previous knowledge mined from PubMed abstracts.[111] The likelihood of interaction through each of these methods is normalized and summed for each predicted partner, and the top-scoring hits are listed as predicted functional associations.

2.3.1. *Accuracy and limitations*

Comparative genomics methods are indirect methods for inference of functional association, and accuracy varies widely with the strength of the conserved association. Phylogenetic profiling in yeast yielded 50% specificity and 58% sensitivity for mitochondrial proteins.[112] This is consistent with contextual genomic clues in *Mycoplasma genitalium*, where approximately 50% of genes can be characterized by strong conserved gene order, neighborhood, fusion, and co-occurrence.[113] It should be cautioned that even strong genomic association does not necessarily indicate functional similarity but could instead indicate an opposite or complementary functional association.

Taking advantage of the increasing number of complete genome sequences, the association-based (comparative genomics) methods have become a promising and rich source of information for protein function prediction. These methods can predict the biological process and/or cellular component of genes.

2.4. *Proteomics experiment-based (cellular context)*

Proteomics is driven by high-throughput experimental techniques that generate tremendous volume of data. Conceptually, the computational analysis of proteomics data falls into one of two categories which are discussed here: interaction-based methods, which utilize protein–protein and genetic interaction data, and expression-based methods, which utilize microarray gene expression data.

The field of proteomics has generated and is generating tremendous amounts of physical interaction data through mass spectrometry[6,114] 2D gel electrophoresis,[7] yeast two-hybrid methods[5,115,116] and protein chips.[117] High-throughput protein–protein interaction screens offer novel and increasingly more accurate and comprehensive association data that can be utilized to predict functions for uncharacterized proteins.[118] Marcotte *et al.*[119] constructed an interaction network for *Saccharomyces cerevisiae* proteins that yielded 93,750 functional links using hard experimental data as well as predicted interactions from genomic and phylogenetic context, including co-expression, domain fusions, and phylogenetic profiles. Using this network, a general function was assigned to 1600 previously uncharacterized proteins. Since then, protein–protein interaction networks have become a focus of functional proteomics.[1]

There are several resources available for the study of protein interactions, including the Database of Interacting proteins (DIP),[120] Biomolecular Interaction Network Database (BIND),[121–123] and Molecular Interaction (MINT) database,[124] all of which serve as an excellent starting point for bioinformatics analyses (Table 3).

Protein–protein interaction data can be represented as interaction graphs or networks. Functional association can be roughly inferred from an interaction graph quite easily just by assuming that proteins that lie within short distances of each other are more likely to share functional qualities than those that are separated further. Early interaction-based methods took this approach, either assigning function to a protein based directly on the most common functions[125] or interaction partners (clusters)[126,127] shared by its direct neighbors in the graph. Another approach is to annotate uncharacterized proteins in order to minimize the number of protein interactions between different functional categories.[128] It is also common to use clustering methods to distinguish groups of proteins sharing a high number of interactions[129,130] or close proximity[131] among themselves. In these interaction networks, molecular function is often shared by proteins interacting with a common partner (e.g., A and B may share a function if both interact with C). These proteins are referred to as second-level interaction partners as they are separated by two edges in an interaction graph.[132]

Table 3. Web resources for association-based, interaction-based, expression-based, and process-based function prediction.

Website	Category	Description	URL
Tools			
PathoLogic	Process	Pathway hole filler software	contact ptools-support @ai.sri.com
GOMiner	Microarray	Microarray analysis software	http://discover.nci.nih.gov/gominer/
GenMAPP	Microarray	Microarray analysis software	http://www.genmapp.org/
Databases			
NCBI GEO	Microarray	Microarray data repository	http://www.ncbi.nlm.nih.gov/geo/
ArrayExpress	Microarray	Microarray data repository	http://www.ebi.ac.uk/ arrayexpress/
Stanford Microarray Database	Microarray	Microarray data repository	http://genome-www5.stanford.edu/
OU Microarray Database	Microarray	Microarray data repository	http://www.ou.edu/microarray/
STRING	Association	Functional association database	http://string.embl.de/
DIP	Interaction	Protein–protein interaction database	http://dip.doe-mbi.ucla.edu/
BIND	Interaction	Biomolecular interaction database	http://bind.ca/
MINT	Interaction	Molecular interaction database	http://mint.bio.uniroma2.it/mint/
MIPS	Interaction	Protein function, expression and interaction database	http://mips.gsf.de/
POINT	Interaction	Orthologous interactions database	http://point.nchc.org.tw/
KEGG Pathway	Process	Metabolic pathway database	http://www.genome.jp/kegg/pathway.html

The concept of protein–protein interactions serves as a base for several derivative methods that can be of use in function annotation. Using the idea that physical interactions are homologous across organisms,[133] Yu *et al.*[134] mapped homologous protein–protein interactions (interologs) and protein–DNA interactions (regulogs) for several organisms. Studies in systems biology also frequently characterize "genetic" interactions, which characterize epistatic, epigenetic, and environmental factors by systematic targeted gene mutation and subsequent phenotypic analysis.[135–139] The results of these genetic interaction studies are analyzed in complex, multi-dimensional maps similar to those based on physical protein–protein interactions.[140]

Microarrays examine expression patterns of hundreds to tens of thousands of genes in a single experiment. Since its development in 1995,[141–143] this technology has become a standard experimental method used in a wide-range of research fields in biology.[144–146] Gene expression data can be analyzed to extract statistically

significant clusters of genes that are likely to be involved in similar or coordinate biological processes.[147–151] Recently, several tools have been created to map microarray data onto biologically significant frameworks. GOMiner[152] translates genes into GO terms and displays them on the GO tree so that enriched functional subgraphs can be easily identified, and GenMAPP[153,154] visualizes microarray data on KEGG or user-customizable biological pathways. Because of the high volume of publicly available data, microarrays have become a rich source of data for functional analysis.

Several large microarray repositories collect this large-scale expression data and make it publicly available. These include NCBI's Gene Expression Omnibus (GEO),[155] EBI's ArrayExpress,[156] the Stanford Microarray Database (SMD),[157] and the Oklahoma University Microarray Database. Together these databases hold data produced from on the order of hundreds of thousands of experiments. Generally, interpretation of these data focuses on clustering genes which are expressed coordinately under certain conditions or along a timescale. Genes which follow similar expression patterns are assumed to be participating in a common biological process or response. Thus, algorithms which aim to identify these statistically significant groups make up the vast majority of available tools. For prediction of gene function, methods examine enrichment of GO terms in significant clusters[152,158] and mine patterns of expression among groups of similar GO terms (so-called "prior-knowledge based clustering").[159] Once a significant function is determined for a cluster of genes, it can be annotated to all the genes in that cluster.

2.4.1. *Accuracy and limitations*

One of the problems in the use of these methods to predict protein function is the presence of intrinsic errors in genome-scale proteomics data. In an earlier study, comparison of the interaction data sets from different yeast genome-wide experiments showed that the overlap of detected interactions among them is not large.[8,160–163] Interaction and expression can be spatially and temporally unique, and readily change in different cell types and under different conditions. Therefore, these data should be used in combination, as Marcotte *et al.* have done (see above). But there are efforts from the computational side to reduce the noise and extract only reliable information from interaction data. Those include assessment of the quality of the interaction data by comparing expression profiles of the protein pairs,[162] comparison of the function category and subcellular locations of the protein pairs using the Gene Ontology (GO) database,[161,164,165] a statistical approach using various data from the experiments,[164] and comparison of topological resemblance between interaction networks.[166] The POINT database[167] uses orthologous interaction pairs, that is, if orthologous proteins in different organisms are detected to form an interaction, those interactions are likely to be more reliable. They combined the orthologous interactions with localization data from GO and microarray expression clusters from both human and yeast cell cycle microarray databases. In the same way, predictions made using microarray data have been cross-validated

and supplemented with predictions made from interaction data.[168,169] In considering any kind of high-throughput preoteomics data in a prediction technique, it is important to consider that these experimental methods are a trade-off between accuracy and scale. Therefore, the most accurate predictive techniques will involve combining several types of heterogeneous data for validation and noise reduction (see Sec. 2.6).

Interaction and coordinate expression data lead to the prediction of biological process association. Second-level interactions could be used to predict a molecular function, but generally the data from large interaction networks do not translate well into specific functions. If a particular process or interaction partner is localized to a distinct region of the cell, however, cellular components can be predicted by interaction-based methods.

2.5. *Process-based (metabolic context)*

Evolution of the cell has established organized networks in the form of metabolic pathways. These can be taken advantage of as an additional method for the prediction of protein functions. The uniqueness of utilizing metabolic pathways to predict the function of uncharacterized proteins is that gaps or holes in known pathways can be and have been intuitively assigned a function, and that function simply awaits a protein to be characterized to match it.[170] The KEGG pathway database of metabolic pathways maps all of the known metabolic pathways, including reactions that do not yet have a protein or enzyme associated with them.[11]

Because this is a fairly new approach, not many attempts have been made to use metabolic pathways to predict the protein function on a genomic scale. A recent publication by Green and Karp[170] analyzed PathoLogic, an algorithm designed to produce a predicted set of metabolic pathways (including holes) using an organism's genomic information, then mine the uncharacterized genes in that organism for potential functional matches to holes in those pathways.[171] Potential matches, or "candidate genes," are found by using BLAST to search isozymes (from other organisms) of missing enzymes in the target organism against the entire genome of the target organism. The matches are then scored using a Bayesian approach according to the probability that they might perform the function of the missing enzyme. The pathway to protein approach takes advantage of being able to search isozyme sequences against the existing, limited genome of the target organism. This kind of global approach is of interest because the most practical limiting boundary to functional genomics is the genome itself, i.e. all of the information needed to assign function to an organism's gene products should exist within the genome.

2.5.1. *Accuracy and limitations*

A process-based approach introduced here does not directly predict the function of a gene by itself, but rather it will indicate that there should be "missing genes"

in a genome that need to fill holes in pathways. If genes can be identified which correspond to missing links in a metabolic pathway, biological process of the genes are literally predicted, most probably together with their molecular function and cellular component.

2.6. *Combined methods (multi-context)*

In this review, we have introduced and summarized the recent developments of computational protein function prediction (Table 4). Stress has been placed on novel and creative methods which directly and indirectly predict biochemical function, interacting protein partners, cellular localization, and even biological pathways of uncharacterized sequences.

Most of these methods predict one or just a few functional aspects of the query protein. In compensating for the limitations of each prediction method, it is effective to integrate many types of function predictions together with available high-throughput proteomic data to form a prediction system (Fig. 3). Deng *et al.*[172] have

Table 4. Web resources for multi-context function prediction.

Website	Category	Description	URL
Tools			
ProKnow	Seq./Struct.	GO function prediction server	http://www.doe-mbi.ucla.edu/Services/ProKnow/

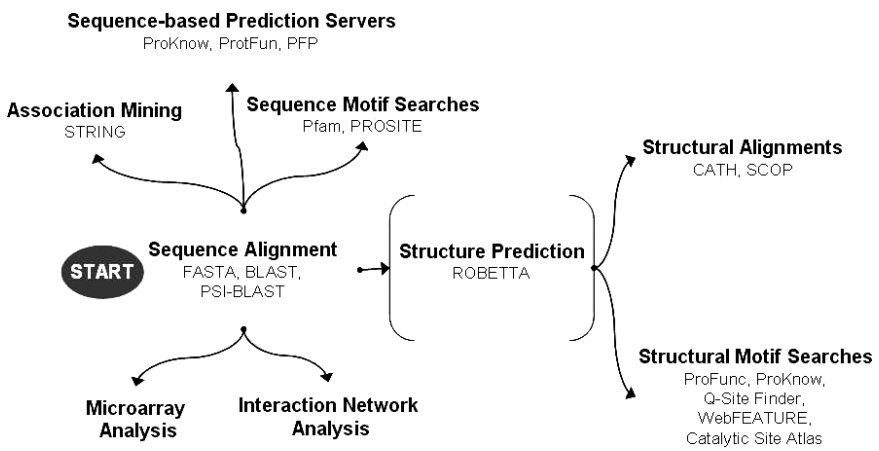


Fig. 3. A practical flowchart for the prediction of a single unknown query sequence. The best starting searches are FASTA, BLAST, and PSI-BLAST, which provide accurate and fast database searching for known similar sequences. If this search does not produce satisfactory results, one may perform several related queries to sequence and structure motif prediction servers (if the structure is not available) or analyses based on high-throughput experimental results (if the experimental data is available).

developed a method that integrates many types of function predictions into an interaction network created from the MIPS Comprehensive Yeast Genome Database[173] using Markov random fields (MRFs).[174] Here, the network weights protein–protein and genetic interaction data, correlated expression data, and protein domain structure information to assign posterior probability of a gene having a particular function. PathoLogic[170] approaches the problem of filling gaps in biological pathways from multiple directions, including sequence similarity and genomic context. Essentially, this approach identifies orthologous pathways and fits gene candidates in the target organism to the profile of the protein performing the orthologous reaction. POINT[167] uses the concept of interologs to identify novel interactions in the human proteome, and references these interactions to available cell cycle microarray expression information and cellular localization. MAGIC[168] incorporates protein–protein and genetic interaction data sets along with transcription factor binding sites into a Bayesian framework to group genes into functional clusters. Marcotte *et al.* combine phylogenetic data with interaction and expression data to predict function.[119]

Certainly, utilizing a variety of contextual clues can filter out noise inherent in computational methods for function annotation. In each of these methods, the significant question involves *how* to combine data from heterogeneous sources. Most of these applications use simple elimination or some form of Bayes' conditional probability to reduce noise, although in the future, artificial intelligence and classification techniques such as neural networks and SVMs are sure to be incorporated in this task. In all of these cases, a common functional vocabulary, e.g. GO, is indispensable for integration of results from varying methods in a function prediction system.

2.7. *Application and evaluation*

There are several cases where application of contextual clues has resulted in functional discovery of a protein or series of proteins. Here we will discuss two examples. First, the thiamine biosynthesis pathway has been found by comparative genomics to include gaps, or reactions for which some species have no identifiable homolog by sequence similarity. Morett *et al.*[175] use the concept of analogous gene replacement to find likely candidates for these gaps. Analogous proteins are functionally equivalent but lack sequence and structural similarity because of evolutionary independence. In their investigation, Morett *et al.* identify analogs by anti-correlation in the phylogenetic profiles of several genes related to thiamine biosynthesis and also use sequence and structural analyses to strengthen their predictions. In this case, the predictions were validated by experimental verification of the predicted functions. Second, the gene frataxin/cyaY, implicated in the neurodegenerative disease Friedreich's ataxia, lacks both experimental characterization and identifiable homology by sequence similarity. Huynen *et al.*[176] identified two genes, hscA and hscB, whose phylogenetic distributions across 56 organisms correlate well with the frataxia gene. All three proteins are known or predicted to localize in the mitochondria, and hscA and hscB are known to be involved in the iron–sulfur cluster protein

assembly. They conclude, therefore, that frataxin may be involved in this process. Both of these are examples of focused computational analyses that utilize a variety of contextual clues to elucidate a putative function for a target protein. Their success indicates strong potential for these same types of analyses to be relevant on a larger scale.

What is currently lacking in the field of function prediction is a real knowledge of how these techniques perform in their ability to provide unique contextual functional clues on a larger scale. This kind of assessment will be vital for understanding how to combine data from heterogeneous sources in a biologically appropriate manner, and subsequently implementing that combined data into a function prediction algorithm. One of the most significant points of concern over the future of the field of protein function prediction is this assessment of predictive performance. Assessment here covers two major points of concern. First is the lack of a "gold standard" data set by which new techniques can prove their merit. Several potential candidates exist, e.g. the *E. coli* or other small, well characterized genomes, but because function prediction methods many times only predict a certain category of function (i.e. biochemical activity or metabolic process), assessment against a universal set of annotations could be difficult and misleading. In several cases, finding the correct function for a gene is difficult even from literature searching, as there are often contradictory results in different papers. Second is the lack of an appropriate method of quantitative assessment of predictions against known annotations. In the context of GO, semantic similarity seems to be the most informative,[177–179] but similarity between GO terms can be assessed purely by the structure of the ontology (edge distance) or textual similarity of term definitions as well. In hierarchical ontologies such as E.C. or SCOP, predictions can be assessed by correct assignment of Superfamily or Family. The problem of assessment is further complicated by the fact that each method of function prediction has a different goal. While some methods attempt to annotate specific biochemical activity to individual residues or groups of residues in a protein structure, others predict low-resolution function for the analysis of high-throughput proteomics data wherein significant clusters may lack any function annotation.

This lack of appropriate assessment methods hinders the practical use of function prediction data by experimentalists and also development of new and better predictive techniques. Thus, assessment solutions are a target of many groups in the function prediction community. Two such efforts are described here. The biennial Critical Assessment of Techniques for Protein Structure Prediction (CASP) has included function prediction as a category in the previous two competitions,[180,181] and the annual Automated Function Prediction (AFP) meeting is trying to organize an ongoing resource for predictions of all types.[179] CASP is a community-wide assessment of protein structure prediction techniques wherein predictors model target proteins which are evaluated against experimentally determined structures. Predictions are made in several categories including structure modeling, domain prediction, and recently, function prediction. AFP is an annual meeting and forum

specifically for those working in the field of function prediction. The organizers of AFP are particularly interested in moderating a CASP-style function prediction assessment. In the end, two questions need to be considered: How useful are these assessments in (1) allowing computational biologists to effectively evaluate algorithm performance and (2) communicating algorithm effectiveness and utility to the experimental biologists who will ultimately use the predictions?

3. Conclusion

The integration of proteomics and computational techniques is vitally important to both computational biologists and experimental biologists alike. Computational techniques enhance the ability of experimentalists to design efficient investigations and ask appropriate questions,[182] and thorough, accurate data sets provide computational biologists with the appropriate resources to develop efficiently targeted algorithms. The near future of biology will see integration of fields — molecular biology, genetics, biochemistry, genomics, proteomics, and bioinformatics — to the end of systems-level investigation and characterization.

Acknowledgments

This work is partially supported by National Institutes of Health (R01 GM075004 and U24 GM077905) and National Science Foundation (DMS 0604776).

References

1. Galperin MY, Koonin EV, Who's your neighbor? New computational approaches for functional genomics, *Nat Biotechnol* **18**(6):609–613, 2000.
2. Andrade MA, Brown NP, Leroy C, Hoersch S, de DA, Reich C, Franchini A, Tamames J, Valencia A, Ouzounis C, Sander C, Automated genome sequence analysis and annotation, *Bioinformatics* **15**(5):391–412, 1999.
3. Venter JC, Smith HO, Hood L, A new strategy for genome sequencing, *Nature* **381**(6581):364–366, 1996.
4. Brown PO, Botstein D, Exploring the new world of the genome with DNA microarrays, *Nat Genet* **21**(1 Suppl):33–37, 1999.
5. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y, A comprehensive two-hybrid analysis to explore the yeast protein interactome, *Proc Natl Acad Sci USA* **98**(8):4569–4574, 2001.
6. Blagoev B, Kratchmarova I, Ong SE, Nielsen M, Foster LJ, Mann M, A proteomics strategy to elucidate functional protein–protein interactions applied to EGF signaling, *Nat Biotechnol* **21**(3):315–318, 2003.
7. Rosen R, Sacher A, Shechter N, Becher D, Buttner K, Biran D, Hecker M, Ron EZ, Two-dimensional reference map of Agrobacterium tumefaciens proteins, *Proteomics* **4**(4):1061–1073, 2004.
8. Bader GD, Hogue CW, Analyzing yeast protein–protein interaction data obtained from different sources, *Nat Biotechnol* **20**(10):991–997, 2002.
9. Eisenberg D, Marcotte EM, Xenarios I, Yeates TO, Protein function in the post-genomic era, *Nature* **405**(6788):823–826, 2000.

10. Genome News Network (GNN), ⟨http://www.genomenewsnetwork.org/⟩, cited October 23, 2004.
11. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M, The KEGG resource for deciphering the genome, *Nucl Acids Res* **32**(Database issue):D277–D280, 2004.
12. Brent MR, Genome annotation past, present, and future: How to define an ORF at each locus, *Genome Res* **15**(12):1777–1786, 2005.
13. Zhang MQ, Computational prediction of eukaryotic protein-coding genes, *Nat Rev Genet* **3**(9):698–709, 2002.
14. Windsor AJ, Mitchell-Olds T, Comparative genomics as a tool for gene discovery, *Curr Opin Biotechnol* **17**(2):161–167, 2006.
15. Do JH, Choi DK, Computational approaches to gene prediction, *J Microbiol* **44**(2):137–144, 2006.
16. Dobson PD, Cai YD, Stapley BJ, Doig AJ, Prediction of protein function in the absence of significant sequence similarity, *Curr Med Chem* **11**(16):2135–2142, 2004.
17. Karaoz U, Murali TM, Letovsky S, Zheng Y, Ding C, Cantor CR, Kasif S, Whole-genome annotation by using evidence integration in functional-linkage networks, *Proc Natl Acad Sci USA* **101**(9):2888–2893, 2004.
18. Gattiker A, Michoud K, Rivoire C, Auchincloss AH, Coudert E, Lima T, Kersey P, Pagni M, Sigrist CJ, Lachaize C, Veuthey AL, Gasteiger E, Bairoch A, Automated annotation of microbial proteomes in SWISS-PROT, *Comput Biol Chem* **27**(1):49–58, 2003.
19. Creating the gene ontology resource: Design and implementation, *Genome Res* **11**(8):1425–1433, 2001.
20. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, Richter J, Rubin GM, Blake JA, Bult C, Dolan M, Drabkin H, Eppig JT, Hill DP, Ni L, Ringwald M, Balakrishnan R, Cherry JM, Christie KR, Costanzo MC, Dwight SS, Engel S, Fisk DG, Hirschman JE, Hong EL, Nash RS, Sethuraman A, Theesfeld CL, Botstein D, Dolinski K, Feierbach B, Berardini T, Mundodi S, Rhee SY, Apweiler R, Barrell D, Camon E, Dimmer E, Lee V, Chisholm R, Gaudet P, Kibbe W, Kishore R, Schwarz EM, Sternberg P, Gwinn M, Hannick L, Wortman J, Berriman M, Wood V, de la CN, Tonellato P, Jaiswal P, Seigfried T, White R, The Gene Ontology (GO) database and informatics resource, *Nucl Acids Res* **32**(Database issue):D258–D261, 2004.
21. Nomenclature committee of the international union of biochemistry and molecular biology (NC-IUBMB), Enzyme Supplement 5 (1999), *Eur J Biochem* **264**(2):610–650, 1999.
22. 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, *Nucl Acids Res* **32**(18):5539–5545, 2004.
23. Jiang L, Lund O, Tan JQ, Selection of proteins for human MHC class II presentation, *Cell Mol Immunol* **2**(1):49–56, 2005.
24. Jensen LJ, Ussery DW, Brunak S, Functionality of system components: Conservation of protein function in protein feature space, *Genome Res* **13**(11):2444–2449, 2003.
25. Eisner R, Poulin B, Szafrom D, Lu P, Greiner R, Improving proetin function prediction using the hierarchical structure of the gene ontology, *IEEE CIBCB 2005* (2005).
26. Myhre S, Tveit H, Mollestad T, Laegreid A, Additional gene ontology structure for improved biological reasoning, *Bioinformatics* **22**(16):2020–2027, 2006.

27. Critical Assessment of Techniques for Protein Structure Prediction 6 (CASP6), ⟨http://predictioncenter.llnl.gov/casp6/Casp6.html⟩, cited October 23, 2004.
28. Pearson WR, Lipman DJ, Improved tools for biological sequence comparison, *Proc Natl Acad Sci USA* **85**(8):2444–2448, 1988.
29. Pearson WR, Rapid and sensitive sequence comparison with FASTP and FASTA, *Methods Enzymol* **183**:63–98, 1990.
30. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ, Basic local alignment search tool, *J Mol Biol* **215**(3):403–410, 1990.
31. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ, Gapped BLAST and PSI-BLAST: A new generation of protein database search programs, *Nucl Acids Res* **25**(17):3389–3402, 1997.
32. Sigrist CJ, Cerutti L, Hulo N, Gattiker A, Falquet L, Pagni M, Bairoch A, Bucher P, PROSITE: A documented database using patterns and profiles as motif descriptors, *Brief Bioinform* **3**(3):265–274, 2002.
33. Henikoff S, Henikoff JG, Pietrokovski S, Blocks+: A non-redundant database of protein alignment blocks derived from multiple compilations, *Bioinformatics* **15**(6):471–479, 1999.
34. Letunic I, Copley RR, Schmidt S, Ciccarelli FD, Doerks T, Schultz J, Ponting CP, Bork P, SMART 4.0: towards genomic data integration, *Nucl Acids Res* **32**(Database issue):D142–D144, 2004.
35. Schultz J, Copley RR, Doerks T, Ponting CP, Bork P, SMART: A web-based tool for the study of genetically mobile domains, *Nucl Acids Res* **28**(1):231–234, 2000.
36. Attwood TK, Bradley P, Flower DR, Gaulton A, Maudling N, Mitchell AL, Moulton G, Nordle A, Paine K, Taylor P, Uddin A, Zygouri C, PRINTS and its automatic supplement, prePRINTS, *Nucl Acids Res* **31**(1):400–402, 2003.
37. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ, Yeats C, Eddy SR, The Pfam protein families database, *Nucl Acids Res* **32**(Database issue):D138–D141, 2004.
38. Attwood TK, The role of pattern databases in sequence analysis, *Brief Bioinform* **1**(1):45–59, 2000.
39. Tian W, Arakaki AK, Skolnick J, EFICAz: A comprehensive approach for accurate genome-scale enzyme function inference, *Nucl Acids Res* **32**(21):6226–6239, 2004.
40. Watson JD, Laskowski RA, Thornton JM, Predicting protein function from sequence and structural data, *Curr Opin Struct Biol* **15**(3):275–284, 2005.
41. Hawkins T, Luban S, Kihara D, Enhanced automated function prediction using distantly related sequences and contextual association by PFP, *Protein Sci* **15**(6):1550–1556, 2006.
42. Hennig S, Groth D, Lehrach H, Automated gene ontology annotation for anonymous sequence data, *Nucl Acids Res* **31**(13):3712–3715, 2003.
43. Zehetner G, OntoBlast function: From sequence similarities directly to potential functional annotations by ontology terms, *Nucl Acids Res* **31**(13):3799–3803, 2003.
44. Khan S, Situ G, Decker K, Schmidt CJ, GoFigure: Automated gene ontology annotation, *Bioinformatics* **19**(18):2484–2485, 2003.
45. Martin DM, Berriman M, Barton GJ, GOtcha: A new method for prediction of protein function assessed by the annotation of seven genomes, *BMC Bioinform* **5**:178, 2004.
46. Vinayagam A, del VC, Schubert F, Eils R, Glatting KH, Suhai S, Konig R, GOPET: A tool for automated predictions of Gene Ontology terms, *BMC Bioinform* **7**:161, 2006.

47. 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* **319**(5):1257–1265, 2002.
48. Cai CZ, Wang WL, Sun LZ, Chen YZ, Protein function classification via support vector machine approach, *Math Biosci* **185**(2):111–122, 2003.
49. Vries JK, Munshi R, Tobi D, Klein-Seetharaman J, Benos PV, Bahar I, A sequence alignment-independent method for protein classification, *Appl Bioinform* **3**(2–3):137–148, 2004.
50. Pearson WR, Searching protein sequence libraries: Comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms, *Genomics* **11**(3):635–650, 1991.
51. Smith TF, Waterman MS, Identification of common molecular subsequences, *J Mol Biol* **147**(1):195–197, 1981.
52. Brenner SE, Chothia C, Hubbard TJ, Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships, *Proc Natl Acad Sci USA* **95**(11):6073–6078, 1998.
53. Hawkins T, Kihara D, PFP: Automatic annotation of protein function by relative GO association in multiple functional contexts, *13th Annual Int Conf Intell Syst Mol Bio* pp. 117, 2005.
54. Koski LB, Golding GB, The closest BLAST hit is often not the nearest neighbor, *J Mol Evol* **52**(6):540–542, 2001.
55. Karp PD, What we do not know about sequence analysis and sequence databases, *Bioinformatics* **14**(9):753–754, 1998.
56. Pal D, Eisenberg D, Inference of protein function from protein structure, *Structure* (*Camb.*) **13**(1):121–130, 2005.
57. Ponomarenko JV, Bourne PE, Shindyalov IN, Assigning new GO annotations to protein data bank sequences by combining structure and sequence homology, *Proteins*, Epub., 2005.
58. Skolnick J, Fetrow JS, Kolinski A, Structural genomics and its importance for gene function analysis, *Nat Biotechnol* **18**(3):283–287, 2000.
59. Moult J, Fidelis K, Zemla A, Hubbard T, Critical assessment of methods of protein structure prediction (CASP)-round V, *Proteins* **53**(Suppl 6):334–339, 2003.
60. Skolnick J, Zhang Y, Arakaki AK, Kolinski A, Boniecki M, Szilagyi A, Kihara D, TOUCHSTONE: A unified approach to protein structure prediction, *Proteins* (53 Suppl 6):469–479, 2003.
61. Fiser A, Sali A, Modeller: Generation and refinement of homology-based protein structure models, *Methods Enzymol* **374**:461–491, 2003.
62. Kihara D, Zhang Y, Lu H, Kolinski A, Skolnick J, Ab initio protein structure prediction on a genomic scale: Application to the Mycoplasma genitalium genome, *Proc Natl Acad Sci USA* **99**(9):5993–5998, 2002.
63. Kihara D, Lu H, Kolinski A, Skolnick J, TOUCHSTONE: An ab initio protein structure prediction method that uses threading-based tertiary restraints, *Proc Natl Acad Sci USA* **98**(18):10125–10130, 2001.
64. Rohl CA, Strauss CE, Chivian D, Baker D, Modeling structurally variable regions in homologous proteins with rosetta, *Proteins* **55**(3):656–677, 2004.
65. Jones DT, McGuffin LJ, Assembling novel protein folds from super-secondary structural fragments, *Proteins* **53**(Suppl 6):480–485, 2003.
66. Kihara D, Skolnick J, Microbial genomes have over 72% structure assignment by the threading algorithm PROSPECTOR_Q, *Proteins* **55**(2):464–473, 2004.

67. Skolnick J, Kihara D, Defrosting the frozen approximation: PROSPECTOR–a new approach to threading, *Proteins* **42**(3):319–331, 2001.
68. Skolnick J, Kihara D, Zhang Y, Development and large scale benchmark testing of the PROSPECTOR_3 threading algorithm, *Proteins* **56**(3):502–518, 2004.
69. Zhou H, Zhou Y, Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition, *Proteins* **55**(4):1005–1013, 2004.
70. Ginalski K, Pas J, Wyrwicz LS, von GM, Bujnicki JM, Rychlewski L, ORFeus: Detection of distant homology using sequence profiles and predicted secondary structure, *Nucl Acids Res* **31**(13):3804–3807, 2003.
71. Shi J, Blundell TL, Mizuguchi K, FUGUE: Sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties, *J Mol Biol* **310**(1):243–257, 2001.
72. Sanchez R, Sali A, Large-scale protein structure modeling of the *Saccharomyces cerevisiae* genome, *Proc Natl Acad Sci USA* **95**(23):13597–13602, 1998.
73. Rost B, Twilight zone of protein sequence alignments, *Protein Eng* **12**(2):85–94, 1999.
74. Gille C, Goede A, Preissner R, Rother K, Frommel C, Conservation of substructures in proteins: Interfaces of secondary structural elements in proteasomal subunits, *J Mol Biol* **299**(4):1147–1154, 2000.
75. 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* **297**(1):233–249, 2000.
76. Murzin AG, Brenner SE, Hubbard T, Chothia C, SCOP: A structural classification of proteins database for the investigation of sequences and structures, *J Mol Biol* **247**(4):536–540, 1995.
77. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM, CATH–a hierarchic classification of protein domain structures, *Structure* **5**(8):1093–1108, 1997.
78. Pearl FM, Lee D, Bray JE, Sillitoe I, Todd AE, Harrison AP, Thornton JM, Orengo CA, Assigning genomic sequences to CATH, *Nucl Acids Res* **28**(1):277–282, 2000.
79. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE, The protein data bank, *Nucl Acids Res* **28**(1):235–242, 2000.
80. Orengo CA, Pearl FM, Bray JE, Todd AE, Martin AC, Lo CL, Thornton JM, The CATH Database provides insights into protein structure/function relationships, *Nucl Acids Res* **27**(1):275–279, 1999.
81. Martin AC, Orengo CA, Hutchinson EG, Jones S, Karmirantzou M, Laskowski RA, Mitchell JB, Taroni C, Thornton JM, Protein folds and functions, *Structure* **6**(7):875–884, 1998.
82. Pearl F, Todd AE, Bray JE, Martin AC, Salamov AA, Suwa M, Swindells MB, Thornton JM, Orengo CA, Using the CATH domain database to assign structures and functions to the genome sequences, *Biochem Soc Trans* **28**(2):269–275, 2000.
83. Fetrow JS, Siew N, Di Gennaro JA, Martinez-Yamout M, Dyson HJ, Skolnick J, Genomic-scale comparison of sequence- and structure-based methods of function prediction: Does structure provide additional insight?, *Protein Sci* **10**(5):1005–1014, 2001.
84. Wallace AC, Laskowski RA, Thornton JM, Derivation of 3D coordinate templates for searching structural databases: Application to Ser-His-Asp catalytic triads in the serine proteinases and lipases, *Protein Sci* **5**(6):1001–1013, 1996.

85. Laskowski RA, Watson JD, Thornton JM, Protein function prediction using local 3D templates, *J Mol Biol* **351**(3):614–626, 2005.
86. Porter CT, Bartlett GJ, Thornton JM, The catalytic site atlas: A resource of catalytic sites and residues identified in enzymes using structural data, *Nucl Acids Res* **32**(Database issue):D129–D133, 2004.
87. Levitt DG, Banaszak LJ, POCKET: A computer graphics method for identifying and displaying protein cavities and their surrounding amino acids, *J Mol Graph* **10**(4):229–234, 1992.
88. Liang J, Edelsbrunner H, Woodward C, Anatomy of protein pockets and cavities: Measurement of binding site geometry and implications for ligand design, *Protein Sci* **7**(9):1884–1897, 1998.
89. Hendlich M, Rippmann F, Barnickel G, LIGSITE: Automatic and efficient detection of potential small molecule-binding sites in proteins, *J Mol Graph Model* **15**(6):359–363, 389, 1997.
90. Binkowski TA, Naghibzadeh S, Liang J, CASTp: Computed atlas of surface topography of proteins, *Nucl Acids Res* **31**(13):3352–3355, 2003.
91. Kinoshita K, Furui J, Nakamura H, Identification of protein functions from a molecular surface database, eF-site, *J Struct Funct Genomics* **2**(1):9–22, 2002.
92. Laskowski RA, Watson JD, Thornton JM, ProFunc: A server for predicting protein function from 3D structure, *Nucl Acids Res* **33**(Web Server issue):W89–W93, 2005.
93. Laskowski RA, Watson JD, Thornton JM, Protein function prediction using local 3D templates, *J Mol Biol* **351**(3):614–626, 2005.
94. Laurie AT, Jackson RM, Q-SiteFinder: An energy-based method for the prediction of protein-ligand binding sites, *Bioinformatics* **21**(9):1908–1916, 2005.
95. Liang MP, Banatao DR, Klein TE, Brutlag DL, Altman RB, WebFEATURE: An interactive web tool for identifying and visualizing functional sites on macromolecular structures, *Nucl Acids Res* **31**(13):3324–3327, 2003.
96. Wei L, Altman RB, Recognizing protein binding sites using statistical descriptions of their 3D environments, *Pac Symp Biocomput* 497–508, 1998.
97. Jacob F, Perrin D, Sanchez C, Monod J, Operon: A group of genes with the expression coordinated by an operator, *C R Hebd Seances Acad Sci* **250**:1727–1729, 1960.
98. Overbeek R, Fonstein M, D 'Souza M, Pusch GD, Maltsev N, The use of gene clusters to infer functional coupling, *Proc Natl Acad Sci USA* **96**(6):2896–2901, 1999.
99. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA, The COG database: An updated version includes eukaryotes, *BMC Bioinform* **4**(1):41–41, 2003.
100. Tatusov RL, Koonin EV, Lipman DJ, A genomic perspective on protein families, *Science* **278**(5338):631–637, 1997.
101. Uchiyama I, MBGD: Microbial genome database for comparative analysis, *Nucl Acids Res* **31**(1):58–62, 2003.
102. Dandekar T, Snel B, Huynen M, Bork P, Conservation of gene order: A fingerprint of proteins that physically interact, *Trends Biochem Sci* **23**(9):324–328, 1998.
103. Kihara D, Shimizu T, Kanehisa M, Prediction of membrane proteins based on classification of transmembrane segments, *Protein Eng* **11**(11):961–970, 1998.
104. Kihara D, Kanehisa M, Tandem clusters of membrane proteins in complete genome sequences, *Genome Res* **10**(6):731–743, 2000.
105. Kihara D, Kanehisa M, Prediction of membrane proteins in post-genomic era, *Recent Res Develop Protein Eng* **1**:179–196, 2001.

106. Yanai I, Derti A, DeLisi C, Genes linked by fusion events are generally of the same functional category: A systematic analysis of 30 microbial genomes, *Proc Natl Acad Sci USA* **98**(14):7940–7945, 2001.
107. Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D, Detecting protein function and protein–protein interactions from genome sequences, *Science* **285**(5428):751–753, 1999.
108. 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* **96**(8):4285–4288, 1999.
109. Korbel JO, Jensen LJ, von MC, Bork P, Analysis of genomic context: Prediction of functional associations from conserved bidirectionally transcribed gene pairs, *Nat Biotechnol* **22**(7):911–917, 2004.
110. von Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P, Snel B, STRING: A database of predicted functional associations between proteins, *Nucl Acids Res* **31**(1):258–261, 2003.
111. PubMed, <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed&itool=toolbar>, cited January 1, 2005.
112. Marcotte EM, Xenarios I, van Der Bliek AM, Eisenberg D, Localizing proteins in the cell from their phylogenetic profiles, *Proc Natl Acad Sci USA* **97**(22):12115–12120, 2000.
113. Huynen M, Snel B, Lathe W, III Bork P, Predicting protein function by genomic context: Quantitative evaluation and qualitative inferences, *Genome Res* **10**(8):1204–1210, 2000.
114. GenoBase, <http://ecoli.aist-nara.ac.jp/>, cited January 12, 2005.
115. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M, Fields S, Rothberg JM, A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*, *Nature* **403**(6770):623–627, 2000.
116. Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E, Vijayadamodar G, Pochart P, Machineni H, Welsh M, Kong Y, Zerhusen B, Malcolm R, Varrone Z, Collis A, Minto M, Burgess S, McDaniel L, Stimpson E, Spriggs F, Williams J, Neurath K, Ioime N, Agee M, Voss E, Furtak K, Renzulli R, Aanensen N, Carrolla S, Bickelhaupt E, Lazovatsky Y, DaSilva A, Zhong J, Stanyon CA, Finley RL, Jr., White KP, Braverman M, Jarvie T, Gold S, Leach M, Knight J, Shimkets RA, McKenna MP, Chant J, Rothberg JM, A protein interaction map of *Drosophila melanogaster*, *Science* **302**(5651):1727–1736, 2003.
117. Arenkov P, Kukhtin A, Gemmell A, Voloshchuk S, Chupeeva V, Mirzabekov A, Protein microchips: Use for immunoassay and enzymatic reactions, *Anal Biochem* **278**(2):123–131, 2000.
118. Jacq B, Protein function from the perspective of molecular interactions and genetic networks, *Brief Bioinform* **2**(1):38–50, 2001.
119. Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D, A combined algorithm for genome-wide prediction of protein function, *Nature* **402**(6757):83–86, 1999.
120. Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM, Eisenberg D, DIP, the database of interacting proteins: A research tool for studying cellular networks of protein interactions, *Nucl Acids Res* **30**(1):303–305, 2002.
121. Bader GD, Hogue CW, BIND–a data specification for storing and describing biomolecular interactions, molecular complexes and pathways, *Bioinformatics* **16**(5):465–477, 2000.

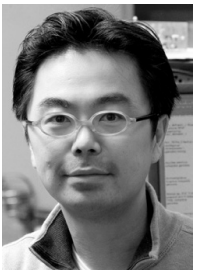
122. Bader GD, Donaldson I, Wolting C, Ouellette BF, Pawson T, Hogue CW, BIND–The biomolecular interaction network database, *Nucl Acids Res* **29**(1):242–245, 2001.
123. Bader GD, Betel D, Hogue CW, BIND: The biomolecular interaction network database, *Nucl Acids Res* **31**(1):248–250, 2003.
124. Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M, Cesareni G, MINT: A molecular interaction database, *FEBS Lett* **513**(1):135–140, 2002.
125. Schwikowski B, Uetz P, Fields S, A network of protein–protein interactions in yeast, *Nat Biotechnol* **18**(12):1257–1261, 2000.
126. Brun C, Chevenet F, Martin D, Wojcik J, Guenoche A, Jacq B, Functional classification of proteins for the prediction of cellular function from a protein–protein interaction network, *Genome Biol* **5**(1):R6–R6, 2003.
127. Samanta MP, Liang S, Predicting protein functions from redundancies in large-scale protein interaction networks, *Proc Natl Acad Sci USA* **100**(22):12579–12583, 2003.
128. Vazquez A, Flammini A, Maritan A, Vespignani A, Global protein function prediction from protein–protein interaction networks, *Nat Biotechnol* **21**(6):697–700, 2003.
129. Bader GD, Hogue CW, An automated method for finding molecular complexes in large protein interaction networks, *BMC Bioinform* **4**(1):2–2, 2003.
130. Rougemont J, Hingamp P, DNA microarray data and contextual analysis of correlation graphs, *BMC Bioinform* **4**(1):15–15, 2003.
131. Brun C, Herrmann C, Guenoche A, Clustering proteins from interaction networks for the prediction of cellular functions, *BMC Bioinform* **5**(1):95, 2004.
132. Chua HN, Sung WK, Wong L, Exploiting indirect neighbours and topological weight to predict protein function from protein–protein interactions, *Bioinform* **22**(13):1623–1630, 2006.
133. Walhout AJ, Sordella R, Lu X, Hartley JL, Temple GF, Brasch MA, Thierry-Mieg N, Vidal M, Protein interaction mapping in *C. elegans* using proteins involved in vulval development, *Science* **287**(5450):116–122, 2000.
134. Yu H, Luscombe NM, Lu HX, Zhu X, Xia Y, Han JD, Bertin N, Chung S, Vidal M, Gerstein M, Annotation transfer between genomes: Protein–protein interologs and protein-DNA regulogs, *Genome Res* **14**(6):1107–1118, 2004.
135. Tong AH, Lesage G, Bader GD, Ding H, Xu H, Xin X, Young J, Berriz GF, Brost RL, Chang M, Chen Y, Cheng X, Chua G, Friesen H, Goldberg DS, Haynes J, Humphries C, He G, Hussein S, Ke L, Krogan N, Li Z, Levinson JN, Lu H, Menard P, Munyana C, Parsons AB, Ryan O, Tonikian R, Roberts T, Sdicu AM, Shapiro J, Sheikh B, Suter B, Wong SL, Zhang LV, Zhu H, Burd CG, Munro S, Sander C, Rine J, Greenblatt J, Peter M, Bretscher A, Bell G, Roth FP, Brown GW, Andrews B, Bussey H, Boone C, Global mapping of the yeast genetic interaction network, *Science* **303**(5659):808–813, 2004.
136. Tong AH, Evangelista M, Parsons AB, Xu H, Bader GD, Page N, Robinson M, Raghibizadeh S, Hogue CW, Bussey H, Andrews B, Tyers M, Boone C, Systematic genetic analysis with ordered arrays of yeast deletion mutants, *Science* **294**(5550):2364–2368, 2001.
137. Fraser AG, Kamath RS, Zipperlen P, Martinez-Campos M, Sohrmann M, Ahringer J, Functional genomic analysis of *C. elegans* chromosome I by systematic RNA interference, *Nature* **408**(6810):325–330, 2000.
138. Tewari M, Hu PJ, Ahn JS, yivi-Guedehoussou N, Vidalain PO, Li S, Milstein S, Armstrong CM, Boxem M, Butler MD, Busiguina S, Rual JF, Ibarrola N, Chaklos ST, Bertin N, Vaglio P, Edgley ML, King KV, Albert PS, Vandenhaute J, Pandey A, Riddle DL, Ruvkun G, Vidal M, Systematic interactome mapping and

genetic perturbation analysis of a *C. elegans* TGF-beta signaling network, *Mol. Cell* **13**(4):469–482, 2004.
139. Winzeler EA, Shoemaker DD, Astromoff A, Liang H, Anderson K, Andre B, Bangham R, Benito R, Boeke JD, Bussey H, Chu AM, Connelly C, Davis K, Dietrich F, Dow SW, El BM, Foury F, Friend SH, Gentalen E, Giaever G, Hegemann JH, Jones T, Laub M, Liao H, Liebundguth N, Lockhart DJ, Lucau-Danila A, Lussier M, M 'Rabet N, Menard P, Mittmann M, Pai C, Rebischung C, Revuelta JL, Riles L, Roberts CJ, Ross-MacDonald P, Scherens B, Snyder M, Sookhai-Mahadeo S, Storms RK, Veronneau S, Voet M, Volckaert G, Ward TR, Wysocki R, Yen GS, Yu K, Zimmermann K, Philippsen P, Johnston M, Davis RW, Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis, *Science* **285**(5429):901–906, 1999.
140. Drees BL, Thorsson V, Carter GW, Rives AW, Raymond MZ, vila-Campillo I, Shannon P, Galitski T, Derivation of genetic interaction networks from quantitative phenotype data, *Genome Biol* **6**(4):R38 2005.
141. Schena M, Shalon D, Davis RW, Brown PO, Quantitative monitoring of gene expression patterns with a complementary DNA microarray, *Science* **270**(5235):467–470, 1995.
142. DeRisi J, Penland L, Brown PO, Bittner ML, Meltzer PS, Ray M, Chen Y, Su YA, Trent JM, Use of a cDNA microarray to analyse gene expression patterns in human cancer, *Nat Genet* **14**(4):457–460, 1996.
143. Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H, Brown EL, Expression monitoring by hybridization to high-density oligonucleotide arrays, *Nat Biotechnol* **14**(13):1675–1680, 1996.
144. Bodrossy L, Sessitsch A, Oligonucleotide microarrays in microbial diagnostics, *Curr Opin Microbiol* **7**(3):245–254, 2004.
145. Liu ET, Kuznetsov VA, Miller LD, In the pursuit of complexity: Systems medicine in cancer biology, *Cancer Cell* **9**(4):245–247, 2006.
146. Wu X, Dewey TG, From microarray to biological networks: Analysis of gene expression profiles, *Methods Mol Biol* **316**:35–48, 2006.
147. Armstrong NJ, van de Wiel, MA, Microarray data analysis: From hypotheses to conclusions using gene expression data, *Cell Oncol* **26**(5–6):279–290, 2004.
148. Shannon W, Culverhouse R, Duncan J, Analyzing microarray data using cluster analysis, *Pharmacogenomics* **4**(1):41–52, 2003.
149. Slonim DK, From patterns to pathways: Gene expression data analysis comes of age, *Nat Genet* **32**(Suppl):502–508, 2002.
150. Krajewski P, Bocianowski J, Statistical methods for microarray assays, *J Appl Genet* **43**(3):269–278, 2002.
151. Kihara D, Yang YD, Hawkins T, Bioinformatics resources for cancer research with an emphasis on gene function and structure prediction tools, *Cancer Inform* **2**:25–35, 2006.
152. Zeeberg BR, Feng W, Wang G, Wang MD, Fojo AT, Sunshine M, Narasimhan S, Kane DW, Reinhold WC, Lababidi S, Bussey KJ, Riss J, Barrett JC, Weinstein JN, GoMiner: A resource for biological interpretation of genomic and proteomic data, *Genome Biol* **4**(4):R28 2003.
153. Dahlquist KD, Salomonis N, Vranizan K, Lawlor SC, Conklin BR, GenMAPP: A new tool for viewing and analyzing microarray data on biological pathways, *Nat Genet* **31**(1):19–20, 2002.
154. Doniger SW, Salomonis N, Dahlquist KD, Vranizan K, Lawlor SC, Conklin BR, MAPPFinder: Using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data, *Genome Biol* **4**(1):R7, 2003.

155. Barrett T, Suzek TO, Troup DB, Wilhite SE, Ngau WC, Ledoux P, Rudnev D, Lash AE, Fujibuchi W, Edgar R, NCBI GEO: Mining millions of expression profiles–database and tools, *Nucl Acids Res* **33**(Database issue):D562–D566, 2005.
156. Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, Abeygunawardena N, Holloway E, Kapushesky M, Kemmeren P, Lara GG, Oezcimen A, Rocca-Serra P, Sansone SA, ArrayExpress — a public repository for microarray gene expression data at the EBI, *Nucl Acids Res* **31**(1):68–71, 2003.
157. Ball CA, Awad IA, Demeter J, Gollub J, Hebert JM, Hernandez-Boussard T, Jin H, Matese JC, Nitzberg M, Wymore F, Zachariah ZK, Brown PO, Sherlock G, The Stanford Microarray Database accommodates additional microarray platforms and data formats, *Nucl Acids Res* **33**(Database issue):D580–D582, 2005.
158. Zeeberg BR, Qin H, Narasimhan S, Sunshine M, Cao H, Kane DW, Reimers M, Stephens RM, Bryant D, Burt SK, Elnekave E, Hari DM, Wynn TA, Cunningham-Rundles C, Stewart DM, Nelson D, Weinstein JN, High-Throughput GoMiner, an 'industrial-strength' integrative gene ontology tool for interpretation of multiple-microarray experiments, with application to studies of Common Variable Immune Deficiency (CVID), *BMC Bioinform* **6**:168, 2005.
159. Radivojac P, Chawla NV, Dunker AK, Obradovic Z, Classification and knowledge discovery in protein databases, *J Biomed Inform* **37**(4):224–239, 2004.
160. Salwinski L, Eisenberg D, Computational methods of analysis of protein–protein interactions, *Curr Opin Struct Biol* **13**(3):377–382, 2003.
161. von C, Mering, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P, Comparative assessment of large-scale data sets of protein–protein interactions, *Nature* **417**(6887):399–403, 2002.
162. Deane CM, Salwinski L, Xenarios I, Eisenberg D, Protein interactions: Two methods for assessment of the reliability of high throughput observations, *Mol Cell Proteomics* **1**(5):349–356, 2002.
163. Janin J, Seraphin B, Genome-wide studies of protein–protein interaction, *Curr Opin Struct Biol* **13**(3):383–388, 2003.
164. Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E, Vijayadamodar G, Pochart P, Machineni H, Welsh M, Kong Y, Zerhusen B, Malcolm R, Varrone Z, Collis A, Minto M, Burgess S, McDaniel L, Stimpson E, Spriggs F, Williams J, Neurath K, Ioime N, Agee M, Voss E, Furtak K, Renzulli R, Aanensen N, Carrolla S, Bickelhaupt E, Lazovatsky Y, DaSilva A, Zhong J, Stanyon CA, Finley RL, Jr., White KP, Braverman M, Jarvie T, Gold S, Leach M, Knight J, Shimkets RA, McKenna MP, Chant J, Rothberg JM, A protein interaction map of *Drosophila melanogaster*, *Science* **302**(5651):1727–1736, 2003.
165. Lu L, Arakaki AK, Lu H, Skolnick J, Multimeric threading-based prediction of protein–protein interactions on a genomic scale: Application to the *Saccharomyces cerevisiae* proteome, *Genome Res* **13**(6A):1146–1154, 2003.
166. Bader JS, Chaudhuri A, Rothberg JM, Chant J, Gaining confidence in high-throughput protein interaction networks, *Nat Biotechnol* **22**(1):78–85, 2004.
167. Huang TW, Tien AC, Huang WS, Lee YC, Peng CL, Tseng HH, Kao CY, Huang CY, POINT: A database for the prediction of protein–protein interactions based on the orthologous interactome, *Bioinform* 3273–3276, 2004.
168. 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* **100**(14):8348–8353, 2003.
169. Troyanskaya OG, Putting microarrays in a context: Integrated analysis of diverse biological data, *Brief Bioinform* **6**(1):34–43, 2005.

170. Green ML, Karp PD, A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases, *BMC Bioinform* **5**(1):76, 2004.
171. Karp PD, Paley S, Romero P, The pathway tools software, *Bioinform* **18**(Suppl 1): S225–S232, 2002.
172. Deng M, Chen T, Sun F, An integrated probabilistic model for functional prediction of proteins, *J Comput Biol* **11**(2–3):463–475, 2004.
173. Mewes HW, Amid C, Arnold R, Frishman D, Guldener U, Mannhaupt G, Munsterkotter M, Pagel P, Strack N, Stumpflen V, Warfsmann J, Ruepp A, MIPS: Analysis and annotation of proteins from whole genomes, *Nucl Acids Res* **32**(Database issue):D41–D44, 2004.
174. Deng M, Zhang K, Mehta S, Chen T, Sun F, Prediction of protein function using protein–protein interaction data, *J Comput Biol* **10**(6):947–960, 2003.
175. Morett E, Korbel JO, Rajan E, Saab-Rincon G, Olvera L, Olvera M, Schmidt S, Snel B, Bork P, Systematic discovery of analogous enzymes in thiamin biosynthesis, *Nat Biotechnol* **21**(7):790–795, 2003.
176. Huynen MA, Snel B, Bork P, Gibson TJ, The phylogenetic distribution of frataxin indicates a role in iron-sulfur cluster protein assembly, *Hum Mol Genet* **10**(21):2463–2468, 2001.
177. Lord PW, Stevens RD, Brass A, Goble CA, Investigating semantic similarity measures across the gene ontology: The relationship between sequence and annotation, *Bioinform* **19**(10):1275–1283, 2003.
178. Lord PW, Stevens RD, Brass A, Goble CA, Semantic similarity measures as tools for exploring the gene ontology, *Pac Symp Biocomput* 601–612, 2003.
179. Friedberg I, Jambon M, Godzik A, New avenues in protein function prediction, *Protein Sci* **15**(6):1527–1529, 2006.
180. Pellegrini-Calace M, Soro S, Tramontano A, Revisiting the prediction of protein function at CASP6, *FEBS J* **273**(13):2977–2983, 2006.
181. Soro S, Tramontano A, The prediction of protein function at CASP6, *Proteins* **61**(Suppl 7):201–213, 2005.
182. Rost B, Liu J, Nair R, Wrzeszczynski KO, Ofran Y, Automatic prediction of protein function, *Cell Mol Life Sci* **60**(12):2637–2650, 2003.

Troy Hawkins is a Ph.D. candidate in the Department of Biological Sciences, Purdue University, Indiana, USA. His work involves automated methods of predicting protein function annotations and functional motifs. He is a member of The International Society of Computational Biology and the American Chemical Society, a participant in the AAAS/Science Program for Excellence in Science and the recipient of an NIH Training Grant in Molecular Biophysics.



Daisuke Kihara received his Ph.D. degree from Kyoto University, Japan, in 1999. From 1999 to 2003, he was with Dr. Jeffrey Skolnick as a postdoctoral researcher in the Danforth Plant Science Center, St. Louis, Missouri, and SUNY Buffalo, New York, USA. In 2003 he joined Purdue University as an assistant professor in the Department of Biological Sciences and the Department of Computer Science. His research projects include protein function prediction from sequence, protein structure prediction, and protein surface shape search for function prediction and docking. He is a member of The Protein Society, The Biophysical Society, The International Society of Computational Biology, The Biophysical Society of Japan, The Molecular Biology Society of Japan, and Japanese Society of Bioinformatics.