Technical Report

Department of Computer Science
and Engineering
University of Minnesota
4-192 EECS Building
200 Union Street SE
Minneapolis, MN 55455-0159 USA

TR 06-028

Computational Approaches for Protein Function Prediction: A Survey

Gaurav Pandey, Vipin Kumar and Michael Steinbach

October 31, 2006

Computational Approaches for Protein Function Prediction: A Survey

Gaurav Pandey, Vipin Kumar and Michael Steinbach
Department of Computer Science and Engineering, University of Minnesota
{gaurav,kumar,steinbac}@cs.umn.edu

Proteins are the most essential and versatile macromolecules of life, and the knowledge of their functions is a crucial link in the development of new drugs, better crops, and even the development of synthetic biochemicals such as biofuels. Experimental procedures for protein function prediction are inherently low throughput and are thus unable to annotate a non-trivial fraction of proteins that are becoming available due to rapid advances in genome sequencing technology. This has motivated the development of computational techniques that utilize a variety of high-throughput experimental data for protein function prediction, such as protein and genome sequences, gene expression data, protein interaction networks and phylogenetic profiles. Indeed, in a short period of a decade, several hundred articles have been published on this topic. This survey aims to discuss this wide spectrum of approaches by categorizing them in terms of the data type they use for predicting function, and thus identify the trends and needs of this very important field. The survey is expected to be useful for computational biologists and bioinformaticians aiming to get an overview of the field of computational function prediction, and identify areas that can benefit from further research.

Categories and Subject Descriptors: J.3 [**Life and Medical Sciences**]: Bioinformatics

Key Words and Phrases: Protein function prediction, bioinformatics, Gene Ontology, multiple biological data types, high-throughput experimental data, data mining, non-homology based methods

Contents

Authors' Address: Department of Computer Science and Engineering, University of Minnesota, 4-192 EE/CS Building, 200 Union Street SE, Minneapolis, MN 55414, USA
Support: This material is based upon work supported by the National Science Foundation under Grant Nos. IIS-0308264 and ITR-0325949. Access to computing facilities was provided by the Minnesota Supercomputing Institute.

1. INTRODUCTION

Proteins are macromolecules that serve as building blocks and functional components of a cell, and account for the second largest fraction of the cellular weight after water. Proteins are responsible for some of the most important functions in an organism, such as constitution of the organs (structural proteins), the catalysis of biochemical reactions necessary for metabolism (enzymes), and the maintenance of the cellular environment (transmembrane proteins). Thus, proteins are the most essential and versatile macromolecules of life, and the knowledge of their functions is a crucial link in the development of new drugs, better crops, and even the development of synthetic biochemicals such as biofuels.

The early approaches to predicting protein function were experimental and usually focused on a specific target gene or protein, or a small set of proteins forming natural groups such as protein complexes. These approaches included gene knockout, targeted mutations and the inhibition of gene expression [Weaver 2002]. However, irrespective of the details, these approaches are low-throughput because of the huge experimental and human effort required in analyzing a single gene or protein. As a result, even large-scale experimental annotation initiatives, such as the EUROFAN project [Oliver 1996], are inadequate for annotating a non-trivial fraction of the proteins that are becoming available due to rapid advances in genome sequencing technology. This has resulted in a continually expanding sequence-function gap for the discovered proteins [Roberts 2004].

In an attempt to close this gap, numerous high-throughput experimental procedures have been invented to investigate the mechanisms leading to the accomplishment of a protein's function. These procedures have generated a wide variety of useful data that ranges from simple protein sequences to complex high-throughput data, such as gene expression data sets and protein interaction networks. These data offer different types of insights into a protein's function and related concepts. For instance, protein interaction data shows which proteins come together to perform a particular function, while the three-dimensional

structure of a protein determines the precise sites to which the interacting protein binds itself. Furthermore, recent years have seen the recording of this data in very standardized and professionally maintained databases such as SWISS-PROT [Boeckmann et al. 2003], MIPS [Mewes et al. 2002], DIP [Xenarios et al. 2002] and PDB [Berman et al. 2000].

The huge amount of data that has accumulated over the years has made biological discovery via manual analysis tedious and cumbersome. This has, in turn, necessitated the use of techniques from the field of bioinformatics, an approach that is crucial in today's age of rapid generation and warehousing of biological data. Bioinformatics focuses on the utilization of techniques from computer science and the development of novel computational approaches for addressing problems in molecular biology and associated disciplines. Indeed, a more recently advocated path for biological research is the creation of hypotheses by generating results from an appropriate bioinformatics algorithm in order to narrow the search space, and the subsequent validation of these hypotheses to reach the final conclusion [Rastan and Beeley 1997; Roberts 2004]. Standard sequence comparison tools such as BLAST [Altschul et al. 1990; Altschul et al. 1997], and databases such as PROSITE [Hulo et al. 2006], Pfam [Sonnhammer et al. 1997] and PRINTS [Attwood et al. 2003] serve as testimonials to the benefits that bioinformatics can provide to molecular biology.

Following the success of computational approaches in solving important problems such as sequence alignment and comparison [Altschul et al. 1997], and genome fragment assembly [Shendure et al. 2004], and given the importance of protein function, numerous computational techniques have also been proposed for predicting protein function. Early approaches used sequence similarity tools such as BLAST [Altschul et al. 1990] to transfer functional annotation from the most similar proteins. Subsequently, several other approaches have been proposed that utilize other types of biological data for computational protein function prediction, such as gene expression data, protein interaction networks and phylogenetic profiles. Indeed, in a short period of a decade, several hundred articles have been published on this topic, including several survey articles that try to provide overviews of different subsets of works at different time points.

According to Hodgman [2000], there were four distinct stages of the growth of this field, namely pairwise sequence matching using BLAST [Altschul et al. 1990], the use of sequence signatures such as motifs, single sequence analysis using data mining approaches, and finally, genome-scale sequence analysis. Rost et al. [2003] analyzed the pros and cons of exploiting biologically important signals, such as sequence homology, subcellular localization, post-translational modifications and protein-protein interactions, for protein function prediction. They also stress the importance of establishing standardized databases such as DIP (Database of Interacting Proteins) [Xenarios et al. 2002], and applying data mining techniques to extract useful information from these databases. Seshasayee and Babu [2005] present a more comprehensive survey of direct function prediction techniques. In this article, the authors discuss the most well-known techniques exploiting genomic and large-scale experimental data, such as protein-protein interaction networks, transcriptional regulatory networks, and gene co-expression networks. They also discuss the earliest approaches that proposed an integration of multiple data types, such as [Marcotte et al. 1999]. Thus, the overall focus of this article is on reviewing approaches that exploit the context information available about a protein. Finally, there has been a string of surveys of the field of functional genomics, which involves the use of genome-wide information for predicting the function and functional associations of proteins [Bork et al. 1998; Teichmann and

Mitchison 2000; Marcotte et al. 2000; Eisenberg et al. 2000; Gabaldon and Huynen 2004; Marcotte 2004]. In addition to discussing the most popular genome-based function prediction techniques (Section 5), these articles also motivate the use of novel representations of the genomic information, such as genome-wide protein functional networks, and biologically relevant features of genome sequences, such as nucleotide frequencies and repeats and regulatory regions, for function prediction. Also, some recent surveys have focused on the functions of more specific types of proteins, such as mitochondrial proteins [Gabaldon 2006] and proteins involved in cancer [Hu et al. 2007].

The early experience in the use of computational techniques to predict protein function from different types of biological data has been encouraging. However, although most of the approaches developed so far highlight the potential of computational techniques for protein function prediction, there have been several real successful cases of functional inference in which the interactions or functions predicted by various computational techniques were verified through experimental work. Table I presents such well-known success cases for the gene fusion, gene neighborhood and phylogenetic profile (PP) approaches (Section 5).

Technique	**Protein/Family of Interest**	**Function and/or Interacting Protein**	**Reference**
Fusion	CHORD-containing proteins	Sgt1 (Disease signaling proteins)	[Shirasu et al. 1999]
	Pur2	Pur3, Purine biosynthesis	[Marcotte et al. 1999]
Neighborhood	Cadherin proteins	Cell adhesion	[Wu and Maniatis 1999]
	Human methylmalonyl-CoA	Racemase enzyme	[Bobik and Rasche 2000]
Phylo. Profile	SmpB family	Protein synthesis	[Karzai et al. 1999]
	Frataxin	Iron-sulfur cluster assembly	[Huynen et al. 2001]

Table I. Success stories of popular genomics techniques for function prediction

Given the wide variety of computational techniques that have been proposed for protein function prediction, it can be very hard to keep track of the field and identify its strengths, weaknesses and needs. Responding to this need, we undertake this survey to provide an extensive overview of the field of protein function prediction. In particular, following are the goals and contributions of our survey:

(1) To provide a broad coverage of the field of computational prediction of protein function using multiple types of biological data. Many of the approaches covered have shown promising results on test data sets, and the results of other approaches are expected to improve with further enhancements.

(2) To highlight the inter-relationships between different types of biological data, and to illustrate how ideas used for the analysis of one type of data could influence those used for the analysis of other data types. For instance, in order to realize the full potential of the available genome sequence data, it would be beneficial to be informed of the novel ideas behind approaches using protein sequence data as well. Similarly, it may be very useful to combine multiple data types and analyze them collectively, rather than analyzing them individually. Indeed, very promising results have been obtained by approaches that have implemented this idea.

(3) To identify the open problems and pressing needs of the field. As will be seen, most of the approaches in this field are ad-hoc in nature, and have several limitations, such as

their applicability to only specific subsets of proteins and/or functional classes. Thus, several conceptual and data issues have to be addressed in order to come up with more complete approaches for the function prediction problem.

(4) To illustrate the potential of data mining and machine learning techniques for addressing the problem of protein function prediction by learning from a wide variety of noisy data. Indeed, the best results in the field have been achieved by approaches based on intelligent learning and prediction techniques.

We believe that this survey will be useful to both computational and experimental biologists working with a wide variety of biological data. The survey contains a section for each of the main types of biological data, as well as their combinations, that have been used for predicting protein function. These data types are as shown below:

(1) Amino acid sequences (Section 3)
(2) Protein structure (Section 4)
(3) Genome sequences (Section 5)
(4) Phylogenetic data (Section 6)
(5) Microarray expression data (Section 7)
(6) Protein interaction networks and protein complexes (Section 8)
(7) Biomedical literature (Section 9)
(8) Combination of multiple data types (Section 10)

However, before we proceed with the discussion of these approaches, it is important to make a couple of notes about protein function. First, protein function is an elusive concept and there has been considerable debate in molecular biology about its definition. Hence, Section 2 includes a detailed discussion of various perspectives on this concept and how different functional schemes embody these perspectives. Second, we do not distinguish gene function, and refer to both of them as protein function. Technically, the true function of a gene is to encode one or more proteins that actually perform the function [Gericke 2005]. However, since it is easier to perform experiments at the genetic level, many times the function of its products are taken as the function of the gene itself. Thus, we do not distinguish gene function from protein function, and refer to both of them as the latter.

2. WHAT IS PROTEIN FUNCTION?

The concept of protein function is highly context-sensitive and not very well-defined. In fact, this concept typically acts as an umbrella term for all types of activities that a protein is involved in, be it cellular, molecular or physiological. One such categorization of the types of functions a protein can perform has been suggested by Bork et al. [1998]:

(1) **Molecular function**: The biochemical functions performed by a protein, such as ligand binding, catalysis of biochemical reactions and conformational changes.
(2) **Cellular function**: Many proteins come together to perform complex physiological functions, such as operation of metabolic pathways and signal transduction, to keep the various components of the organism working well.
(3) **Phenotypic function**: The integration of the physiological subsystems, consisting of various proteins performing their cellular functions, and the interaction of this integrated system with environmental stimuli determines the phenotypic properties and behavior of the organism.

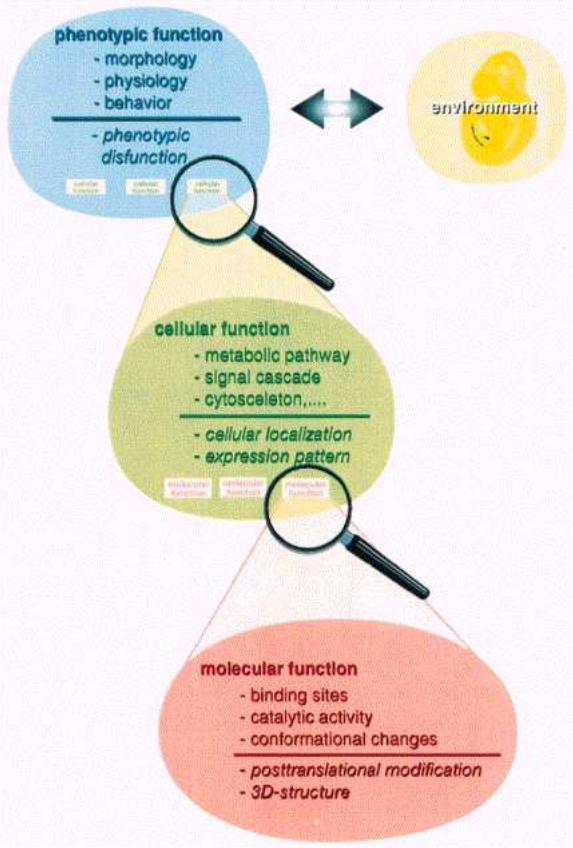


Fig. 1. A possible hierarchical organization of the categories of protein function (taken from [Bork et al. 1998])

Clearly, these three categories are not independent, but rather are hierarchically related as shown in Figure 1. Also, this is not the only categorization that has been proposed. For instance, the Gene Ontology classification scheme categorizes protein function into cellular component, molecular function and biological process [Ashburner et al. 2000]. Confronted with such a variety of formalizations for the concept, we chose to follow the following definition proposed by Rost et al. [2003] in this survey: *function is everything that happens to or through a protein*. In fact, we extend this definition by considering functional relationships and modules as forms of information about the function of a protein as well.

2.1 Functional Classification Schemes

From the above discussion, protein function appears to be a very subjective concept, and different researchers may denote the functions of proteins differently. The first approach to this naming may be to assign natural language labels to proteins, as and when their function is determined. Indeed, this is the case, but such a naming convention sometimes leads to highly atypical labels such as *Yippee* and *Starry Night* [Lan et al. 2002].

Clearly, such as naming system is not amenable to analysis by a human, much less a

computer, because of its large variability. Thus, the need for a standardized functional labeling scheme was paramount, and several groups responded to this need with very innovative proposals. Before discussing these proposed schemes, it is of merit to list some desirable properties of such schemes [Riley 1998; Rison et al. 2000; Ouzounis et al. 2003].

(1) **Wide coverage**: This is the most important property, since any functional scheme should cover as many of the functional phenomena across as many the organisms as possible.

(2) **Standardized format**: Having minimal variability in the functional labels and adopting a standard data structure for the scheme makes the scheme easily readable by computer programs and signficantly enhances their impact.

(3) **Hierarchical structure**: As was seen [Bork et al. 1998], the possible functions do not form a flat list, but are instead arranged hierarchically at a conceptual level. Functional classes range from specific functions to very general functional categories, thus allowing a researcher to choose the appropriate level(s) for his analysis.

(4) **Disjoint categories**: Functions can be of different types, such as cellular component, molecular function and biological process. Thus, a separate hierarchy should be constructed for each type, with no links between them. This allows the choice of the appropriate type of function to be studied.

(5) **Multiple functions**: In order to model the biological possibility of a protein being involved in multiple biological processes depending on the context, it is necessary for a functional scheme to allow the labeling of a single protein with multiple functions.

(6) **Dynamic nature**: Last but not the least, the scheme should not be static, but should be modified as and when new functional knowledge is discovered.

As mentioned, several functional schemes have been proposed to address these issues, with each being successful to some extent, and each having a different scope. The earliest systematic scheme proposed in this arena was the Enzyme Classification (EC) proposed by the International Union of Biochemistry and Molecular Biology [Webb 1992]. This scheme divides the class of enzymes, which are essential proteins responsible for the catalysis of metabolic reactions, into six classes based on their chemical composition. These classes are then further subdivided into three hierarchical levels that further specify the precise reaction a particular enzyme is involved in. However, this scheme had a limited scope, since it was was essentially a classification of reactions and not properties of various catalyst enzymes [Riley 1998].

Subsequent to EC, many functional schemes were proposed for a wider class of proteins. Ouzounis et al. [2003] and Rison et al. [2000] present excellent reviews of some of these schemes, listed in Table II. Many of these schemes, such as EcoCyc [Keseler et al. 2005] and SubtiList [Moszer et al. 2002], were originally designed for specific organisms, in order to study the properties of their genomes and the consituent genes. However, they were subsequently generalized and became more widely applicable. The most popular of these functional schemes are those which were not designed for any specific organism, but were based on general biological phenomena taking place in a wide variety of organisms, both eukaryotes. MIPS/PEDANT [Mewes et al. 2002] (now FunCat [Ruepp et al. 2004]) is currently one of the most popular scheme for the validation of function prediction techniques because of its wide coverage and a standardized hierarchical structure. However, the Gene Ontology (GO) [Ashburner et al. 2000; GO Consortium 2006] is a recently proposed

Scheme	Reference	Scope
EcoCyc	[Keseler et al. 2005]	*E. coli* genes
TIGRFAM	[Haft et al. 2003]	Complete genomes
SubtiList	[Moszer et al. 2002]	*B. subtilis* genes
MIPS/PEDANT	[Mewes et al. 2002]	General
FunCat	[Ruepp et al. 2004]	General
KEGG	[Kanehisa et al. 2004]	Metabolic Pathways
WIT	[Jr et al. 1998]	Metabolic Pathways
Gene Ontology	[GO Consortium 2006]	General

Table II. Popular functional schemes having varying scopes

functional classification system which is based on solid computer science and biological principles and is rapidly being recognized as the most general scheme for functional annotation techniques across a wide variety of biological data [Jensen et al. 2003; Letovsky and Kasif 2003; Hvidsten et al. 2001]. TIGR FAMilieS (TIGRFAMs) [Haft et al. 2003] is another scheme designed for the functional annotation of complete genomes. Overall, almost all of these schemes posess a good subset of above mentioned properties of a global functional classification scheme, and the validation of an approach according to one of them gives a reasonably good estimate of the general applicability, thus alleviating the concern of *overfitting* to a particular labeling scheme.

A very interesting one-of-a-kind quantitative comparison of the first six schemes listed in Table II is reported by Rison et al. [2000]. This is a hard task, since all these schemes were developed almost independently of each other, and thus, it is hard to compare one against the other. Still, Rison et al. [2000] worked out a two-step unification-based strategy for this comparison. In the first step, a combined scheme (CS) is created by manually mapping functional classes up to level three in each of the schemes, and applying filtering techniques to reduce the bias towards any particular scheme. In the second step, a representative subset of each the original schemes was selected by mapping CS back to the scheme. Thus, a representative and comparable version of all the schemes was retrieved. Upon evaluation, it was found that MIPS had the largest overlap with CS, showing that it had the best coverage and generality. This is a quantitative justification for the wide usage of the MIPS functional classification in the protein function prediction literature. Another conclusion of this study was that the overall overlap of all the schemes with CS was high, thus showing that all of them are reasonably similar to each other at the conceptual level. This conclusion is echoed by Ouzounis et al. [2003], who remark that the overlap between functional classification schemes is much higher than that between structural classifications such as SCOP [Andreeva et al. 2004] and CATH [Orengo et al. 2002], though the variability is much higher in the former than in the latter. Thus, these studies provide a justification for the above remarks that the evaluation of a function prediction technique made according to any of these schemes, if conducted correctly, will provide robust results. However, an effort should always be made to use the best available alternative.

Today, any review of functional classification schemes would be incomplete without the discussion of GO and its many desirable properties. These properties have been exhibited by the large number of studies which have used GO for different types of functional classes. A quantitative proof of this popularity is the fact that the GO bibliography[1] currently lists

[1] `http://www.geneontology.org/cgi-bin/biblio.cgi`

1081 publications describing studies following Gene Ontology (as of Oct 19, 2006), which is impressive. Here, we intend to provide a detailed discussion of why the Gene Ontology is the most appropriate scheme for the functional analysis of genes and proteins.

2.2 GO is the Way to Go!

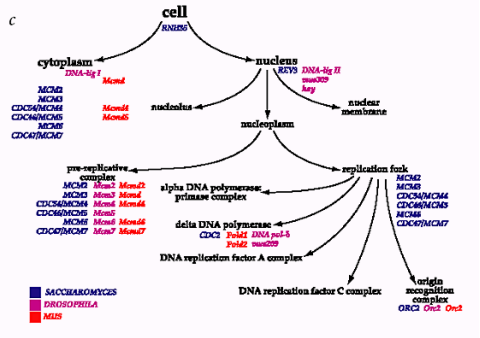
An ontology is defined as a systematic arrangement of all of the important categories of objects and concepts that exist in some field of discourse, together with the relations between them[2]. This concept, which comes originally from philosophy [Smith 2003], is a very effective approach for the organization of the available knowledge in a domain. Owing to these merits, ontologies have found wide applicability in various fields of computer science such as data mining, artificial intelligence, software engineering, electronic commerce and e-commerce [Tan et al. 2005; Davies et al. 2003; Gomez-Perez et al. 2004].

The recognition of the ability to effectively organize knowledge, which is crucial for biology, where the research is highly decentralized, led to the constitution of the Gene Ontology (GO) [GO Consortium 2006]. At the highest level, GO is a functional classification system composed of three disjoint functional ontologies corresponding to *cellular component* (Figure 2(a)), *molecular function* (Figure 2(b)) and *biological process* (Figure 2(c)), each of which addresses a different aspect of a protein's function, as mentioned earlier [Ashburner et al. 2000]. Each of these ontologies is hierarchically structured and is modeled as a directed acyclic graph (DAG), in which each node corresponds to a functional label and each directed edge correponds to either an *is:a* or a *part:of* relationship. Thus, even though GO seems similar in methodolgy and scope to the other functional schemes such as MIPS and TIGR, there is a fundamental difference which makes GO much more general than the others. Almost all the other schemes were designed to aid the functional annotation of specific genome(s), and were generalized later. However, the designers of GO set out with the goal of creating a common multi-dimensional functional ontology which could be applied irrespective of the genome being considered [Bada et al. 2004], thus ensuring the wide applicability of the scheme. This underlying shift in ideology led to the recognition of Gene Ontology as a radical rethink of gene product functional classification [Rison et al. 2000].

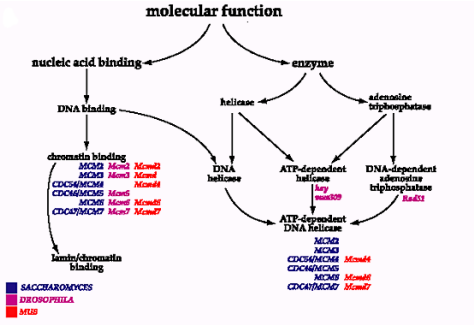
Interestingly, GO posesses all the desirable properties of a functional classification system listed earlier. In fact, its design ideology incorporated all these properties. The following description of how GO satisfies these properties also illustrates its various aspects and provides historical information.

(1) **Wide coverage**: GO was formed by a collaboration of three leading organism-specific genomic databases FlyBase [FlyBase Consortium 2003], SGD [Dwight et al. 2002] and MGI [Blake et al. 2003], who first realized the need to create a cross-species functional classification system in order to solve the data integration problem created by the huge number of indepedent genome projects in the early genome sequencing period [Lewis 2005]. Very soon, other databases, such as TAIR [Huala et al. 2001], also joined the GO consortium, and thus, the coverage of GO became very wide, since biological phenomena occurring in a wide variety of biological systems were considered when adding new labels to the ontology. A proof of this coverage is the large number of genomes, including the human genome, that have been annotated with

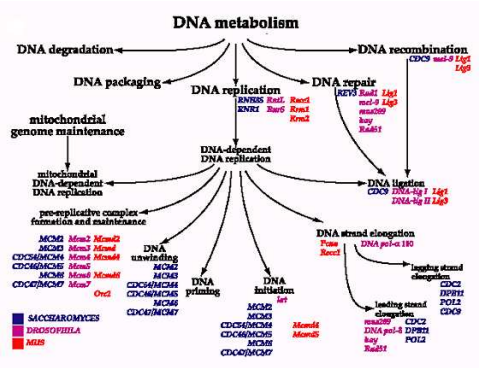
[2]`http://http://www.answers.com/topic/ontology-computer-science`



(a) Cellular component ontology



(b) Molecular function ontology



(c) DNA Metabolism portion of biological process ontology

Fig. 2. Snapshots of the three GO functional ontologies (Figure adapted from [Ashburner et al. 2000])

GO labels [Camon et al. 2003].

(2) **Standardized format**: The ontologies constituting GO are modeled as a generic category of graphs known as directed acyclic graphs (DAGs), which have numerous applications in computer science, such as Bayesian networks and parse trees created by compilers [Edwards 2000]. Each node in these graphs represents a specific functional label and is assigned a unique GO id of the form $GO:XXXXXXX$, and each edge represents either an *is:a* or a *part:of* relationship. This well-defined structure makes GO easily usable by both humans and computers.

(3) **Hierarchical structure**: As shown in Figure 2, all the ontologies in GO are hierarchical in nature. However, they are more complex than other schemes such as MIPS, which model this hierarchy as a tree [Mewes et al. 2002; Ruepp et al. 2004]. The ontologies in GO are modeled as DAGs, which allows a node to have more than one parent. This is biologically appropriate, since a specific function can be a part of more than one higher functions.

(4) **Disjoint categories**: GO is comprised of three disjoint ontologies corresponding to *cellular component* (Figure 2(a)), *molecular function* (Figure 2(b)) and *biological process* (Figure 2(c)), each of which is a different aspect of a protein's function. There does not exist a link between any two of these ontologies, thus satisfying the disjointness condition. This is also in accordance with the multi-dimensional nature of a classification scheme as justified by Riley [1998], in order to treat the different functional aspects of a protein separately, depending on the context of the study.

(5) **Multiple functions**: The structure of GO is inherently multi-dimensional, as discussed above. In addition, within a single ontology, a protein may be labeled with multiple nodes at different levels in hierarchy. The well-defined structure of each hierarchy makes it possible to either extend an annotation to all the ancestors, or subsume them in the opposite direction. In fact, the extension of annotation to all parents is the basis of the validation of several function prediction strategies.

(6) **Dynamic nature**: Last but not the least, GO is an open-source endeavour and has a public interface at the Sourceforge website[3], which acts as the channel for submitting new functional labels and other forms of functional knowledge. These submissions are continuously reviewed by the curators and scientifically correct information is incorporated into the GO database [Bada et al. 2004; GO Consortium 2006].

The above discussion enumerates a detailed list of reasons that made Gene Ontology a success [Bada et al. 2004; Clark et al. 2005]. This success has not been achieved only due to the strong conceptual foundations of GO, but also because its applications in function prediction that have produced great results, both in a quantitative and qualitative sense. Numerous protein function prediction strategies that have used Gene Ontology as a source of functional classes and for the purpose of validation [Jensen et al. 2003; Letovsky and Kasif 2003; Hvidsten et al. 2001], and now the use of these classes is almost a de-facto standard. Also, the rapidly expanding army of easy-to-use tools for manipulating GO, such as the AmiGO browser [GO Consortium 2006], has enhanced the utility of GO for experimental and computational biologists substantially.

The rich knowledge embedded in GO, and its more complex structure as compared to other simpler schemes, has motivated studies that focus on making a better use of this

[3] `http://geneontology.sourceforge.net`

knowledge and structure. Lord et al. [2003; Lord et al. [2003] performed a unique study of GO, where they investigated whether proteins with similar characteristics, such as similar amino acid sequences, are annotated with *similar* functional classes in GO. While the similarity of sequences is reasonably easy to estimate using BLAST, the estimation of similarity between GO classes that are arranged in a DAG-based hierarchy is considerably harder. Thus, for this task, Lord *et al* used several semantic similarity measures, such as that of Lin [1998], that have been used for a similar purpose in other ontologies, such as WordNet [Fellbaum 1998]. Using these measures, they determined that there is a significant correlation between the similarity between the biological characteristics of proteins and the similarity of their GO annotations. Similar results were reported for microarray data [Sevilla et al. 2005]. This is important from a function prediction viewpoint, since now models for a function can utilize data not only from the proteins annotated with the same function, but also from *similar* functions.

A related issue in the use of GO in functional genomics studies is the fact that most of the ontologies in GO contain several thousand functional terms at different levels of specificity. In this situation, a hard task for function prediction studies is to choose which terms should be used to evaluate the efficacy of the proposed prediction method. Myers et al. [2006] have made useful suggestions for this choice, by evaluating terms in the biological process ontology for their relevance to experimental studies. This evaluation is done by obtaining votes from biologists in response to the following questions: "Given a protein p and a function f, if a method M predicts that p performs f in an organism, can this prediction be tested in wet-lab experiments?". This evaluation is very important for the field of protein function prediction, since the real utility of the computational methods in this field lies in making predictions that can be verified experimentally. Thus, although the results of this evaluation are currently available only for yeast [4] and human proteins[5] for the GO biological process ontology, this study marks a significant advance in the use of GO for function prediction studies.

Motivated by these advances in making an effective use of the knowledge-rich but complicated Gene Ontology, several machine learning methods have been proposed for explicitly incorporating the structure of GO into function prediction methods. Following is a list of some of the methods used for this problem:

(1) Bayesian network modeling of the hierarchical DAG structure [Barutcuoglu et al. 2006].

(2) Probabilistic chain graphs for modeling the hierarchical DAG structure [Carroll and Pavlovic 2006].

(3) Incorporation of the semantic similarities between functional classes into standard classification algorithms [Pandey and Kumar 2007; Tao et al. 2007].

In particular, Barutcuoglu et al. [2006] were able to achieve significant improvements in performance over several classes, by augmenting a standard SVM classifier with hierarchical relationships between classes using their Bayesian network. Even better results are expected as more rigorous methods of complementing biological data with the information in the structure and contents of GO are developed.

[4] `http://www.biomedcentral.com/content/supplementary/1471-2164-7-187-s1.txt`
[5] Unpublished

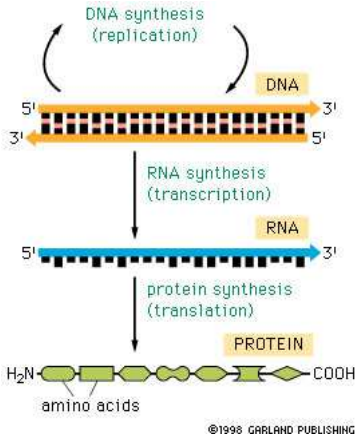


Fig. 3. The central dogma of molecular biology: conversion of gene to protein via mRNA [Alberts et al. 2003]

This discussion makes it clear that the use of Gene Ontology in any function prediction project, whether for validation or for algorithm design, naturally incorporates important biological concepts into the strategy, thus making it more robust and biologically useful, in addition to providing other advantages such as the improvement of coverage and accuracy. This is why we suggest that **GO is the way to go** for the field of protein function prediction.

2.3 Discussion

In the preceding discussions, an attempt was made to more precisely define the meaning of protein function. However, protein function is an umbrella concept that has various aspects such as molecular function, cellular function and phenotypic function, and the most appropriate formalization in general is to treat function as a hierarchical, multi-dimensional concept. This model has been adopted by numerous functional classification schemes, each with their own strengths and weakness. The most important conclusion of this section is the superiority of the Gene Ontology over all the other classification schemes with respect to an extensive list of desirable properties that any biological meaningful scheme should posess. Because of its coverage, generality and biological relevance, it would be beneficial for protein function approaches to incorporate GO into their strategy in some form or the other.

3. PROTEIN SEQUENCES

3.1 Introduction

The central dogma of molecular biology is the conversion of a gene to protein via the transcription and translation phases, as shown in Figure 3. The result of this process is a sequence constructed from twenty amino acids, and is known as the protein's primary structure. This sequence is the most fundamental form of information available about the protein since it determines different characteristics of the protein such as its sub-cellular localization, structure and function.

The most popular experimental method for the identification of protein sequences is mass spectrometry [Sickmann et al. 2003], which, in combination with algorithms such as ProFound [Zhang and Chait 2000], comes in various flavors, such as peptide mass finger-

printing, peptide fragmentation and other comparative methods. However, these methods are low-throughput, and thus, with the exponential generation of genome sequences, the focus has shifted to computational approaches that can identify genes from these genomes. Once a gene has been identified, it is a trivial task to apply the codon-to-amino acid translation code [Weaver 2002] to predict the sequence of the protein encoded by the gene. Some of the most popular tools for this gene identification task in eukaryotic organisms are GenScan [Burge and Karlin 1997] and GeneParser [Snyder and Stormo 1995], which employ hidden Markov models and dynamic programming, respectively, to combine the signals corresponding to the various components of a gene's structure.

Since its amino acid sequence is the most fundamental information available about a protein, such sequences have been accumulated in large numbers in several standardized databases. The most popular of these are the SWISS-PROT and TrEMBL databases [Boeckmann et al. 2003]. SWISS-PROT is a comprehensive, manually curated database that provides a wide variety of information about the constituent proteins, such as their functional annotation, amino acid sequence and other information in the form of keywords and features. TrEMBL (Translated EMBL) is an automatically curated supplement of SWISS-PROT that contains the resultant translations of all nucleotide sequences present in the EMBL/GenBank/DDBJ databases [Brunak et al. 2002], as well as their automated classification and annotation. As of May 2, 2006, the number of entries in SWISS-PROT and TrEMBL were 217551 and 2851442 respectively. Because of the associated confidence in the assigned functional classes, several approaches that employ data from these two databases use SWISS-PROT as the source of the training sequences, while a subset of TrEMBL is used as the test set. Other extensive databases of protein sequences are MIPS [Mewes et al. 2002], PIR [Wu et al. 2003] and IPI [Kersey et al. 2004].

Database	Reference	Organism
SGD	[Dwight et al. 2002]	*S. cerevisiae*
FlyBase	[FlyBase Consortium 2003]	*D. melanogaster*
WormBase	[Harris et al. 2004]	*C. elegans*
TAIR	[Huala et al. 2001]	*A. thaliana*
TubercuList	[Camus et al. 2002]	*M. tuberculosis*
GenProtEC	[Serres et al. 2004]	*E. coli*
EnsEMBL	[Hubbard et al. 2005]	Several mammals

Table III. Organism-specific databases of protein sequences

Database	Reference	Type of proteins
GPCRDB	[Horn et al. 2003]	G-coupled protein receptors
MEROPS	[Rawlings and Barrett 1999]	Peptidases
TCDB	[Saier Jr. 2000]	Transport membrane proteins
LGICdb	[Novre and Changeux 2001]	Ligand-gated ion channels
BRENDA	[Schomburg et al. 2004]	Enzymes
NucleaRDB	[Horn et al. 2001]	Nuclear receptors

Table IV. Type-specific databases of protein sequences

Besides the above general-purpose databases, many groups have created organism- and type-specific databases of protein sequences. Tables III and IV present a list of some

of these databases. Most of these databases also contain functional annotations for the member proteins. Finally, yet another category of databases consists of those that provide functional annotations for genes, such as GOA [Camon et al. 2003] and GenBank [Benson et al. 2004]. This widespread availability of information about and associated with protein sequences has lead to a rapid increase in the use of protein sequences in bioinformatics research.

3.2 Annotation transfer from homologues: How good is it for function prediction?

The first major breakthrough in the field of computational biology was the design of sequence similarity systems such as FASTA [Pearson and Lipman 1988] and BLAST [Altschul et al. 1990] (which was later enhanced into PSI-BLAST [Altschul et al. 1997]). These systems search standard databases such as SWISS-PROT to find proteins homologous to the subject protein, i.e., a similar protein in another organism, using approximate sequence alignment algorithms. In addition, BLAST and PSI-BLAST also produce an E-value for each match S in the database, which denotes the probability of achieving an alignment scores equivalent to or better than S in a database of random sequences of the same size as the target database[6], and can be used as a metric for ranking the search results. This probability is calculated using an extreme value distribution [Gumbel 2004]. An immediate consequence of the development of these systems was a method for the prediction of function of unclassified proteins, namely annotation transfer from homologues. In this method, the functions of the most homologous proteins (the results in a BLAST search with an E-value greater than a pre-specified threshold) are transferred to the protein under consideration. Though early applications of this method produced promising results, subsequent studies discovered several limitations [Gerlt and Babbitt 2000; Devos and Valencia 2000; Whisstock and Lesk 2003].

The most significant factor causing an inconsistency of function between homologues is duplication during evolution, where a duplicate of the original gene adopts a new function in response to selective pressure [Gerlt and Babbitt 2000]. For such genes and their products, annotation transfer by homology produces erroneous results, as has been confirmed by several studies [Gerlt and Babbitt 2000; Whisstock and Lesk 2003].

In order to quantify the early indications that sequence homology is not equivalent to functional identity, some studies were conducted to systematically evaluate the correlation between sequence and functional similarity [Devos and Valencia 2000; Wilson et al. 2000]. Devos and Valencia [2000] evaluated this correlation for four distinct levels of protein function:

—Enzymatic function classification, represented by the enzyme classification (EC) number

—Functional annotations in the form of SWISS-PROT keywords

—Cell functional class

—Conservation of the type of amino acid in the binding site

In addition, the authors of this study also evaluated how sequence homology is correlated with the conservation of three-dimensional protein structure. The structure of a protein is considered closer to its sequence than its function, as discussed in Section 4. Indeed, structural comparison is considered the gold standard in the evaluation of remote sequence

[6] `http://www.ncbi.nlm.nih.gov/blast/tutorial`

homology [Kuang et al. 2005; Rangwala and Karypis 2005], which operates primarily on sequence data.

In this framework, the evaluation on the *E. coli* genome resulted in the following order for the sequence-function correlation

$$Structure \succ EC\ number \succ SwissProt\ Keyword \succ Functional\ class \succ Binding\ site$$

which shows that sequence similarity is more highly correlated with structural than more specific notions of functional similarity. This result is in agreement with those reported by other researchers [Whisstock and Lesk 2003], and thus highlights the limitations of the annotation transfer strategy. However, on the positive side, it suggests a new route for function prediction from sequence, i.e. $sequence \rightarrow structure \rightarrow function$, since the two segments of this route have been reported to have a strong correlation than the whole [Whisstock and Lesk 2003].

Finally, on the computational side of things, the inability of the annotation transfer technique to accurately determine protein function had an indirect effect of database contamination [Devos and Valencia 2000], also sometimes referred to as propagation of error. Since a major fraction of the annotations provided in the sequence databases that are automatically annotated, were derived using this technique, this led to the creation of an erroneous reference set for function prediction approaches. Due to these issues, the focus shifted from to using more sophisticated forms of sequence similarity than simple alignment to predict function. The following section describes these approaches in detail.

3.3 Existing Approaches Beyond Simple Homology-based Annotation Transfer

In the domain of automated function prediction, sequences have been heavily utilized, in both direct homology-based and indirect subsequence- and feature-based approaches. Specifically, techniques that predict protein function from sequence can be categorized into three classes, namely, sequence homology-based approaches, subsequence-based approaches and feature-based approaches, which are explained below:

—**Homology-based approaches**: As discussed in Section 3.2, the results from simple homology-based approaches are not always accurate. Hence, approaches in this category attempt to make the homology search process more sensitive by multiple means, such as making the search probabilistic and adding evidence from other sources of data to obtain more accurate and confident annotations for the query proteins.

—**Subsequence-based approaches**: It has been shown in several studies that often not the whole sequence, but only some segments of it are important for determining the function of a given protein. Consequently, the approaches in this category treat these segments or subsequences as features of a protein sequence and construct models for the mapping of these features to protein function. These models are then used to predict the function of a query protein.

—**Feature-based approaches**: The final category of approaches attempts to exploit the perspective that the amino acid sequence is a unique characterization of a protein, and determines several of its physical and functional features. These features are used to construct a predictive model which can map the feature-value vector of a query protein to its function.

An important observation that may be observed from the above categorization is that the subsequence- and feature-based approaches are very similar at the fundamental level,

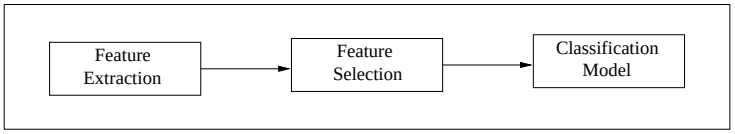


Fig. 4. General route adopted by the model-based approaches

since all these approaches involve construction of a model for the feature-to-function mapping. Hence, these categories can be grouped into the category of model-based approaches, which essentially follow the route shown in Figure 4. Following is a description of the three stages in this route:

—**Feature Extraction**: This stage involves the definition of features from a sequence, that can be used to encode the desired properties of a protein. Some of the popularly used features are motifs derived from a set of functionally or evolutionarily related proteins, functional domains, n-grams and more biologically meaningful features such as the isoelectric point, the Van der Waals volume and post-translational modifications.

—**Feature Selection**: Often, all the features used to encode a protein are not useful, since some features may be noisy and/or redundant. To handle this, some, but not all, approaches use feature selection techniques from data mining, such as χ^2 and Backward Elimination [Dash and Liu 1997].

—**Classification Model**: Finally, a classification model is constructed by training a classifier with feature-values vectors and their correponding functional classes. This model can then be used to assign functional labels to query proteins that have been converted into their corresponding feature-value vectors. Some classifiers that have been popularly used in this field are support vector machines (SVMs), neural networks (NNs) and the naive Bayesian classifier.

However, there are also significant differences between subsequence- and feature-based approaches, the most fundamental of them being that subsequence-based approaches extract the features, i.e., meaningful subsequences, such as motifs and domains, from a set of functionally related sequences. On the other hand, feature-based approaches derive and evaluate their features on the basis of individual protein sequences. Thus, the latter approaches are more "direct" than those based on subsequences. Another related difference is that the features used by the feature-based approaches are more biologically meaningful, since they are defined on the basis of the available knowledge about protein function, and model factors which may affect a protein's function. On the other hand, it is known that subsequences such as motifs and domains represent biologically meaningful portions of a protein, but it is hard to attach a specific meaning with them. As will be discussed in subsequent sections, this is the primary reason for the success of the feature-based approaches.

Now, with a high-level view of the field of sequence-based function prediction, we proceed to discuss in detail the approaches falling within the three categories of homology-based, subsequence-based and feature-based approaches.

3.3.1 *Homology-based approaches.* In Section 3.2, it was discussed that simple transfer of annotation from the most homologous sequence may not produce very accurate results, primarily because of the weak correlation between a protein's sequence and function.

This section discusses several approaches that attempt to make this technique more accurate by using various methods that make the homology search process more sensitive.

GeneQuiz [Andrade et al. 1999] was the first completely automated system for sequence analysis and annotation. The annotation module of GeneQuiz used the standard sequence comparison systems such as FASTA and PSI-BLAST, but also performed additional functions such as sequence filtering to identify the most significant portions of a sequence, using methods such as pattern discovery, multiple alignment and structural inference. Thus, this system focused both on utilizing off-the-shelf software for function prediction, as well as adding multiple evidence for the inferences made. The result was a more confident estimation of the function of the query protein. PEDANT [Riley et al. 2005] and AutoFACT [Koski et al. 2005] are other genome database systems focussed on similar goals and based on similar techniques. Together, these systems have enabled an integrated comparative analysis of sequences, often including proteins from other organisms a well.

The next major development in homology-based function prediction was the integration of Gene Ontology (GO) categories into the annotation process. The use of GO standardized the process since now organism-independent and hierarchically-structured functional categories were used. Consequently, several homology-based annotation systems were proposed on the basis of this idea, such as GOblet [Hennig et al. 2003], OntoBlast [Zehetner 2003], GOFigure [Khan et al. 2003], GoAnno [Chalmel et al. 2005] and GOPET [Vinayagam et al. 2006], which are essentially simple extensions of the similarity-based annotation technique. GOtcha [Martin et al. 2004] is a more sophisticated system that utilizes the GO hierarchical structure to find the most relevant annotations for a query sequence. For this process, first a set of homologues is found for the sequence using a BLAST search across various organisms, and the annotation set for these homologues is arranged in a set of GO-like DAGs (Directed Acyclic Graphs) [West 2001]. Based on the frequency of occurrence of the respective annotations and the E-values of the corresponding matches, a new score called the P-score is calculated for each annotation. This score acts as a measure of the confidence attached to the annotation of the query sequence with that term, and thus, the final set of annotations are retrieved by simply thresholding this score. Experiments on *D. melanogaster* (Malaria parasite) showed that the results were more sensitive and specific than those obtained by transferring the annotations of the top BLAST match. Thus, compared to the earlier systems, GOtcha was better able to integrate GO categories into the annotation process.

Besides direct annotation transfer, sequence homology has also been used in more indirect approaches for function prediction. Abascal and Valencia [2003] discuss some of the problems with the traditional annotation process, which are as follows:

—Function prediction errors introduced into the standard protein databases by the “classical” annotation strategies [Devos and Valencia 2000].

—Presence of multiple domains in a sequence, which contribute individually to the protein’s function, thus making it essential for any function prediction strategy to take into account the domain structure of a protein.

—Inconsistency between the levels of detail in different functional annotations.

In order to overcome these obstacles,several approach have proposed a multi-step strategy for functional annotation based on clustering of protein sequences according to their sequence similarities [Xie et al. 2002; Abascal and Valencia 2003; Sasson et al. 2006]. Figure 5 shows the flowchart of the basic strategy adopted in these two studies. The algorithm

starts with the construction of the a similarity matrix that stores the BLAST similarity values between the protein sequences in the original training set. This matrix is then used to cluster these sequences, and the annotation of a sequence in these approaches depends not on individual homologous sequences but a cluster composed of many such sequences. This makes the process more robust to errors in individual entries. Thus, at a higher level, it can be observed that these two approaches use homology detection only as an intermediate step in the complete annotation process, thus reducing the effect of the problems associated with the traditional annotation process.

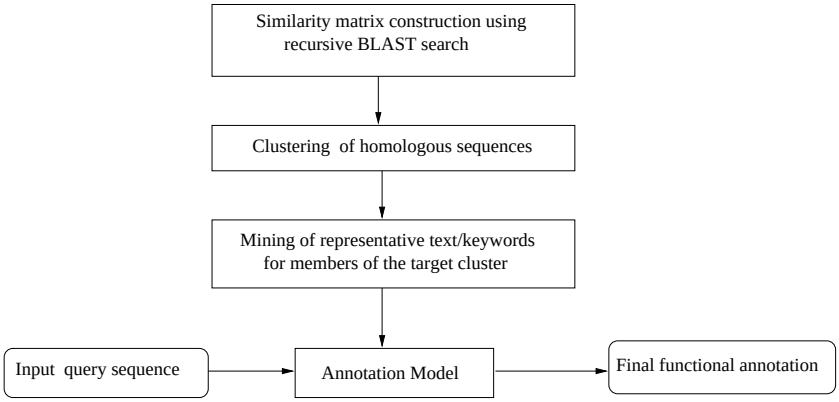


Fig. 5. Basic strategy adopted by approaches based on clustering of protein sequences according to their pairwise sequence similarities [Xie et al. 2002; Abascal and Valencia 2003; Sasson et al. 2006]

Another direction in which homology-based function transfer can be improved is by making the process probabilistic. Levy et al. [2005] do this by hypothesizing that a protein can only belong to a functional class if its BLAST score distribution with the members of the class is the same as that of these members themselves. In order to model this hypothesis, a univariate and a multivariate probabilistic scheme is proposed. The former scheme makes predictions simply on the basis of the total score of the target protein, by assigning it a probability of belonging to each class. However, this often leads to ambiguous results, and hence this scheme is extended to a multivariate one by constructing a vector of scores for all classes for the target protein and then comparing it against the distribution in each class. Results on a set of enzymes indicated a high accuracy of 90.6%. However, these results should be seen in the light of the fact that enzyme functions are more tightly correlated with their sequences than other proteins. Also, this scheme is expected to work only for very specific classes, for example, the most specific level of GO, since more general classes have significant overlap between them and thus the prediction may become ambiguous. Hence, more work is needed in this direction to rid homology-based prediction of its conceptual problems.

3.3.2 *Subsequence-based approaches.* Often only specific parts of whole sequence are crucial for the protein to perform its functions. A related example is that of *exons* in a gene sequence, which are substrings that are translated into an amino acid chain, and *introns*, which are subsequences that are excluded from translation, and hence do not have

a clear function in the sequence. Thus, in order to accurately model a protein's function, many approaches try to identify useful portions of the protein sequence that may contribute to the accomplishment of the function by the protein. This section reviews several such approaches. However, these approaches define "useful portions" in different ways, and some definitions are discussed before continuing with the discussion:

—**Motif:** Motifs are defined as sub-sequences which are conserved across a set of protein sequences belonging to a family [Bork and Koonin 1996]. Owing to their conservation property, they are candidates for functional sites in proteins, such as sites for ligand binding, DNA binding and interactions with other proteins, and thus are useful as clues for predicting the function of a protein [Bork and Koonin 1996; Huang and Brutlag 2001].

—**Domain:** It has been strongly hypothesized that the multiple functions performed by a protein are due to different regions of the protein sequence having different structural and functional characteristics [Servant et al. 2002]. These regions are known as functional domains, and a protein's function is a combination of the functions of each of these domains. However, this gives rise to the multi-domain problem, since now it is important to identify all the domains in a protein sequence in order to elucidate its function completely.

The above definitions indicate that identifying domains and motifs can be useful for predicting protein function. As mentioned earlier, these subsequences provide a new way of encoding the protein sequence in terms of features that encode whether a certain motif or domain is present in a sequence, and a confidence value of the match if desired. Once such a feature vector has been calculated for each protein in the target set, various statistics and data mining techniques, such as classfication, could be used. Many approaches based on this idea have been proposed, starting with [Hannenhalli and Russell 2000]. This approach tried to identify regions of a sequence that best distinguish a certain function or sub-type. This is done by identifying positions in a multiple sequence alignment of proteins in a family s, and finding the relative entropy of each position with respect to s and $\bar{s}$. The most discriminating positions are those having the highest total relative entropy with respect to each family considered. Thus, once this set of discriminating positions, which are the features in this case, has been constructed, the classification of a new sequence is carried out using the HMMER program [Eddy 1998]. Experiments on four particular enzyme classes and 42 Pfam [Sonnhammer et al. 1997] families showed that this method almost always had a higher accuracy than HMMER, though the difference was not very high.

The solution proposed by Wang et al. [2001] represents a mid-point between n-gram [Wu et al. 1992] and motif-based approaches, since they use both these types of features for classification using a Bayesian neural network (BNN) [MacKay 1992]. The features that were used encode two types of similarities between sequences, namely *global* and *local* similarity, modeled via n-grams and motifs respectively. Appropriate feature selection techniques are also used to reduce the total number of features. The final classification is carried out using the resulting features. The results on a set of four superfamilies from the PIR database [Wu et al. 2003] are better than BLAST [Altschul et al. 1997] and two version of SAM [Karplus et al. 1998]. However, for each class, a negative class is explicitly used, which may make the classification easier. Nevertheless, this study showed the merit of combining motif-based features with sequence-based features for protein classification.

Moving forward, a completely motif-based strategy is adopted by Liu and Califano [2001]. This is a decision tree-inspired approach to cluster proteins into functional families. Since motifs are strong signals for common family membership, the set of given sequences are characterized in terms of presence or absence of motifs derived by the SPLASH algorithm [Califano 2000]. This initial set of motifs is refined and expanded to find a substantial number of statistically significant motifs. An unsupervised top-down tree is then constructed by dividing the set of proteins at each node according to whether they contain the next most significant motif or not. The leaves so obtained are hypothesized to contain sets of proteins belonging to the same functional family, which represents a top-down clustering of the sequences. Upon validation using the set of G protein-coupled receptors (GPCRs), a classification rate in the range of $57 - 72\%$ is achieved. This is a reasonable performance since GPCRs are known to be a highly diverse family at the sequence level, and thus are hard to classify using automated methods [Moriyama and Kim 2006].

A very similar approach is presented in [Wang et al. 2003], which also proposes the characterization of proteins using the motifs they may or may not contain. However, this approach differs from from [Liu and Califano 2001] in that this is a supervised approach in which training examples are labeled with functional classes. Thus, once the proteins have been converted into binary vectors using the approach above, a decision tree is constructed on the training set, and is used for the classification of the test set. The use of manually curated protein families (from the MEROPS database [Rawlings and Barrett 1999]) and motifs (from the PROSITE database [Hulo et al. 2006]) for training gave this approach a significant edge over [Califano 2000] in terms of classification accuracy.

Another motif-based approach to protein classification that uses neural networks is presented in [Blekas et al. 2005]. Here, two ways of using motifs are proposed:

—**Class-independent motifs**: Motifs extracted from the entire set of training sequences.

—**Class-dependent motifs**: Motifs extracted separately for each class and then combined to make a global set.

In both cases, the MEME [Bailey et al. 1999] algorithm was used to extract about 30 such motifs were used to construct vectors for each sequence, which were fed into a neural network to build a classification model. In experiments on PROSITE [Hulo et al. 2006], it is found that class-dependent motifs form the best encoding scheme, which is expected since the motifs calculated are more class-sensitive. For the classification of the GPCR superfamily, this approach was shown to surpass SAM [Karplus et al. 1998] and MAST [Bailey et al. 1999] with respect to the ROC50 measure. However, now it has been shown that the naive Bayes classifier, in combination with the χ^2 feature selection algorithm is the most effective technique for the GPCR superfamily [Cheng et al. 2005]. This work is discussed in Section 3.3.3

A rigorous machine learning-oriented study of motif-based protein classification is reported in [Ben-Hur and Brutlag 2005]. Here, a *motif kernel* that simply uses the occurrence count of each motif in a sequence as a similarity measure between the motif and the sequence, is proposed, and is used for classification with an SVM. In addition, this paper also investigated issues such as feature selection for SVM, multiclass classification using SVM, and the multifunctional nature of proteins. The most optimal configuration was determined to be composed of (i) feature selection using the RFE method [Guyon et al. 2002], (ii) combination of many one-against-the rest classifiers [Rifkin and Klautau 2004] and (iii) counting the multiple classes of a protein as a single class. In this optimal configuration, on

a data set of enzymes, the results showed that SVM performs better than a k-NN classifier. Though enzymes are not the best benchmark for this application, the performance tuning of an SVM using a motif kernel will be valuable for future motif-based approaches. Yet another appproach that reports good results for motif-based SVM classification of enzymes is presented by Kunik et al. [2005].

The above descriptions show that motif-based approaches have come a long way from the initial idea that sequence motifs may represent functional units of a protein. However, a more direct approximation of these units are protein domains, whose use for the function prediction task is discussed next.

The first use of domains for funtion prediction appears to have been made in the simple strategy presented in [Schug et al. 2002]. In this approach, domains were extracted from two standard databases, namely ProDom [Servant et al. 2002] and CDD (Conserved Domain Database) [Marchler-Bauer et al. 2005], and rules for function assignment were constructed on the basis of BLAST searches on a set of $11,679$ GO annotated proteins from three popular genomes, namely *D. melanogaster*, *M. musculus* and *S. cerevisiae*. Application of these rules to a set of 4357 manually curated human proteins resulted in a recall of 81% and a precision of 74%, while on data sets for other organisms, both these figures were around the 50% mark. In a similar approach [Cai and Doig 2004], domains from the SBASE library of protein domains [Vlahovicek et al. 2002] are used as attributes, and vectors constructed the values of these attributes are classified using both the nearest neighbor algorithm (NNA) and support vector machines (SVM). The results obtained for thirteen functional classes from the MIPS database indicate that NNA is better for this task than SVM. This reversed order of classification accuracy (in most applications related to protein sequences, SVMs have produced better results than NNA) may be an artifact of the simple vector representations of proteins adopted in this study, i.e., binary vectors, which may not be the most informative input for a classification algorithm. Yet another approach based on the idea of domains is presented in [Perez et al. 2002]. However, in this case, instead of experimentally determined domains, constant-length statistically significant amino acid patterns called *protomotifs* [Thode et al. 1996] are used. Correlations between SWISS-PROT keywords assigned to the sequences and the positions of these patterns are found, and these are used to establish rules for function assignment. Coverage and precision close to [Schug et al. 2002] are obtained for a set of PROSITE [Hulo et al. 2006] sequences. Similarly, a low sensitivity of less than 50% for the three GO ontologies is achieved by the decision tree based classification of sequences [Hayete and Bienkowska 2005] expressed in terms of the PFAM domains [Sonnhammer et al. 1997], indicating the insufficiecy of the assumption of independent behavior of domains. Thus, even though substantial work has been done on the idea of predicting protein function using functional domains, the results obtained have not been very exciting . A fundamental reason for this is the assumption made by all the above approaches that each domain contributes a function to the protein independent of the other domains, which may not always be true. This assumption needs to be relaxed in order to get better results from this approach.

From the above discussion of the approaches for protein function prediction based on the identification of significant subsequences, it can be observed that the results obtained are not as impressive as expected. One of the reasons for this is the lack of a precise definition of subsequences such as motifs and domains. For instance, each of the above approaches modeled these patterns in a manner different from the others, and the results

varied accordingly. In addition, the programs used to extract these newly-defined patterns are only approximations, and hence, add a degree of error to the prediction process. Also, this two-step procedure leads to the additional issue of the most optimal encoding of sequences with respect to motifs and domains, which is acknowledged to be a hard problem. Hence, there is a severe need for a unified standard mathematical definition of sequence patterns, and standard databases containing high-confidence sets of such patterns.

The WILMA system [Prlic et al. 2004] addresses the above problems, and thus employs a different route for improving the accuracy of subsequence- or pattern-based functional classification. WILMA integrates various protein databases such as SWISS-PROT, IPI [Kersey et al. 2004] and WORMPEP [Harris et al. 2004] with sequence pattern databases, such as PROSITE [Hulo et al. 2006], Pfam [Sonnhammer et al. 1997], and PRINTS [Attwood et al. 2003]. The searching of patterns in sequences is also performed using an ensemble of methods such as RPS-BLAST [Altschul et al. 1997], PROSITE scans and FingerPRINTScan [Scordis et al. 1999]. This design amounts to sequence searching at multiple levels, and thus leads to a more confident annotation.

Yet another problem of subsequence-based methods is their coverage, i.e., the fraction of the entire protein sequence space covered by these subsequences. Just as the dictionary helps us make sense of any text in a language, a dictionary of sequence patterns that covers the entire sequence space can be used for annotating proteins with functions. This view is adopted by the creators of Bio-Dictionary [Rigoutsos et al. 1999]. The Bio-Dictionary consists of amino acid patterns known as *seqlets*, that are subsequences of a certain maximum length containing a number of don't care characters, thus making seqlets more flexible than strict subsequences. Interestingly, the TEIRESIAS algorithm [Rigoutsos and Floratos 1998] used to discover these pattern is inspired by frequent pattern discovery algorithms from the field of association analyis, which is a part of the subject of data mining [Tan et al. 2005]. The primary merit of seqlets is their extensive coverage of the protein sequence space, which is illustrated by the fact that the version of BioDictionary constructed for the May 14, 2001 release of SWISS-PROT, contained $42,996,454$ seqlets, which covered 98.2% of the processed input at amino acid level.

Rigoutsos et al. [2002] used these seqlets for annotating protein sequences using the approach shown in Figure 6. Essentially, this figure denotes the transfer of the attached field values of a seqlet to all the query sequences that it matches strongly. The approach is validated rigorously using the DE[7] field of SwissProt as the annotation, and it is found that many kinds of sequences and fragments, both long (human protein $UL78\_HCMVA$) and short ($VVVTAHAF$) were annotated correctly. In addition, the generalization capabilities of the algorithm were shown by an annotation accuracy of about 90% for three genomes which were not used in the construction of the Bio-Dictionary. Thus, this approach illustrates the benefits that data mining in particular, and computer science in general, can provide to the systematic study of biology.

Finally, it is important to mention a recent class of techniques that have been proposed for the detection of remote homologies among proteins. These are sequence- and motif-based approaches that use machine learning techniques such as SVMs [Kuang et al. 2005; Rangwala and Karypis 2005; Rangwala et al. 2006; Ben-Hur and Brutlag 2003] and HMMs [Jaakkola et al. 2000]. An extensive survey of these techniques appears in [French 2005]. Although in principle, these approaches can be used for function prediction, this

[7] http://www.expasy.org/sprot/userman.html#DE_line

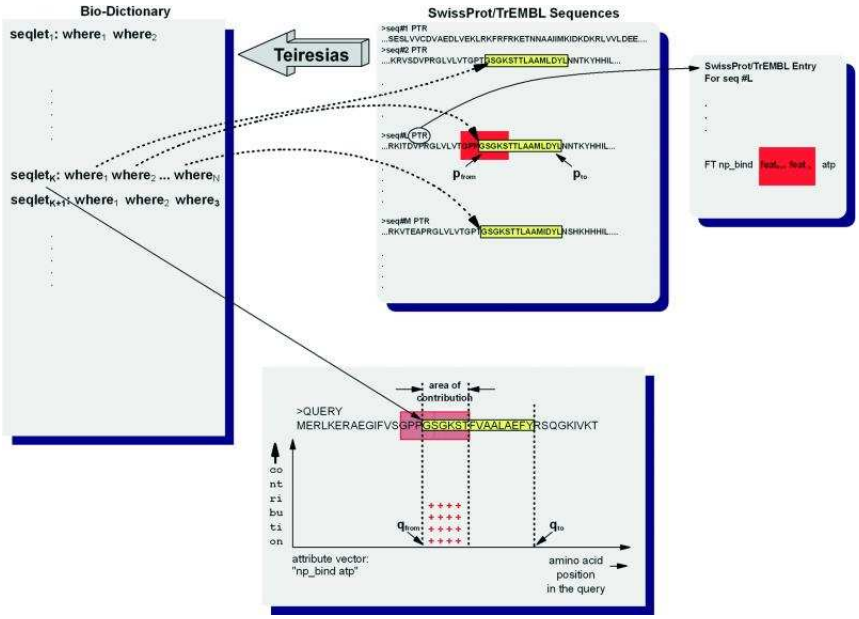


Fig. 6. Dictionary-driven protein annotation [Rigoutsos et al. 2002]

is not the preferred validation methodology for them. Instead, since it is more widely acknowledged that sequence and structural similarities are more tightly correlated, the SCOP superfamily classification [Andreeva et al. 2004] is considered the gold standard for validation here [Riley 1998]. Consequently, an estimate of the efficacy of these approaches for function prediction is not readily available, and hence these approaches are not covered in detail here. Nevertheless, SVMs have been shown to give the best performance in this field [French 2005], and an extension to function predicition may yield useful results.

3.3.3 *Feature-based approaches.* The approaches discussed in the above sections predict protein function from sequences in their raw form, i.e. as a string of characters. However, it is possible to transform these sequences into more biologically meaningful features, which make it easier to distinguish between proteins from different functional classes. This is the perspective adopted by this category of feature-based approaches, which use standard classification algorithms to learn models of functional classes from the transformed set of features, and then utilize this model to make predictions for uncharacterized proteins. The most commonly used classifiers in this class of approaches are support vector machines (SVM), neural networks (NN) and the naive Bayesian classifier. This section describes several such approaches.

It is well known that neural networks [Duda et al. 2000] were one of the earliest classification models. Hence, in early research in this field, ProCANS (Protein Classification Artificial Neural System) and its successors [Wu et al. 1992; Wu et al. 1995] were designed to used n-grams of protein sequences as features for a one hidden layer neural network. However, because of the limited availability of protein sequences, these systems were tested only on classes of enzymes. The observed performance of these models on these limited sets showed that choosing the right representation for sequences is an impor-

tant issue for this problem, a conclusion which was echoed in recent research studies [King et al. 2001].

In another study, King et al. [2000] attempt to demonstrate the utility of data mining for solving biological problems by proposing a solution for the function prediction problem via inductive logic programming (ILP). Their approach works on protein sequences and creates a binary feature vector for each of them. The features used for this transformation correspond to a certain form of frequent item sets, which essentially are sets of characteristics that are common to a significant fraction of proteins in the consideration set. Once these feature vectors have been created for all the proteins, the C4.5 decision tree learning algorithm is used to construct a rule set for predicting function from these features. Thus, this strategy is a combination of ILP and propositional data mining algorithms. Experiments on ORFs from *M. tuberculosis* and *E. coli* showed that the prediction accuracy achieved was about 65%, which was encouraging as this was among the first applications of data mining to protein sequences for the function prediction problem.

In an interesting extension of the above work, the same authors investigated the most suitable representation of a protein sequence for function prediction [King et al. 2001]. The three types of representations evaluated were the following:

—Sequence based attributes (SEQ) such as the number of residues of type R in the sequence, the length of the sequence, the molecular weight etc.

—Phylogeny based attributes (SIM) computed through the results of a PSI-BLAST search.

—Structure based attributes (STR) computed from the secondary structure prediction made by the Prof program [Ouali and King 2000].

The strategy proposed in [King et al. 2000] is applied to these representations, as well as their combinations, namely SEQ+STR, SEQ+SIM, SEQ+SIM and the last one consisting of all of them. Evaluation of the results on *E. coli* ORFs indicated that SIM is the most accurate representation of a protein sequence for function prediction. This is further evidence for the hypothesis presented in Section 6 that modeling evolution adds great strength to the prediction process.

PRED-CLASS is another feature-based approach to function classification of protein sequences [Pasquier et al. 2001]. However, it is more focused in scope since it models only the transmembrane (TM), fibrous (FIB) and globular (GLOB) classes. The model adopted is a three-level cascaded neural network, with TM proteins being classified at the first level, FIB at the next and the last level classifying between GLOB and undecided classes of proteins. The features used at the first level only depict the compositional features of the sequence, while the second level additionally employs the top thrity Fourier transform intensities, which model the periodicities of residues or groups in a sequence. Training and testing are carried out using small sets of proteins, 11 and 387 in size respectively, and a fairly high sensitivity-selectivity figure is achieved.

By far, the most cited work in this category of approaches is [Jensen et al. 2002]. This paper presents the ProtFun method for predicting function from sequence, which is based on the hypothesis that a protein has to undergo different types of modifications and sortings using the cellular machinery, before it performs its function. These are knowns as post-translational modifications (PTMs) [Mann and Jensen 2003], some of which include N- and O-glycosylation, (S/T/Y) phosphorylation and cleavage of N-terminal signal peptides controlling the entry to the secretory pathway. ProtFun uses 14 such attributes computed

for various tools available for measuring individual factors, and trains a set of neural networks on such attribute-value pairs for the available training set.

Another interesting component of this paper was that the validation of ProtFun was performed on a set of 5500 human proteins from the TrEMBL database. The functional categories were automatically assigned on the basis of their SWISS-PROT keywords using the EUCLID system [Tamames et al. 1998]. The results of these experiments were very encouraging, with a sensitivity of 90% and a 10% false positive rate being achieved for some functional categories. Similar results were obtained when ProtFun was extended to cover GO categories corresponding to human protein functions [Jensen et al. 2003], thus further validating the strength of the PTM approach to function prediction.

SVM-Prot [Cai et al. 2003] is another function prediction tool based on SVMs. Here also, every protein sequence is represented by a set of residue-specific features such as normalized Van der Waals volume, polarity, charge and surface tension, which are averaged over all the residues to in the sequence obtain the feature-value vector for the protein. Two-way classification (positive or negative) is then carried out for each functional family considered using an SVM. Accuracies obtained on standard databases such as BRENDA [Schomburg et al. 2004], GPCRDB [Horn et al. 2003] and NucleaRDB [Horn et al. 2001] are in the high range of $69.1 - 99.6\%$. The capability of SVM-Prot is also shown in the classification of 49 novel plant proteins, of which it was able to predict the classes of 31 accurately and approximately accurately for 4, thus leading to a reasonable accuracy of 71.4% [Han et al. 2005].

In a recent publication, Eisner et al. [2005] performed a detailed machine learning-oriented study of the problem of assigning hierarchical GO classes to proteins effectively. Their CHUGO (Classification in a Hierarchy Under Gene Ontology) system employs the following three ways of handling this problem. It should be noted that, in general, these are important issues for any classification study and thus should be addressed carefully in order to obtain accurate results.

—*Training set design*: Since a protein being assigned to a GO node N implicitly assigns it to all the ancestors of N, the ideal training set for a learning algorithm should label the protein with the whole GO subgraph terminating at N. This naturally, leads to an improvement in recall, and hopefully in precision as well.

—*Classification model*: Each GO node in CHUGO is attached to a separate binary classifier, which is actually an ensemble of classifiers, since a protein can belong to multiple functional classes.

—*Evaluation methodology*: As in the training set design, expanded label sets should be used for all the proteins for calculating the evaluation metrics such as precision and recall.

Though the above ideas have intellectual merit, they were not supported substantially by the experimental results. For instance, the classifier had almost the same performance as BLAST for 89% of the proteins, which does not illustrate the power of classification for functional annotation. Similarly, the most appropriate method for training set design, i.e. the all inclusive approach, led to an obvious improvement in recall, with no noticable improvement in precision, but instead, a significant increase in computation time. These results suggest that these ideas need to be formulated more carefully for effective use.

The PANTHER database [Mi et al. 2005] expands the general framework of protein sequence analysis databases by indexing protein families and subfamilies according to their

GO functional labels. In its basic form, it consists of two parts: *PANTHER/LIB*, a library of protein families and subfamilies, and *PANTHER/X*, a set of ontology terms describing protein function. The families and subfamilies are created by clustering the $256,413$ constituent proteins using a single-link hierarchical clustering algorithm, and are manually annotated with functional labels. A SAM profile [Karplus et al. 1998] is constructed for each family, and is used for classifying novel proteins. Efforts are currently underway to integrate PANTHER with the InterPro [Apweiler et al. 2000] and PIR [Wu et al. 2003] databases.

One of the most important of these classes is that of the G-protein coupled receptor proteins (GPCR), since they are the largest family of proteins found in the human body, and are the targets of approximately 60% of the approved drugs in the market [Gether 2000]. In addition, this superfamily is known to be very diverse in terms of sequence homology, and hence are particularly attractive targets for sequence classification research [Moriyama and Kim 2006]. Until recently, the state-of-the-art classification technique for this family was considered to be SVM, as concluded by Karchin et al. [2002], who showed that a complex classifier such as an SVM is more effective for the classification of diverse families such as GPCRs, than simpler classifiers such as HMMs, which are more suitable for coarser classification tasks, such as superfamily classification. However, in a systematically revealing study inspired by document classification techniques [Cheng et al. 2005], it was shown that even a simple classifier such as the naive Bayesian classifier, in combination with the χ^2 feature selection technique, was able to surpass the performance of SVM at the task, with a lower computational cost. The χ^2-based technique was used here since it has been shown to work best for text classification applications [Yang and Pedersen 1997]. Also, this study showed that in combination with feature selection algorithms, simple encoding schemes such as n-grams, as used in very early studies [Wu et al. 1992], may be more effective than alignment-based schemes that also take into account the ordering of amino acids in a protein sequence. Indeed, the utility of feature selection for creating better encodings of protein sequences has also been shown by other systematic studies [Al-Shahib et al. 2005], this time for a wider set of functional classes and with a more diverse set of features.

Overall, from the above discussion of the approaches, it is clear that feature-based approaches are better able to handle the function prediction task than homology- or subsequence-based approaches, because of the inclusion of more biologically meaningful features, such as post-translational modifications [Jensen et al. 2002]. This enables the construction of a more robust model for the sequence-function mapping. However, there is still much scope for work and better results in this field.

3.4 Discussion

Previous sections showed how protein sequence data can be exploited for function prediciton using various homology-, subsequence- and feature-based approaches. In many cases, good results were also obtained at the task. However, an important caveat that should be stated here is that even though a sequence forms a unique characterization of protein, it is still a weak representation for complex operations such as the prediction of its function. In comparison, more complex forms of data such as gene expression and protein interaction networks offer a deeper insight into the mechanisms leading to the performance of a protein's function, and are thus more useful for predicting function (See Sections 7 and 8 for details). In fact, it has been suggested in the literature that the sequence- structure correlation is much stronger than the sequence- function correla-

tion [Devos and Valencia 2000; Whisstock and Lesk 2003], and hence many approaches take the $sequence \rightarrow structure \rightarrow function$ route for function prediction [Fetrow and Skolnick 1998]. Details of such approaches can be found in the next section.

4. PROTEIN STRUCTURE

4.1 Introduction

A protein is an organic biopolymer that is comprised of a set of amino acids, and assumes a configuration in three-dimensional space due to interactions between these constituents. Protein structure may be specified at multiple levels. Usually, it is specified at three levels, with a fourth level being specified for some cases [Schulz and Schirmer 1996]. Following is a brief description of these levels, which are also illustrated graphically in Figure 7:

(1) **Primary structure**: The primary structure of a protein is simply a sequence of amino acids. This level has been discussed earlier in Section 3.

(2) **Secondary structure**: The sequence of a protein influences its conformation in three-dimensional space via the formation of bonds between spatially close amino acids in the sequence. This process is popularly known as *protein folding*, and leads to the creation of substructures such as α-helices, β-sheets, turns and random coils, of which the first two are the most common, while the last two are formed very rarely. The collection of these substructures forms the secondary structure of a protein.

(3) **Tertiary structure**: The attractive and repulsive forces among the substructures caused by the folding balance each other and provide the protein with a relatively stable, though complicated, three-dimensional structure. This structure is known as the *tertiary structure* of the protein.

(4) **Quarternary structure**: Some proteins, such as the *spectrin* protein [Fuller et al. 1974], consist of multiple amino acid sequences, also known as protein subunits. Each of these sequences folds to form its own tertiary structure, which come together to produce the *quarternary structure* of the protein.

Owing to the relatively systematic nature of the protein structure formation process, several experimental methods have been devised for the determination of the tertiary structure of a protein, including X-ray crystallography [Drenth 1999] and nuclear magnetic resonance (NMR) [Cavanagh et al. 1996]. However, the cost of these approaches have prompted a rapid growth of automated structure prediction techniques, such as PHD [Rost 1996], PROF [Ouali and King 2000] and NNSSP [Salamov and Solovyev 1995]. Many of these tools have been integrated into the ProteinPredict server [Rost et al. 2003].

In addition to these individual systems, a particularly interesting initiative in the field of protein structure prediction is the CASP[8] (Critical Assessment of Structure Prediction) experiment [Bourne 2003; Moult 2005]. CASP is a bi-annual contest, in which several participants submit potential structures for a given set of proteins, which have been pre-classified into three classes on the basis of the expected level of difficulty: comparative modeling, fold recognition or threading, and new fold recognition or *ab initio* methods. Recently, an additional event called CAFASP (Critical Assessment of Fully Automated Structure Prediction) has been initiated, which focuses on evaluating the fully automated techniques for predicting the structure of a given set of proteins within a limited amount of

[8] http://predictioncenter.gc.ucdavis.edu/

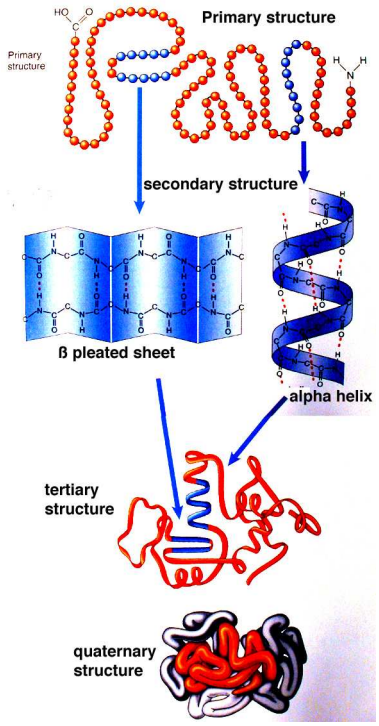


Fig. 7. Four levels of protein structure (image taken from [Campbell and Reece 2004])

time (two days). Together, CASP and CAFASP have played an integral role in identifying the issues and future needs of protein structure prediction.

Owing to the significance of protein structure, structural data collected using these experimental and computational methods have been collected in several standardized databases. However, since structural bioinformatics [Bourne and Weissig 2003] is a relatively new field compared to its sequence counterpart, the sources for protein structural data are not as diverse as those for sequences. Three standard databases dominate the structure data landscape: PDB [Berman et al. 2000], SCOP [Andreeva et al. 2004] and CATH [Orengo et al. 2002]:

—**PDB** (**Protein Data Bank**): PDB [Berman et al. 2000] is by far the most extensive and popular repository of experimentally determined protein 3D structures. As of May 11, 2006, it contains structures determined by various experimental methods such as X-ray crystallography, NMR spectroscopy and electron microscopy, for about $33,300$ proteins. In addition, there are about 3000 structures of other large molecules such as nucleic acids, protein complexes and others. There are additional tools for structural analysis available on the PDB website.

—**SCOP** (**Structural Classification of Proteins**): SCOP [Andreeva et al. 2004] can be considered an extension of PDB since it supplements a subset of the latter with useful functional information. The main motive of SCOP is to organize the available structures in a hierarchy so as to elicit the evolutionary relationships between them. The three

levels of hierarchy are:

—**Family**: Clear evolutionary relationship observed from high sequence similarity.

—**Superfamily**: Probable common evolutionary origin despite low sequence similarity, observed from similar structural and functional features.

—**Fold**: Major structural similarity, observed from the same secondary structure elements arranged in the same manner with the same topological connections.

The functional underpinnings of this hierarchy are clear, and for this reason, SCOP has become a gold standard for remote homology detection techniques [Jaakkola et al. 2000; Ben-Hur and Brutlag 2003; Kuang et al. 2005; Rangwala and Karypis 2005; Rangwala et al. 2006].

—**CATH** (**Class, Architecture, Topology and Homologous superfamily**): The expansion of the acronym CATH itself describes the primary purpose of this database, which is to organize a subset of the structures in PDB according to their similarity [Orengo et al. 2002]. The main differences from SCOP are the more detailed nature of the classification and the automated method for classifying structures according to this hierarchy.

Together, these databases form the core of data sources for structural proteomics and bioinformatics studies. Additional specialized databases are also available, such as DALI [Dietmann et al. 2001] and FSSP [Holm and Sander 1994].

Finally, having seen a detailed description of protein structure and how it can be inferred using experimental and computational methods, the interesting issue is how to use it for inferring the function(s) of the corresponding protein. The subsequent sections discuss this issue from several angles, namely the different forms and sources of structural data, the relationship between protein structure and function, and approaches that exploit this relationship to infer function from structure. However, before we proceed with these discussions, it is important to make a special mention of [Bartlett et al. 2003], which is a review article that has systematically studied the field of structural genomics and provides significantly useful information about its various aspects. In particular, it covers the following aspects:

—**Forms of Structural Data**: Three-dimensional structure and protein-ligand complexes.

—**Structure-Function Relationship**: Relations between function and structural classes, folds, homologous families and analogues.

—**Assigning Function from Structure**: *Ab initio* prediction, structural comparisons and structural motifs, and several programs to find the latter, such as TESS [Wallace et al. 1997], FFF [Fetrow and Skolnick 1998] and SPASM [Kleywegt 1999].

This discussion of function prediction from structure is a more detailed and updated version of [Bartlett et al. 2003].

4.2 Is Structure Tied to Function?

In many biological processes, the interacting entities have to come into physical contact in order to accomplish the desired task. This indicates a connection between structure and function, since the structure of a protein determines several of its functional features, such as it cellular location, the types of ligands it binds to and other proteins it can interact with. A very important example of these features, which can be discovered effectively on the basis of protein structure, is that of active sites in enzymes. These are parts of the enzyme to

which the reaction substrate binds itself, and thus are fundamental for catalysis by the enzyme, which is the basic function of an enzyme. This example illustrates that the structure of a protein is expected to be of great utility in inferring its biological function [Skolnick et al. 2000]. This section presents a review of several approaches that present evidence for the $structure \rightarrow function$ route for function prediction.

In a landmark paper, Martin et al. [1998] broke the ground for this field by exploring if the structural fold of a protein is correlated to its function. This work was conducted in the restricted domain of enzymes, since they are very well studied and their structures are abundant in PDB [Berman et al. 2000]. Also, the analysis procedure was very simple, namely the construction of a specialized form of pie charts known as CATH wheels [Orengo et al. 2002], which essentially show the distribution of the different types of folds among a given set of related proteins. When this procedure was applied to the six Enzyme Classification classes [Serres and Riley 2000] individually, it was seen that the enzyme function, represented by the first digit of the EC number, could not be tightly correlated with the over-representation of any of the three structural classes, α, β and $\alpha\beta$. This indicated that there isn't a very strong correlation between the structural and functional classes of an enzyme. Zooming further, for some important types of enzymes, this analysis indicated that there is significant correlation between structural class and lower level functional features such as the type of ligand the enzyme binds, and whether it is an intra- or extra-cellular enzyme. Combining the above findings for enzymes, Martin et al. [1998] concluded that even though the structure of a protein is not tightly correlated directly with its biological function, it is correlated with lower-level functional features. Indeed, researchers have advocated the use of these features for predicting the function of a protein from its structure [Thornton et al. 2000], as shown in Figure 8.

The conclusions made in [Martin et al. 1998] were further confirmed by several subsequent studies [Hegyi and Gerstein 1999; Orengo et al. 1999; Thornton et al. 1999]. Hegyi and Gerstein [1999] conducted approximately the same correlation analysis as Martin et al. [1998] for all the single-domain single-function proteins in SWISS-PROT. From this analysis, it was found that more than half of the functions are associated with at least two different structural classes, while almost half of the structural classes are association with least two functions. Very similar results were obtained when the analysis was subjected to the following variations in the base data:

—Individual genomes, such as Yeast and E. Coli.

—Different functional classification schemes, such as MIPS [Mewes et al. 2002] and COGs [Tatusov et al. 1997] and ENZYME [Bairoch 2000].

—Different structural classification schemes, such as SCOP [Andreeva et al. 2004] and CATH [Orengo et al. 2002].

Thus, this study confirmed the findings of Martin et al. [1998] at a more global scale. Finally, Orengo et al. [1999] and Thornton et al. [1999] discuss the construction of the CATH database [Orengo et al. 2002], and report that they did not observe any strong correlation between structure and function during this construction. Overall, these studies demonstrated that the structure-function correlation is not strong enough to enable the inference of function directly from a protein's structure. However, a possible path suggested was the conversion of protein's structure into lower-level functional features, which could be mapped to its function more robustly [Martin et al. 1998; Orengo et al. 1999; Thornton et al. 1999].

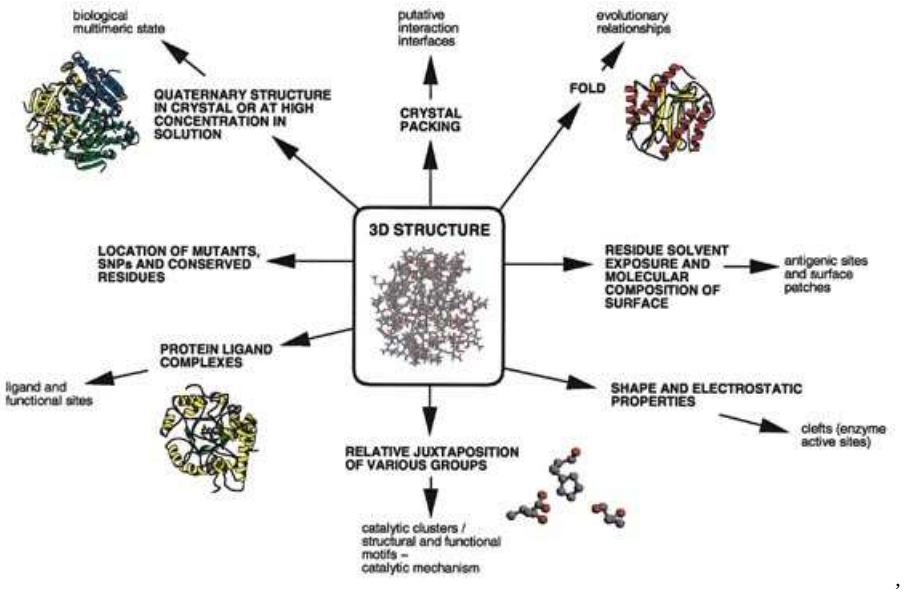


Fig. 8. Possible approaches for deriving functional information from protein structure (Figure taken from [Thornton et al. 2000])

In light of the results mentioned above, new ideas for inferring functional features from the structure of a protein were proposed. Moult and Melamud [2000] discussed the derivation of function from three forms of structural information, namely fold, structural features and models of protein structure, and various sub-categories thereof. Jones and Thornton [2004] zoomed in on the second category, namely the use of structural features, and discussed approaches such as identifying functionally important sites in proteins, predicting enzyme active sites, and predicting DNA-binding sites such as the helix-turn-helix (HTH) motif, which is found widely in DNA-binding proteins [Luscombe et al. 2000]. Skolnick and Fetrow [2000] and Fetrow et al. [2001] also discuss the importance of active site identification for function prediction. However, their analysis is more tied to their Fuzzy Functional Form (FFF) technique [Fetrow and Skolnick 1998], which is discussed in detail in Section 4.3.2. Najmanovich et al. [2005] promote the use of local similarity measures to predict protein function and popular techniques used for this task. Finally, Wild and Saqi [2004] complete the spectrum by reviewing the major hints provided by structure for determining function, and cites biological examples from the literature as evidence for each of these hints. An apt summary of the ideas discussed in these studies is presented graphically in Figure 8, which identifies eight different types of information that can be derived from the structure of a protein and can be used to derive various types of functional information about it [Thornton et al. 2000]. Thus, motivated by the immense potential shown by the field of structural genomics for function prediction, several techniques based on different representations of protein structure have been proposed. We review these techniques in the next section.

4.3 Existing Approaches

After the discussion presented in the previous section, it is clear that structure can be utilized in various ways to predict protein functions. Correspondingly, many groups have proposed various structural features and approaches for exploiting them for prediction. These approaches can be largely classified into the following four categories:

—**Similarity-based approaches**: Given the structure of a protein, these approaches identify the protein with the most similar structure using structural alignment techniques, and transfer its functional annotations to the query protein.

—**Motif-based approaches**: The approaches in this category attempt to identify three-dimensional motifs, that are substructures conserved in a set of functionally related proteins, and estimate a mapping between the function of a protein and the structural motifs it contains. This mapping is then used to predict the functions of unannotated proteins

—**Surface-based approaches**: It is sometimes necessary to analyze the structure of a protein at a higher resolution than that of distances between consecutive amino acids. This corresponds to the modeling of a continuous surface for the structure and identifying features such as voids or holes in these surfaces. The approaches in this category utilize these features to infer a protein's function.

—**Learning-based approaches**: This category of recent approaches employ effective classification methods, such as SVM and k-nearest neighbor, to identify the most appropriate functional class for a protein from its most relevant structural features.

In the subsequent sections, we review each of these categories in detail. However, a caveat that must be made before we proceed with this discussion is that the form of function that most of the approaches below work with is the biochemical function of a protein [Laskowski et al. 2003] (molecular function in GO [Ashburner et al. 2000]), since the structure gives clues only to the chemical processes a protein undergoes before achieving its function. Of course, information like this could be extended to other forms of function, such as biological process.

Also, from the above description, it may be noted that some of these categories, such as those corresponding to similarity-based and motif-based approaches, are motivated by ideas from protein sequences, since the three-dimensional structure of a protein may also be considered as a sequence of tuples of coordinates, each tuple corresponding to an amino acid in the protein sequence. With this background, we now proceed to the discussion of the approaches in the above categories and how they exploit various structural features to infer the function of a given protein.

4.3.1 *Structural Similarity-based Approaches.* The easiest way of functionally annotating a protein based on its structure is to find another protein with a similar structure and transferring the latter's function to the former, just as in the case of protein sequences (Sections 3.2 and 3.3.1). A useful insight here, as mentioned above, is that a protein's structure is a sequence of tuples of three coordinates corresponding to the location of each of its amino acids (or their constituent atoms) in space. One way of solving this similarity estimation problem is by mapping it to the well-known alignment problem, that has received significant attention in the sequence alignment case, and for which many tools are now available [Altschul et al. 1997; Higgins et al. 1996]. Several programs have been designed for solving the structure alignment problem [Wolfson et al. 2005], the most popular of which are listed in Table V.

Program	Reference
DaliLite	[Holm and Park 2000]
CE-MC	[Shindyalov and Bourne 1998]
SSAP	[Orengo and Taylor 1996]
SSM	[Krissinel and Henrick 2004]
STRUCTAL	[Kolodny and Linial 2004]
LSQMAN	[Kleywegt 1996]
MultiProt	[Shatsky et al. 2004]
3DCoffee	[O'Sullivan et al. 2004]

Table V. Popular structure alignment programs

The first six entries in Table V are pairwise alignment algorithms, while the last two are designed for multiple alignment. Kolodny et al. [2005] provide an extensive and systematic comparison of pairwise alignment programs, concluding that STRUCTAL [Kolodny and Linial 2004] and SSM [Krissinel and Henrick 2004] perform the best. This result and the comparison procedure can significantly aid the identification of the most appropriate structure alignment algorithm for a given application, such as function prediction.

However, for the function prediction task, even though more information is available from the use of structure, alignment-based approaches suffer from problems similar to their sequence counterparts. Some of these problems include the unavailability of a sufficiently similar protein that has been annotated with a function, and the not-so-tight correlation between structural and functional similarity, which was discussed in Section 4.2. Hence, more specialized approaches have been proposed for solving this problem. These approaches adopt more sensitive similarity estimation methods, and derive their inferences from a larger set of similar proteins, instead of a single protein in a simple alignment-based approach. Thus, the ideas used here are similar to those used by advanced sequence homology-based approaches discussed in Section 3.3.1.

The PHUNCTIONER system [Pazos and Sternberg 2004] utilized the structural alignments from the FSSP database [Holm and Sander 1994] to find the positions in a structure of a protein which are functionally most important for a particular GO category. This importance is calculated using a Z-score based on the conservation of residues in these alignments. In cross-validation experiments on 121 GO terms at different levels of the hierarchy, an accuracy in the range of 75% to $> 90\%$ was obtained, which is much higher than that obtained by simply using sequence identity. ROC analysis of the two methods also gives similar results. These results proved the higher sensitivity obtained by using structural instead of sequence profiles for similarity searching.

Hou et al. [2005] build upon their earlier work on constructing a protein structure space map (SSM) [Hou et al. 2003] using the multi-dimensional scaling (MDS) technique. The hypothesis underlying this mapping is that proteins sharing similar molecular function are located in the vicinity of each other in this structure space map. The implementation consists of the conversion of the structure data into a dissimilarity matrix using the DaliLite scores [Holm and Park 2000] and selection of the most informative dimensions in this matrix using the strategy proposed by Williams [2002]. Finally, in the constructed structure space, a novel protein is classified as the GO category of the proteins lying within a distance threshold. The ROC analysis on the proteins in the PDB_SELECT data set [Hobohm and Sander 1994], which is a representative subset of size 1949 of the PDB database, showed that the SSM method is superior to those based on simple sequence similarity

(BLAST [Altschul et al. 1997]) and DALI scores [Dietmann et al. 2001]. Thus, this approach presented an effective Cartesian embedding of protein structures, which could be useful for many other purposes such as evaluating the pairwise similarity and subsequent clustering of these structures.

Some studies have also tried to combine structural similarity measures with other measures of similarity. One such study [Shakhnovich 2005] analyzed the correlation of the functional similarity of two protein, with their phylogenetic and structural similarity. It was concluded from this analysis that even though phylogenetic similarity is a better determinant of functional similarity than structural similarity, a combination of the two methods, i.e. considering structural similarity in a phylogenetic context can improve the precision of functional annotation. Thus, this study attempts to build a case for combining structural information with other forms of data in order to solve the function prediction problem accurately.

4.3.2 *Three-dimensional Motif-based Approaches.* Another category of approaches that closely mirror their sequence counterparts are those based on motifs in protein structures (Section 3.3.2). Just like a sequence motif, a structural motif is a three dimensional substructure of a protein that occurs in the structures of several related proteins. A very well known example of a structural motif is the helix-turn-helix (HTH) motif, which is found in many DNA-binding proteins [Luscombe et al. 2000]. However, it is notable that structural motif finding programs, such as TESS [Wallace et al. 1997], FFF [Fetrow and Skolnick 1998] and SPASM [Kleywegt 1999] rely on their own definitions of a structural motif, since there does not exist a universally accepted definition of the concept. This also holds for approaches that infer function from such motifs, as will be seen below.

By far, the most widely cited work in this area is the use of Fuzzy Functional Forms (FFFs) [Fetrow and Skolnick 1998] for the prediction of function from structure using the $sequence \rightarrow structure \rightarrow function$ paradigm, which has been motivated by the finding that the sequence similarity is correlated more with structural similarity than functional similarity [Devos and Valencia 2000; Whisstock and Lesk 2003] (Section 3.2). FFFs are fuzzy three-dimensional descriptors of specific protein functions, and are based on the geometry, residue identity and confirmation of protein active sites. They are constructed by superimposing structural information from several functionally related sequences, using the algorithm shown in Figure 9.

The specificity and uniqueness of FFFs, even when constructed from low-resolution structural data obtained from *ab initio* and threading experiments, was shown via their application to the glutaredoxin/thioredoxin and T_1 ribonuclease classes of enzymes. In subsequent papers [Skolnick and Fetrow 2000; Fetrow et al. 2001; Gennaroa et al. 2001], very impressive results were obtained on larger and more varied sets of protein structures, thus demonstrating the ability of this technique to build characterization of function from structure. However, it must be noted that even though the algorithm for the construction of FFFs is very systematic, the action(s) in each step has (have) to be performed by a human expert. For instance, the very first step in the algorithm, i.e., the identification of residues important for a function, and the examination of functionally related structures, require a structural biologist. This extensive human involvement reduces the coverage of the approach, and explains why the above results are only presented for a couple of enzyme classes. Thus, in order to extend the coverage of this technique, automated methods for extracting active sites from structures, such as those presented by Pazos and Sternberg

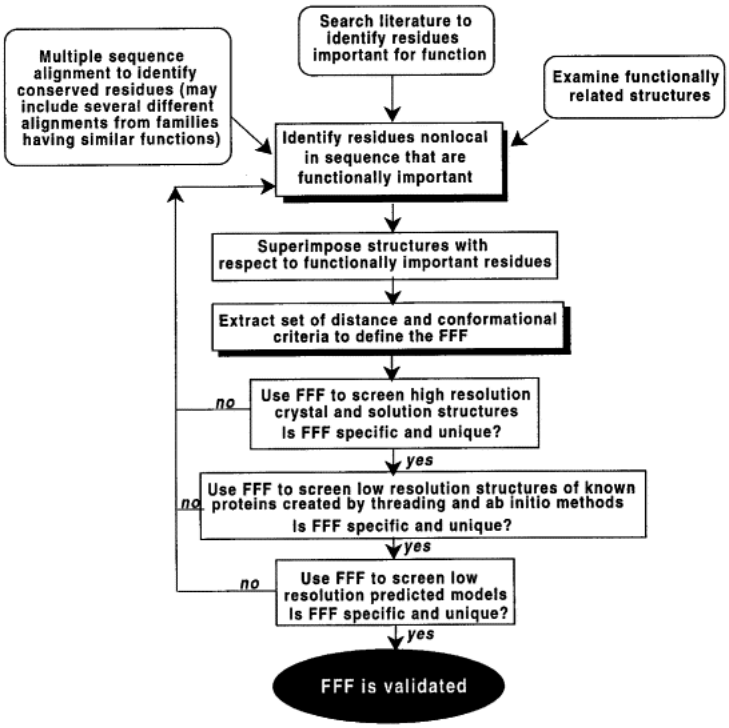


Fig. 9. The FFF construction algorithm (Figure taken from [Fetrow and Skolnick 1998])

[2004], are urgently needed. This will ensure the generation of FFFs on a large scale.

In a similar approach, Suzuki et al. [2005] used their FCANAL (Fast Calculable protein function ANALyzer) system to define a few (two to four) functionally important residues constituting the core of a functional site for a given function, construct a local similarity matrix for the other residues, and then assign a novel protein to the function corresponding to functional site most similar to it. Their technique was tested on a small set of 31 enzymes, on which it gave reasonable results. However, for a more robust evaluation, a bigger test set needs to be used.

Finally, the PROCAT database, which is constructed from the 3D enzyme active site templates extracted using the TESS algorithm [Wallace et al. 1997], is a significant step in the direction of enhancing the scalability of structural motif-based function prediction algorithms, and can be considered the structural counterpart of the commonly used PROSITE [Hulo et al. 2006] and PRINTS [Attwood et al. 2003] sequence motif databases. SMoS (Structural Motifs of Superfamilies) [Chakrabarti et al. 2003] and DSMP (Database of Structural Motifs in Proteins) [Guruprasad et al. 2000] are other useful structural motif databases. These databases will greatly assist in identifying regions of protein structures that are essential for the accomplishment of a given function and the subsequent annotation of novel proteins.

4.3.3 *Surface-based Approaches.* Traditionally, the structure of a protein has been defined as a sequence of tuples of three coordinates, each tuple corresponding to the location

of each of its amino acids (or their constituent atoms) in space. This definition indicates that the intermolecular interactions that lead to a certain biochemical function being performed occur at the level of amino acids or their atoms. However, in many cases, such interactions occur due to the complementarity of the molecular surfaces of the proteins. For instance, it has been shown in studies that hydrophobic surfaces often act as the interfaces between interacting molecules [Oda et al. 1998], and electrostatic molecular surfaces are also often used to explain protein functions [Honig and Nicholls 1995; Nakamura 1996]. These studies have motivated the need to model protein surfaces, which essentially is the specification of protein structure at a resolution finer than that of distances between consecutive amino acids, and computational approaches for this task have been proposed [Connolly 1983; Ferre et al. 2004]. Also, the additional information provided by this higher-resolution structure of a protein has been utilized by several approaches for the prediction of function prediction. These approaches, are based essentially on the idea of matching local patterns between two protein surfaces.

The first of these approaches adopts a graph theoretical approach to the problem of surface matching [Kinoshita et al. 2001]. In this work, the surfaces for a set of proteins structures from PDB is calculated using the MSP program [Connolly 1983], and points at small distances (1.4 $\mathring{A}$ here) are annotated with their electrostatic potential and hydrophobicity, properties that are important for the accomplishment of biochemical functions. A database known as *eF-site* (*e*lectrostatic-surface of *F*unctional site) was developed, consisting of such annotated surfaces for important functional sites of commonly found proteins, such as enzymes. Overall, this database currently contains a rich collection of molecular surfaces corresponding to over $19,000$ functional sites [Kinoshita and Nakamura 2004]. As for the matching stage, the global similarity of two protein surface is calculated by converting the surface comparisons at each point into a graph, from which a clique is then extracted, denoting the portion of maximum match between the surfaces. Two proteins are said to be similar if their largest clique has a similarity score above a certain threshold. Since the clique operation may become intractable due to the potentially large number of vertices in the comparison graph, two heuristics were also applied to the search procedure. The robustness of the algorithm was shown by individual examples of success for each of the four components of the eF-site database. and is a useful structural genomics resource for future studies.

In another approach, Binkowski et al. [2003] focus on the voids and pockets in a protein surface, working under the hypothesis that these are regions in a protein structure to which a solvent or a ligand can gain access, and thus aid the protein in performing its function(s). The pockets are computed from a protein structure using the approach of Edelsbrunner et al. [1998], and the amino acid residues composing them are composed into motifs known as *pvSOAR* patterns, which are in turn collected in the pvSOAR database. This database can now be used to perform one-against-all and all-against-all alignment-based searches for prediction and evaluation respectively. Indeed, in the latter type of experiments, with SCOP [Andreeva et al. 2004] and CATH [Orengo et al. 2002] classifications as gold standard, numerous examples were found that could be functionally classified just on the basis of pocket surface patterns. However, many cases were found in which the matching patterns belonged to different folds or functional families which may indicate a remote relationship between the two groups. Another useful contribution of this work was a method for calculating the statistical significance of the match between two short residue patterns,

based on the extreme value distribution (EVD). This could also be useful for other studies.

Ferre et al. [2004] also adopt the same strategy as Binkowski et al. [2003], i.e., the identification of local surface patterns(known as *clefts* here) which could be used for the functional annotation of query proteins. However, it differed from Binkowski et al. [2003]'s approach in three ways. First, the SURFNET algorithm [Laskowski 1995] was used to find surface clefts. Second, the patterns were annotated with GO codes using the PROSITE database [Hulo et al. 2006]. Finally, the matching algorithm considered both the structural and residue similarity, measured using the RMSD measure and the PAM matrix [Dayhoff et al. 1978], respectively. An all-against-all evaluation on the set of patterns constructed from the structures in the PDB_SELECT [Hobohm and Sander 1994] data set gave an accuracy of about 90%. This whole system has been organised into the SURFACE database [Ferre et al. 2004], which is freely accessible via the web.

In the latest approach in this category, Espadaler et al. [2006] proposed the use of loop motifs for the identification of protein function from its three-dimensional structure. These loops are responsible for connecting the elements of the secondary structure of a protein, and there is sufficient experimental evidence for these loops being involved in important protein functions such as tyrosine sulfation and prohormonal cleavage [Fetrow 1995]. This motivated Espadaler *et al* to investigate if the presence of loop motifs in a set of proteins indicated their enrichment with one or more GO functions. For this task, they identified several loop motifs from a a non-redundant set of proteins structures in the SCOP database, and obtained the most probable amino acid sequence patterns representing these motifs. Next, they identified 73 protein families in the Pfam database, whose protein signatures matched the sequence pattern of the different loop motifs identified, and investigated the enrichment of these families with certain GO functional classes. Indeed, a strong correlation was observed. The results on prediction of functions of unannotated proteins using the loop-derived patterns was also found to produce more accurate results than a BLAST sequence similarity search. This showed the utility of using loops, which are less studied as compared to α-helices and β-plates, for inferring protein function.

Overall, it can be seen that approaches which make use of structural surfaces achieve higher specificity since they are more aligned to known biological knowledge about ligand binding and functional sites. However, this also requires more computation since surfaces are also harder to model than the coordinates of individual residues. Hence, approaches in this category must address both considerations to be effective. However, the brighter side of the picture is that several features of a protein structure's surface that are related to protein function have been identified, that can be combined to enhance the effictiveness of each individual feature. Research in this direction is expected to yield accurate and biologically meaningful results.

4.3.4 *Learning-based Approaches.* We have seen, or will see, in other places in this survey, such as Sections 3, 7 and 9, that machine learning-based approaches have achieved great success in function prediction, because of the natural mapping of this problem to the problem of building classification models that have been studied extensively by the data mining and machine learning communities [Tan et al. 2005]. In particular, kernel- and SVM-based techniques have been shown to be very effective for the functional classification problem [Brown et al. 2000; Cai et al. 2003; Tsuda and Noble 2004], owing to their flexibility of modeling the similarity between two data objects. Maintaining this trend, some kernel-based techniques have also been proposed for predicting function from struc-

ture. Here also, the structure of a protein can be treated as its attribute-value vector and its functional class as the class label. The techniques discussed below adopt this perspective in various forms.

Kin et al. [2004] discuss a mathematically sound solution for using sequence similarity results to estimate a structural similarity kernel matrix, which can then be used for predicting functional class using an SVM. Specifically, a sequence kernel matrix K_B is estimated for the training set of proteins using the marginalized count kernel (MCK) [Tsuda et al. 2002]. Similarly, a structure kernel matrix K_I is estimated using the MATRAS program [Kawabata 2003]. However, K_I is incomplete, since structural data may not be available for all proteins in the data set. Hence, an EM-algorithm based algorithm is used to estimate K_I accurately, based on the hypothesis that such a matrix should minimize the Kullback-Leibler divergence between itself and K_B, i.e. $KL(K_B, K_I)$ should be minimum [Amari 1995]. In experiments on three classes from the SCOP database [Andreeva et al. 2004], it was found that this solution works best for cases in which less than 50% entries of K_I are missing. Nevertheless, the ability to estimate structural similarity from sequence similarity can be useful for many applications.

A more application-oriented study is presented in [Dobson and Doig 2005], in which the authors used a simple SVM classifier for classifying the enzymes into the six classes at the top level of the Enzyme Classification [Webb 1992]. The focus of this study is on the definition of attributes that can be easily extracted from the sequence and structure of a protein using various methods, and subsequent supervised attribute subset selection using the Backward Elimination algorithm [Dash and Liu 1997]. With these subsets of attributes, and a one-against-one classification of the six classes, average accuracies of 35% and 60% were obtained when considering the first and the first two choices as the classification for a test protein.

In a recent paper, Wang and Scott [2005] propose three kernels for comparing two proteins structures are proposed, as discussed below:

—$\mathbf{K_{Pattern\_Sim}}(\mathbf{S}, \mathbf{T})$: This kernel defines the similarity between two substructures on the basis of the best corresponding pairs of amino acids in them and their proximity.

—$\mathbf{K_{Redox\_Func}}(\mathbf{S}, \mathbf{T})$: This kernel is a tailored version of the one above for the thiol/disulfide oxidoreductase proteins, based on the fact that all such proteins contain a $CxxC$ motif.

—$\mathbf{K_{3Dball}}(\mathbf{P_1}, \mathbf{P_2})$: The previous kernels were limited in that they were defined only with respect to amino acid positions in the structure. This kernel tries to relax this limitation by considering the protein as a set of balls of a given radius around each amino acid. The final similarity is a sum of the similarities between the best pairs of balls in the two structures.

These kernels model the similarity between two structures in different ways, and were used as the backbone of two classifiers in this paper, namely k-NN (k-nearest neighbor) and SVM. The validation was conducted in two separate experiments, one on ten superfamilies from SCOP [Andreeva et al. 2004], and the other on 21 thiol/disulfide oxidoreductase structures from PDB [Berman et al. 2000]. In the first experiment, using K_{3Dball}, k-NN showed a significantly higher true positive rate than SVM. In the second, K_{Redox_Func} gave the best results, since it was customized for the class being tested. In general, all the above kernels gave results better than using sequence-based HMMs or alignment methods such as DALI [Holm and Sander 1994] and CE [Shindyalov and Bourne 1998]. Thus, in

addition to providing three novel kernels for comparing protein structures, this study also suggested the superiority of the k-NN classifier over SVM for functional classification using protein structure.

Finally, Bandyopadhyay et al. [2006] have proposed a novel approach for finding structural templates in functional families, using techniques from the area of frequent subgraph mining, which falls within the larger field of association analysis in data mining [Tan et al. 2005]. Here, using a previously published technique, the three-dimensional structures of proteins within a SCOP family are converted in a graph representation. Next, using the fast frequent subgraph mining algorithm [Huan et al. 2005], several frequently occuring subgraphs in this set of graphs are derived. Additional selection of these subgraphs is also performed so as to ensure their statistical significance for this family, by comparing their frequency to the frequency of their occurrence in the entire set of proteins across all families. The resultant set of subgraphs are expected to be functionally important substructures for the concerned family, and thus can be used for the predicting which families an unannotated protein belongs to by matching them with its structure. Indeed, in an evaluation of 442 novel proteins that were added to 94 SCOP families, this approach could automatically discover the true family assignments of as many as 71% of these proteins, while the corresponding number for BLAST was only 53%. Overall, this study showed the utility of association mining for analysis of protein structure, due to the computationally efficient methods available and the biological relevance of the substructures derived.

The above discussions show that data mining techniques, such as classification, kernel estimation and association analysis, have great potential for functional classification using structural data. However, the main hindrance for this potentially accurate method is the not-so-extensive availability of protein structures. Once more data is available, the performance of these approaches is expected to surpass that of approaches in other categories.

4.4 Discussion

The previous sections discussed various perspectives on the prediction of function from structure, and the ways these perspectives have been formulated in various approaches. Some systems, such as ProFunc [Laskowski et al. 2003; 2005] have now started integrating the most popular of these approaches in order to infer a consensus annotation for the query protein. Thus, with further research in this field, more successful results are expected in the near future.

5. GENOMIC SEQUENCES

5.1 Introduction

The basic hereditary information about an organism is encoded in DNA molecules, which are dominantly organized as chromosomes in the cell, and are also found in the mitochondria of the cell to some extent. This set of chromosomal and mitochondrial DNA constitutes the genome of an organism. DNA itself is typically a double stranded molecule, where one of the strands is constituted of four characters, namely A, T, C and G, which denote the four nucleotides adenosine, guanine, cytosine and thymine, and other strand is complimentary to the first, owing to the complimentarity of the $A-C$ and $T-G$ nucleotide pairs. An illustration of a DNA molecular is shown in Figure 10.

Genomes contain genes and non-coding regions, both of which can be represented as strings of the four characters A, T, C and G. Proteins are synthesized from genes through

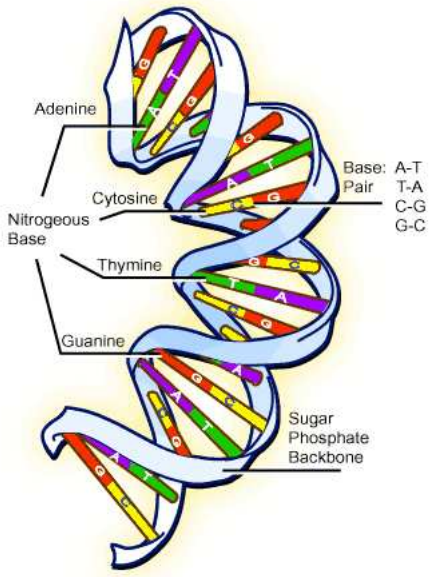


Fig. 10. Illustration of a DNA molecule (Figure taken from [Secko 2007])

a process consisting of two steps, namely transcription and translation. The genomes of eukaryotic organisms are typically several million base pairs in length, and typically contain several thousand genes. For instance, the yeast genome is 14 million base pairs in length, and contains about 6000 genes, while the human genome is over 3 billion base pairs long, and contains over 20000 genes. Also, the positions of these genes in the genome are typically known, thus providing information about the context of a gene in an organism's genome.

The genomes of nearly a thousand organisms have been sequenced till date, and hundreds others are now in progress. Several public data sources have been established, which make the public access to these sequences very convenient. The most prominent of these databases are the NCBI Entrez Genome database [Tatusova et al. 1999], the Genome Sequence DataBase of the National Center for Genome Resources [Harger et al. 1998] and the Genome Sequencing Project of the Sanger Institute[9]. Several useful tools have also been developed for visualizing and analyzing these large databases, such as the UCSC Genome Browser [Hinrichs et al. 2006] and databases such as GenBank [Benson et al. 2004] that organise these genomes into their constituent genes, and accompanying information. This wide availability of genome sequence data has spurred research in the field of genomic context-based protein function prediction.

5.2 Existing Approaches

This section discusses some fundamental ideas that have arisen from the new genome resource, and several approaches that have utilized these ideas for function prediction. However, it should be noted that in this domain, most of the studies fall in the field of

[9] http://www.sanger.ac.uk/Projects/

comparative genomics [Marcotte 2000], since they involve the comparison of genes across several genomes. As a result, the primary form of results derived from these studies is that of functional associations between genes or proteins rather than annotations for individual proteins. Also, it must be remarked that the approaches in this field are often justified on the basis of evolutionary mechanisms, since the availability of the complete genomes for a wide variety of organisms offers an insight into the ways in which genes may have evolved from each other [Koonin and Galperin 2002].

Several approaches have been proposed to accomplish the target of deriving functional associations from genomic data, and possible function prediction subsequently. These approaches largely fall into one of the following three categories [Marcotte 2000]:

—**Genome-wide homology-based annotation transfer**: This category consists simply of the use of larger databases for searching proteins homologous to the query proteins, and the transfer of functional annotation from the closest results.

—**Gene neighborhood- or gene order-based approaches**: These approaches are based on the hypothesis that proteins, whose corresponding genes are located "close" to each other in multiple genomes, are expected to interact functionally. This hypothesis is supported by the concept of an *operon*, and its relevance to protein function [Salgado et al. 2000].

—**Gene fusion-based approaches**: These approaches attempt to discover pairs or sets of genes in one genome that are merged to form a single gene in another genome. The underlying hypothesis here is that these sets of genes are functionally related, and is supported by biochemical and structural evidence [Marcotte et al. 1999].

As can be seen, approaches in the latter two categories exploit genomic context, i.e. the location of a gene on the genome with respect to that of other genes [Huynen et al. 2000].

Yet another category of approaches that make heavy use of multiple genomes simultaneously are those based on the concept of phylogenetic profiles. However, because of the distinct evolutionary underpinnings of these approaches, and since much more work has been done in this category as compared to the others, it is discussed separately in Section 6. Hence, the following subsections discuss the details of the categories of approaches listed above, followed by a discussion of the comparative advantages and disadvantages of each of them against the other.

5.2.1 *Genome-wide homology-based annotation transfer.* The most immediate impact of large-scale genome sequencing projects has been the wider application of existing sequence-homology based approaches for functional annotation transfer. The availability of complete genomes of many organisms led to the creation of databases of gene sequences, such as GenBank [Benson et al. 2004], and protein sequences, such as SwissProt [Boeckmann et al. 2003] and PIR [Wu et al. 2003]. These databases also contained the available experimental knowledge about some of these genes and proteins in the form of keywords and descriptions. Thus, it was straightforward to use sequence search systems, such as PSI-BLAST [Altschul et al. 1997], for searching homologous sequences in these large databases. Thus, functional information could be obtained from a larger number of proteins and organisms, and this became a popular method for the prediction of protein function.

Another very significant development in this category was the creation of the database of Clusters of Orthologous Genes (COGs) [Tatusov et al. 1997]. This study aimed at

constructing gene families by identifying orthologous genes across several genomes. The underlying idea was that orthologs, which are genes in different organisms that evolved from a common ancestor through speciation events, are expected to perform the same function, and thus each COG would represent a functionally coherent group of genes. This idea was implemented using a simple clustering scheme that ensured that there was a high pairwise similarity between any two genes in the same cluster. Upon validation, it was found that, for most of the 720 COGs so constructed, a specific cellular function could be deduced from available functional knowledge and high levels of sequence similarity. Also, the constituent genes of many clusters showed very consistent phylogenetic profiles, indicating a common evolutionary origin of the cluster. Thus, the COG database [Tatusov et al. 2003] marked an important starting point for functional genomics, and have been used for the validation of other function prediction algorithms [Zheng et al. 2002]. However, one important problem with this database was that most of the functional categories were split into several clusters, which may complicate their use in further studies.

5.2.2 *Approaches exploiting gene neighborhood.* One of the most basic signals offered by the genomic data is the relative positioning of genes on a genome. It may be hypothesized that two or more proteins, whose correponding genes are "close" to each other on a genome, are functionally related [Dandekar et al. 1998]. This hypothesis finds evidence in the well-known concept of an *operon*, which is a contiguous portion of the DNA that includes an operator, a common promoter, and one or more genes that are expressed as a unit to produce messenger RNA (mRNA)[10]. Interestingly, proximity between the constituent genes is the dominant strategy for identifying operons in a genome [Salgado et al. 2000]. Thus, this is a viable strategy for inferring functional associations between genes and their corresponding proteins.

Dandekar et al. [1998] were the first group to explore the use of this signal for discovering pairs of proteins that are expected to interact functionally. In this study, they found all the pairs of genes in a set of nine genomes, such that the two genes were close to each other and occurred in the same order in at least three genomes. Among the small number of gene pairs so discovered, at least 75% were known to interact physically, and the others also represented potential interactions. Thus, even though the scope of this strategy was not very broad, it made a strong case for the gene neighborhood idea.

The first extensive study based on this idea was reported by Overbeek et al. [1999a], where they used it to infer functional coupling between genes in 24 genomes, and then conducted an extended analysis of the approach and its results [Overbeek et al. 1999b]. Their analysis is based on the concept of a Pair of Close Bidirectional Best Hits (PCBBHs), which are essentially a pair of gene pairs (Xa, Ya) and (Xb, Yb) such that Xa and Ya in genome Ga are orthologous to Xb and Yb in genome Gb, respectively, and Xa and Ya and Xb and Yb are close to each other in Ga and Gb respectively. This formulation of a PCBBH is depicted graphically in Figure 11. Notably, the order to genes is not significant in a PCBBH, thus ensuring a larger coverage for this approach as compared to that of [Dandekar et al. 1998]. The occurrence of such a PCBBH indicates that evolution has preferred to keep Xa and Ya close in the genome since they are expected to interact functionally, and the strength of this deduction increases with the number of PCBBHs that a pair of genes participates in. Thus, a score is calculated for each PCBBH on the basis

[10] http://en.wikipedia.org/wiki/Operon

of the evolutionary distance between the original genome containing the PCBBH and the genome in which a match is found. This score denotes the strength of the functional relationship between the two genes, and the predictions are those PCBBHs whose score is above a pre-defined threshold. However, though this was a feasible approach, an important weakness of the assumption was that gene proximity is not sufficient for functional coupling. As a result, a precision of less than 35% was obtained when this idea was applied to a set of 31 genomes [Overbeek et al. 1999b].

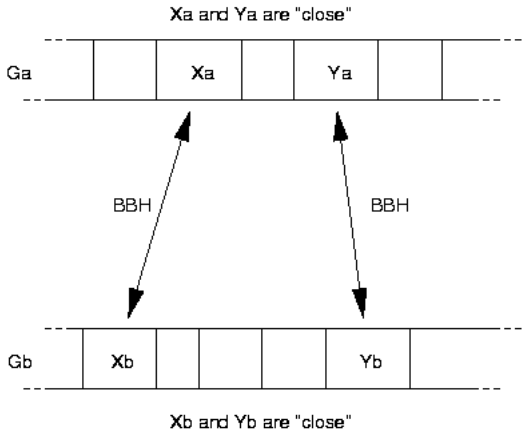


Fig. 11. A graphical illustration of the concept of PCBBHs (Taken from [Overbeek et al. 1999a])

In a complementary study [Korbel et al. 2004], it was hypothesized that neighboring genes which are bidirectionally transcribed, i.e. whose transcription start sites are very close, and whose direction of transcription are opposite of each other, may be functionally associated. The motivation for this hypothesis came from a study of the human genome, which showed that for certain classes of genes, such as the DNA repair genes, bidirectionally transcribed pairs of genes are functionally linked [Adachi and Lieber 2002]. From experiments on the *E. Coli* genome, the observed accuracy and coverage of this method was not very high, since examples of such bidirectionally transcribed gene pairs are not too common, except for few classes [Adachi and Lieber 2002].

An innovative system that tries to relax the proximity definition of [Overbeek et al. 1999a] is SNAPper [Kolesova et al. 2001; 2002]. This system builds an SN-graph for the genes in the given set of genomes by iteratively finding similarity or orthology (S) and neighborhood (N) relations between genes and adding corresponding an edges for each relationship to the graph. The authors' hypothesis says that all genes involved in a cycle in this graph, named as an SN-cycle, are functionally related. Intuitively, this hypothesis is viable since such a cycle corresponds to groups of close genes which are conserved across several genomes, and thus, these genes may be related functionally. Formal experimental verification of the claim was done by measuring two coefficients, K_p and K_f, based on the KEGG [Kanehisa et al. 2004] and FunCat [Ruepp et al. 2004] databases respectively, for all the SN-cycles found. These coefficients respectively estimate the fraction of an

SN-cycle involved in the same metabolic pathway and functional class respectively. This validation showed that the claim was reasonably valid for a good fraction of the cycles. However, the results were better for the K_p as compared to K_f, indicating that SNAPper is more effective at reconstructing metabolic pathways than directly predicting functions for unannotated proteins.

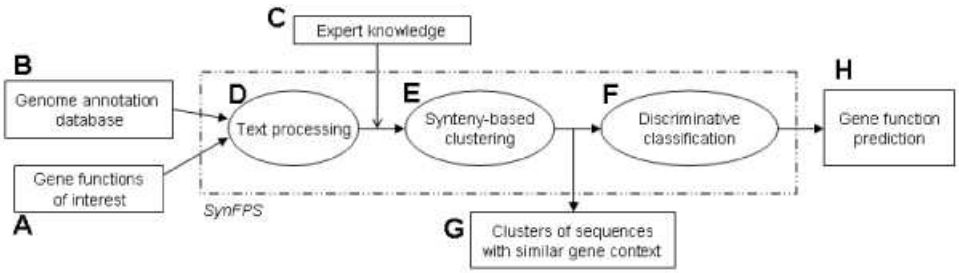


Fig. 12. Data mining approach to inferring gene function from gene order information

Finally, Li et al. [2007] have presented an innovative approach involving the application of data mining techniques to the inference of gene function using genomic context information. This is useful since these techniques are better able to capture the variability in the input data, and are robust to noise, issues that initial ad-hoc approaches for this problem did not address very robustly. Figure 12 shows an overview of the approach adopted by Li *et al*. The first step of this approach is the selection of genes in the given set of genomes by searching in the available literature for genes that are known to be annotated with the functions of interest. Once this step is complete, two data mining steps are applied to predict gene function:

(1) **Clustering**: In the first step, the complete set of genomes is clustered into clusters, such that the genomes in each cluster has similar patterns of distances between the genes identified in the first step. This helps identify groups of evolutionarily close genomes, which are most useful for inferring gene function.

(2) **Classification**: Next, a classification model is built for each functional class of interest within each gene cluster. Several biological features relevant to protein function, such as amino acid composition, van der Waals volume, hydrophobicity and polarity, are extracted from the sequence of the all the gene products. Also, in order to obtain a highly discriminative classifier, the positive examples for each function are chosen to be the genes identified in the first step, while the negative examples are composed of genes in the neighborhood of the positive examples, which are not yet annotated with that function. With this setup, an SVM classifier is constructed for each functional class in each cluster, and is used for predicting the functions of currently unannotated genes.

The overall system, named SynFPS (Synteny-based Function Prediction System), is tested on 296 bacteriophage genomes for nine functional classes that are relevant for these genomes. A high accuracy of 80% is achieved in cross-validation experiments, which validates the utility of using data mining and machine learning methods for analyzing genomic

context data. Also, several novel functional annotations are obtained, that are otherwise impossible to find using simple sequence similarity techniques.

It is evident from the above discussions that gene neighborhood has been defined in different ways by different groups. However, it is interesting to find that variations based on this simple concept have been able to find functional links between genes more accurately [von Mering et al. 2003; Mellor et al. 2002; Bowers et al. 2004] than some methods, such as gene fusion and phylogenetic profiles, that are discussed in the next few sections.

5.2.3 *Approaches exploiting gene fusion.* Gene fusion is another very innovative method for exploiting the relative gene positioning on a genome, and was proposed for the first time by Marcotte et al. [1999]. This idea simply states that if two separate genes in one genome are merged, or "fused", as a single gene in another, then these genes are expected to be functionally related. Interestingly, this hypothesis is backed by very strong biological reasoning [Marcotte et al. 1999]:

—The fusion of two genes greatly reduces their entropy of dissociation [Erickson 1989], which indicates that they may have existed very stably as two domains of a polypeptide in another organism, and consequently evolved into independent genes in a descendant organism. A possible mechansim for the evolution of protein-protein interactions is also proposed in [Marcotte et al. 1999] on the basis of this reasoning.

—At a structural level, protein-protein interfaces have strong similarity to interdomain interfaces within single protein molecules [Tsai and Nussinov 1996]. This basically implies that two separate proteins may interact in the same way as two domains interact within the same protein.

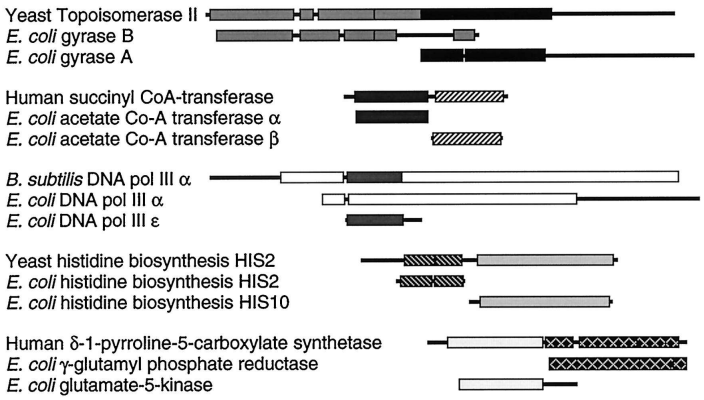


Fig. 13. Example of pairs of *E. coli* proteins predicted to interact functionally by the gene fusion method (Taken from [Marcotte et al. 1999])

In order to validate the hypothesis using real data, a rigorous procedure consisting of three independent tests was adopted by Marcotte et al. [1999], and was applied to 6809 pairs of non-homologous genes found in the *E. coli* genome by this method. These tests were based on SWISS-PROT keywords, the Database of Interacting Proteins [Xenarios et al. 2002] and phylogenetic profiles [Pellegrini et al. 1999], respectively. Results showed

that a very significant fraction of the pairs were actually reported to interact physically or functionally, thus demonstrating the practical efficacy of the method. Figure 13 shows five examples of pairs of *E. coli* proteins predicted to interact functionally by the gene fusion method. Also, owing to the information hidden in the fused protein sequence, and its utility for function prediciton, this approach has also been given the interesting name of the Rosetta Stone method[11].

The scope of the above study was expanded by Yanai et al. [2001], who systematically applied the Rosetta Stone method to 30 microbial genomes. Upon validation of the predicted functional links using the COGs database [Tatusov et al. 2003], very high average sensitivity and specificity [Tan et al. 2005] of 72% and 90% respectively were observed, which is significantly higher than those observed for a random distribution of fusion links. A parallel study working with a reference set of 24 genomes reported similar results [Enright and Ouzounis 2001]. Together, these studies illustrated both the wide coverage and high accuracy of this method, which may be extended easily to function prediction for individual genes.

Finally, Marcotte and Marcotte [2002] presented some very interesting enhancements to the basic fusion method. First, they argued that using orthology as the basis for finding fusions lowers the coverage significantly, and does not necessarily discover accurate linkages. Hence, in this study, the more general concept of homology was used for finding fusion, in order to increase the coverage, and hence the likelihood of finding correct functional linkages. Next, a scoring function for the discovered fusions, was formulated, based on the hypergeometric distribution. This score reflects the probability of the chance occurrence of a given number of fusion events between a given pair of genes. Thus, the smaller this score, the more reliable is a functional linkage discovered by this method. Also, the functional similarity of two genes was found to be linearly related to the log of their association score, thus showing that the scoring function is indeed a robust estimator of the reliability of a functional association found by the gene fusion method. In essence, this study provided a framework for the design of algorithms based on gene fusion, and more such algorithms are expected in the future.

5.3 Comparison and Assimilation of the Approaches

The previous section presented details of the two most common categories of approaches in functional genomics: neighborhood- or order-based, and fusion-based. An obvious question to ask then is: Given the same initial set of genomes, which kind of approach is the most effective at finding functional associations? Here, we discuss some studies that have tried to answer this question. It should be noted that these studies consider phylogenetic profiles also as a functional genomics method, but for reasons mentioned earlier, we have covered these in detail elsewhere (Section 6).

Huynen et al. [2000] report a comparison of the approaches for the genes in the *M. genitalium* genome, which has been used in many other benchmarking studies because of its favourable characteristics [Teichmann et al. 1999; Brenner 1999; Hutchison III et al. 1999]. The results of the analysis of the the functional linkages detected in this genome using the above methods are shown graphically in Figure 14, which shows the distribution of these linkages among seven types of interations that may provide a clue to a protein's function. It is evident from these distributions that for all the approaches that a large fraction of

[11] http://en.wikipedia.org/wiki/Rosetta_Stone#Use_as_metaphor

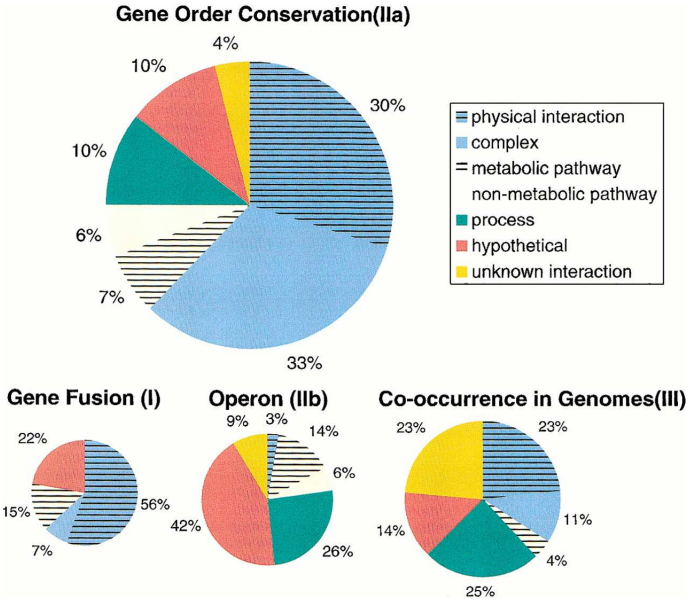


Fig. 14. Distribution of the links discovered by various genomic context methods (Taken from [Huynen et al. 2000]). Note that methods IIa and IIb refer individually to the approaches presented in [Dandekar et al. 1998] and [Overbeek et al. 1999b] respectively, and collectively refer to the gene neighborhood method. Also, method III refers to the phylogenetic profile approach.

the detected interactions are functionally meaningful, since they either represent physical interactions, or co-membership in a complex, a metabolic/non-metabolic pathway or a biological process. Thus, this study qualitatively justified the potential of these approaches for discovering functional links between proteins and protein function.

A very useful development in the field of functional genomics has been the development of databases which collect and compare functional associations discovered by each of these approaches. Table VI captures the various characteristics of these databases. These databases are available freely to the research community for their use.

Name	**Reference**	**# Genomes**	Validation Method(s)
STRING	[von Mering et al. 2003]	89	KEGG [Kanehisa et al. 2004]
Prolinks	[Bowers et al. 2004]	83	COG [Tatusov et al. 2003]
Predictome	[Mellor et al. 2002]	44	KEGG [Kanehisa et al. 2004], COG [Tatusov et al. 2003], GeneQuiz [Andrade et al. 1999]
Phydbac "Gene Function Predictor"	[Enault et al. 2005]	*E. coli*	COG [Tatusov et al. 2003]

Table VI. Popular databases of functional links discovered by functional genomics methods

A few words about these databases are in order concerning the strategies they adopt to detect functional associations. Interestingly, all of them used the gene order (neighborhood), gene fusion and phylogenetic profile approaches. However, while the others were

based on traditional implementations of these methods, Phydbac "Gene Function Predictor" (PGFP) [Enault et al. 2005] was built on top of Phybac database [Enault et al. 2003b], which is based on the idea of consensus phylogenetic profiles (CPPs), a more sensitive version of the basic phylogenetic profiles (PPs). However, it should also be noted from the third column of Table VI that while STRING, Prolinks and Predictome derive their results from a significant number of genomes, while those of PGFP are based only the *E. coli* genome. Thus, the results of PGFP are less reliable than the others.

Finally, and most importantly, as mentioned in the last column of Table VI, these databases also adopted certain methodologies for the validation of the collected links. The most common of these methodologies are COGs [Tatusov et al. 2003] which is a functional classification, and KEGG [Kanehisa et al. 2004], which indicates that the detected functional associations may be used to reconstruct metabolic pathways. Interestingly, all the databases reported that **gene neighborhood** or **gene order** [Dandekar et al. 1998; Overbeek et al. 1999b] was the most accurate approach for both the tasks, which is a significant result and can be used to guide further research in this field.

6. PHYLOGENETIC DATA

6.1 Introduction

The biological species existing today have evolved from primitive forms of life over millions of years [Darwin 1909], and this process of evolution continues today. The changes in the physiologies of different organisms have been driven by the changes at the cellular level, which include the adoption and surrender of functions by proteins due to the changes in the genes encoding them. Thus, it is essential to include the evolutionary perspective in any complete understanding of protein function. As a result, several approaches for predicting protein function using evolution-based data have recently been proposed. The two most common forms of this data are known as phylogenetic profiles and phylogenetic trees, and the field of biology that deals with the evolutionary relationships among living organisms is also known as phylogenetics [Bittar and Sonderegger 2004].

The phylogenetic profile of a protein is (generally) a binary vector whose length is the number of available genomes. The vector contains a 1 in the i^{th} position if the i^{th} genome contains a homologue of the corresponding gene, else a 0. Some variations of these vectors use real numbers that reflect the extent of similarity between the original gene and the best match in the genome being searched, instead of 0s and 1s. Thus, these profiles provide a way of capturing the evolution of genes across various organisms. This information becomes useful for functional genomics when seen in the light of the phenomenon of *speciation*, which is the evolutionary mechanism by which new species are created from currently existing ones [Coyne and Orr 2004]. Now, it may be hypothesized that proteins which interact functionally are under strong selective pressure, and thus their corresponding genes are inherited across several genomes during speciation events [Gaasterland and Ragan 1998]. Phylogenetic profiles are a powerful mathematical way of modeling this phenomenon, and thus offer a very innovative method for inferring functional associations between proteins, since functionally associated proteins are expected to have very similar phylogenetic profiles [Pellegrini et al. 1999]. This is the basic assumption made by all the approaches for function prediction on the basis of phylogenetic profiles. In addition, it can be seen that the construction of these profiles involves running a simple BLAST [Altschul et al. 1997] search again well known databases of completed genomes mentioned in Sec-

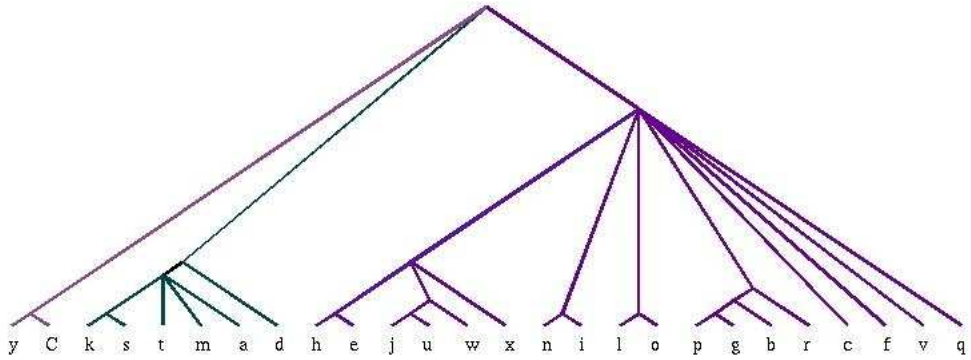


Fig. 15. Phylogenetic tree constructed for twenty six completed genomes (each represented by a letter) constituting the PhylProM database [Thoren 2000]

tion 5, such as TIGR and NCBI, thus enhancing the appeal of phylogenetic profiles further.

In several other studies, a more extensive representation of evolutionary knowledge is used [Bittar and Sonderegger 2004]. This representation is known as a phylogenetic tree [Baldauf 2003], which is a standard tree with respect to the graph theoretical definition, but whose nodes and branches carry special meaning. Its leaves correspond to the organisms that were used to build the tree. The internal nodes denote the hypothetical last common ancestor (LCA) of all its descendents and the branches represent *has evolved from* relationship. This indicates the complexity of the problem of constructing phylogenetic trees from genomes. Most available tools for this construction, such as PHYLIP [Felsenstein 1989], PAML [Yang 1997] and BETE [Sjolander 1997], apply various data mining and probabilistic methods for the task, and are mostly based on a hierarchical clustering of the given set of sequences from different organisms. One such reconstructed phylogenetic tree for twenty four fully sequenced organisms is shown in Figure 15.

It is easy to see that phylogenetic trees embody a much richer source of knowledge than phylogenetic profiles since the latter are constructed only on the basis of the leaf nodes of the former, hence ignoring the hierarchical structure of the evolutionary knowledge. This additional knowledge of the internal nodes tree can be used to extract further information about the pattern of evolution of a set of proteins. In addition, this knowledge is multi-level in nature, since the information extracted depends on the depth of the node in the tree. Thus, phylogenetic trees, if constructed accurately, can provide strictly richer information than simple profiles. Still, both these forms of phylogenetic data together constitute a very rich pool of knowledge about evolution that can be utilized effectively for the prediction of protein function [Kensche et al. 2007].

6.2 Existing Approaches

The evolution of one species of organisms from another has been an active area of research in biology [Darwin 1909], and, of late, has come to be known as the field of *phylogenetics*. Recently, several studies have been conducted for utilizing and predicting the implications of evolution at the molecular and cellular levels [Bittar and Sonderegger 2004]. Of most interest to us are the studies that try to uncover gene/protein functions and functional link-

ages using phylogenetic data such as profiles and trees. This section describes several such studies. However, before we embark on this discussion, it would be useful to categorize these studies into three categories:

—**Approaches Using Phylogenetic Profiles**: This category consists of a large number of approaches that are based on the hypothesis that proteins with similar phylogenetic profiles are functionally related. Thus, most of the approaches here are comparative in nature, and model this hypothesis using ways to measure the similarity of two profiles.

—**Approaches Using Phylogenetic Trees**: As noted earlier, phylogenetic trees embody a richer knowledge of genetic evolution than simple profiles. Thus, a recent category of approaches have started using this knowledge to predict function. Most of these approaches use various data mining and machine learning approaches to achieve this task, and produce better results than those based only on profiles.

—**Hybrid Approaches**: Recently, some approaches have started using SVM-based techniques to combine the two forms of evolutionary knowledge in phylogenetic profiles and trees. This category corresponds to these approaches.

The following sections discuss the approaches in each of these categories in detail, and also propose ideas for enhancing these approaches further.

6.2.1 *Approaches Using Phylogenetic Profiles.* The first study to analyze protein function using phylogenetic profiles was presented by Pellegrini et al. [1999]. The underlying hypothesis was that proteins that function together in a pathway or a protein complex, are likely to have a similar evolutionary path. To test this hypothesis, phylogenetic profiles were constructed from the fully sequenced genomes of sixteen organisms, other than *E. coli*, which was the model organism used in this study. Using three *E. coli* proteins, $RL7$, $FlgL$ and $His5$, it was verified that proteins with profiles differing by at most one out of sixteen bits are indeed functionally related as per the SwissProt annotation. Similar encouraging results were derived from the EcoCyc database [Keseler et al. 2005] of metabolic pathways. This was a seminal study in this area, and it opened the floodgates for protein function prediction using phlogenetic profiles.

A further examination of the feasibility of prediction function from phylogenetic profiles is reported by Liberles et al. [2002], wherein many important conclusions were made. First and foremost, phylogenetic profiles were shown to perform much better than homology-based approaches for the SwissProt keyword recovery task [Marcotte et al. 1999]. With respect to the same measure, it was also concluded that profiles that are constructed using a larger number of genomes are more informative for the function prediction task. Finally, another type of profile, known as an inverse phylogenetic profile, was proposed to model those gene families whose members may have undergone replacement by each other during evolution, and thus occur in disjoint sets of organisms. In many case, this replacement occurs since the genes perform the same function, and thus become redundant, are eventually lost in a genome. Thus, genes which have complementary phylogenetic profiles may belong to the same functional family. Indeed, this was shown to be true for three functional classes, DNA-directed DNA polymerases, DNA repair proteins and Isomerases. These results and other useful phylogenetic data has been compiled into the PhylProM database by the authors [Thoren 2000].

Wu et al. [2003] advocate the use of more general measures of similarity for pairs of phylogenetic profiles. Three popularly used measures of similarity [Tan et al. 2005], namely

the Hamming distance (D), Peason's Correlation Coefficient (r) and mutual information (MI) are evaluated for this task. It is concluded from the analysis that, although the three measures are strongly related to each other, MI is the most informative measure of profile similarity for inferring functional relationship between two proteins. This relationship is judged by membership of the proteins in the same metabolic pathway in KEGG [Kanehisa et al. 2004]. In addition, it is argued that proteins with complimentary profiles may suggest that they are functionally similar, which is likely to be missed if exact similarity of profiles is required.

The idea of relaxing phylogenetic profiles is carried forward by Enault et al. [2003a], particularly for the annotation of bacterial genomes. The modification suggested here is to use the normalized BLAST score [Altschul et al. 1997] denoting the best match for a protein in a genome, instead of using a 0 or 1. Annotation is carried out by finding the statistically dominant class of the MultiFun database [Serres and Riley 2000] in the neighborhood of a protein induced using cosine similarity. Results better than [Pellegrini et al. 1999] are shown, thus showing the potetial of real-valued phylogenetic profiles. This annotation procedure is available via the website of the Phydbac database [Enault et al. 2003b]. In Phydbac2 [Enault et al. 2004], a more recent version of the original database, the annotation procedure is strengthened further by combining predictions based on other comparative genomics methods discussed in Section 5. Here, a consensus phylogenetic profile (CPP) is constructed for a protein by incorporating the profiles of other genes that are detected to be "close" to this protein by gene neighborhood or gene fusion methods described in Sections 5.2.2 and 5.2.3.

Two modifications of phylogenetic profiles are proposed in [Bilu and Linial 2002] in order to improve the inference of functional linkages. The first modifications is to use partially complete genomes also for the construction of phylogenetic profiles, so as to enhance the knowledge contained in these profiles. This idea, though useful, is less relevant today when several hundred genomes have been sequenced, as compared to the total of 85 genomes used in this study. The more important contribution of this paper were two modifications that are suggested for the basic Hamming distance measure. These modifications are based on the following novel ideas:

(1) The distance between two profiles that are constructed using a larger number of genomes should be assigned greater significance than that between profiles constructed using fewer number of genomes.

(2) This distance should also take into account the evolutionary history of the proteins. More specifically, the column in the profile corresponding to a genome that is evolutionarily distant from the one containing the query gene should be assigned higher weight in distance calculations.

Experiments to test these ideas are carried out for all the *E. coli* enzymes and functional relationship is measured by co-occurrence in a pathway of the KEGG database [Kanehisa et al. 2004]. Results derived using two tests designed specifically for this study suggest that the first modified similarity measure has a much more positive impact on the performance of the linkage annotation procedure than the second. However, the reconstruction of metabolic pathways using the proposed similarity measure is not very encouraging.

Date and Marcotte [2005] describe the PLEX (Protein Link EXplorer) system, that adopts an iterative strategy for searching proteins with similar phylogenetic profiles. This approach simply uses the results for one iteration of similarity search as input for the next

iteration, and thus iteratively searches for proteins with phylogenetic profiles similar to that of the query protein. By combining the predicted functional links with gene fusion links and gene neighborhood links, PLEX was able to reconstruct two important protein systems in *M. tuberculosis*, namely the urease enzyme complex and the isoprenoid biosynthesis pathway.

Lastly, an innovative method for improving the predictions from simple phylogenetic profiles is proposed by Zheng et al. [2002]. The hypothesis here is that, due to selective pressure, pairs of genes close to each other on a genome are expected to be preserved across genomes, since such genes are expected to be functionally related [Overbeek et al. 1999a; 1999b]. This hypothesis is modeled by finding several pairs of *E. coli* genes that are close to each other and constructing a single phylogenetic profile for each of these pairs. This profile simply records whether the pair is found to be close in each genome or not. Finally, all the pairs with exactly the same profiles are collected, and the functional coherence of these clusters is tested with respect to COGs functional categories [Tatusov et al. 2003]. Quantitative evaluation using the purity and Jaccard coefficient show the superiority of the method over single gene phylogenetic profiles. Another conclusion of this study was that mutual information is the most appropriate measure for the similarity of two profiles. This agrees with the general perception that proteins occurring in very few or too many organisms are not very informative for predicting other proteins' functions. However, one surprising finding of this study was that using a larger set of organisms for constructing the profiles does not necessary improve prediction. The reason for this is hypothesized to be the simultaneous amplification of both the coevolution and noise signals when more genomes are used. This finding needs to be investigated further, since it may have implications for the number and choice of genomes used to construct these profiles.

Having covered a wide range of approaches that attempt to predict protein function using phylogenetic profiles, it is important to observe that none of the above approaches involve significant application of techniques from the field of data mining [Tan et al. 2005], which has been extensively made use of for the analysis of other types of biological data, such as protein sequences (Section 3) and gene expression profiles (Section 7). Data mining has significant potential for the analysis of phlogenetic profiles, since these profiles are binary vectors, to which an entire field of data mining known as *association analysis* [Tan et al. 2005] has been dedicated. Two ideas that demonstrate how the application of association analysis principles can aid in the use of phylogenetic profiles are as follows:

—Several studies have concluded that mutual information (MI) is the most appropriate similarity measure for phylogenetic profiles [Wu et al. 2003; Date and Marcotte 2003; Zheng et al. 2002]. These studies directly map to the data mining problem of finding the right objective measure for association patterns. Tan et al. [2004] have discussed several objective measures in detail, including five properties that a measure should posess. Some of these properties are null invariance, symmetry under variable permutation and invariance to scaling. It turns out that measures such as odd ratio and its normalized versions Yule's Q and Yule's Y, the Piatetsky-Shapiro measure and collective strength, satisfy more of these properties than MI. Thus, it may be useful to investigate these similarity measures also, in order to improve the useful results already obtained using MI as the measure.

—As noted earlier, a set of phylogenetic profiles can be treated as a binary matrix, which maps directly to the concept of a *market basket* in association analysis [Tan et al. 2005].

Many algorithms have been designed for accurately and efficiently extracting frequently occurring and meaningful patterns from these baskets. In the domain of phylogenetic analysis, these patterns could reveal important biological knowledge such as groups of genes following the same evolutionary rate and genes evolving through the same set of organisms. Such knowledge could aid in the construction of the global evolutionary tree, which is one of the most important goals of biological research.

These ideas demonstrate the utility of data mining in general, and association analysis in particular, for phylogenetic analysis. However, no such study has been reported in the literature yet.

Finally, as an ending note, the following high-level conclusions can be drawn about the use of the phylogenetic profile paradigm on the basis of the approaches discussed above:

(1) In general, the use of a large number of genomes for constructing phylogenetic profiles improves the performance of function prediction methods [Bilu and Linial 2002; Liberles et al. 2002].
(2) Real-valued profiles offer more flexibity and hence more reliability in function prediction [Wu et al. 2003; Enault et al. 2003a].
(3) Mutual information (MI) is the most appropriate similarity measure for phylogenetic profiles [Wu et al. 2003; Date and Marcotte 2003; Zheng et al. 2002].
(4) Phylogenetic profiles are effective at the task of the reconstruction of metabolic pathways, since the proteins appearing together in a pathway are highly likely to evolve together [Pellegrini et al. 1999; Bilu and Linial 2002].

These conclusions are expected to be useful for future research involving phylogenetic profiles. Nevertheless, the use of phylogenetic profiles has already helped uncover useful functional information about several proteins. For a detailed list of proteins for which predictions were made using their phylogenetic profiles, and were experimentally verified, see [Kensche et al. 2007].

6.2.2 *Approaches Using Phylogenetic Trees.* As was discussed in Section 6.2, approaches that use phylogenetic trees are far fewer than those that use phylogenetic profiles. Two major factors contribute to this. First, trees are more difficult to use than simple profiles, and hence demand more intricate algorithms, as justified in Section 6.1. Second, the precise evolutionary tree for a set of organisms is not known apriori and is constructed from their genomic sequences using systems such as PHYLIP [Felsenstein 1989]. This creates to an additional source of error. However, if these issues are handled appropriately, more reliable predictions can be made about protein function and/or functional linkage, as shown by the following approaches.

In an early theoretical study of the use phylogenomic data [Eisen 1998], a possible approach for finding the functions of uncharacterized proteins from phylogenetic trees is outlined. In this approach, a phylogenetic tree is constructed for the protein under consideration by finding their homologs, and using one of the known methods for tree construction. Next, events such as gene duplication and gene speciation may be identified on this tree, and using the structure of the tree, functional predictions can be made. The author also identified some conditions under which this method is expected to perform better than homoogy-based methods, such as the following:

—Functional change between homologs during evolution.

—Variation of the rate of functional change during evolution.

—Variation of the rate at which gene duplication occurs.

Soon after the publication of the above arguments, it was quantitatively shown by a subsequent study [Doerks et al. 1998] that the inclusion of phylogenetic trees led to an improvement in the annotation of the currently uncharacterized protein families (UPFs) in SWISS-PROT [Boeckmann et al. 2003]. Iterative BLAST searches were unable toannotate these diverse families, due to well-known problems with sequence homology-based approaches [Whisstock and Lesk 2003]. However, when phylogenetic trees were constructed for a set consisting of selected members of these UPFs and other characterized families, many of the unannotated proteins clustered amazingly well with proteins that performed the same function, thus providing a confident annotation for the former. Thus, this study quantitatively exhibited the potential of phylogenetic trees for providing accurate functional predictions.

Following this strategy, the earliest approaches attempted to identify functional interactions between proteins. Pazos and Valencia [2001] attempted to identify these these interactions at three possible levels, namely interactions between structural domains of proteins, between individual proteins and between all the proteins in a complete genome. This approach involved the derivation of phylogenetic similarity matrices from a multiple sequence alignment of the set of entities being studied, using the appraoch of Goh et al. [2000]. The following results were obtained from the analysis of these matrices:

(1) **Domains**: Nine of thirteen known interactions were identified [Pazos et al. 1997].

(2) **Proteins**: Again, Seven out of eight experimentally known interactions among Dandekar et al. [1998]'s test set of 53 *E. coli* proteins.

(3) **Genome**: Here, functional associations were sought between all the 4300 proteins in the *E. coli* genome, and 2700 strong interactions were proposed, some of which were previously well known, such as the ATP synthases α and β.

Thus, this approach showed both high accuracy and good coverage, in addition to demonstrating the applicability of the phylogenetic approach at various levels.

A more direct use of phylogenetic trees for classification is made by Qian et al. [2003], where a tree-based HMM (T-HMM [Qian and Goldstein 2003]) is learnt for the class of GPCR proteins. It is hard to construct a classifier for this class of proteins since they are very diverse at the sequence level and hence are hard to annotate using sequence homology-based methods. However, Qian et al. [2003] handle this problem by modeling the evolutionary history of this class by a phylogenetic-tree based HMM [Qian and Goldstein 2003]. This model essentially constructs a hidden Markov model at each node of the tree by using the multiple sequence alignment of the reconstructed sequences at its child nodes. Since it is unknown apriori which class and evolutionary stage a novel protein comes from, a score is calculated for each node of each tree, and the protein is classified as the class whose tree scores the highest. This model is suitable for diverse families such as GPCR, where the sequences are not similar and the members are at different evolutionary stages. Indeed, almost 99% accuracy is achieved for a set of 1749 GPCRs, which is impressive.

So far, the largest project involving phylogenetic analysis was undertaken by the Berkeley Phylogenomic Group, and a detailed review of the strategy developed by them for the inference of molecular function appears in [Sjolander 2004]. This multi-step strategy is implemented in the GTREE software and its various steps are listed in Figure 16 along

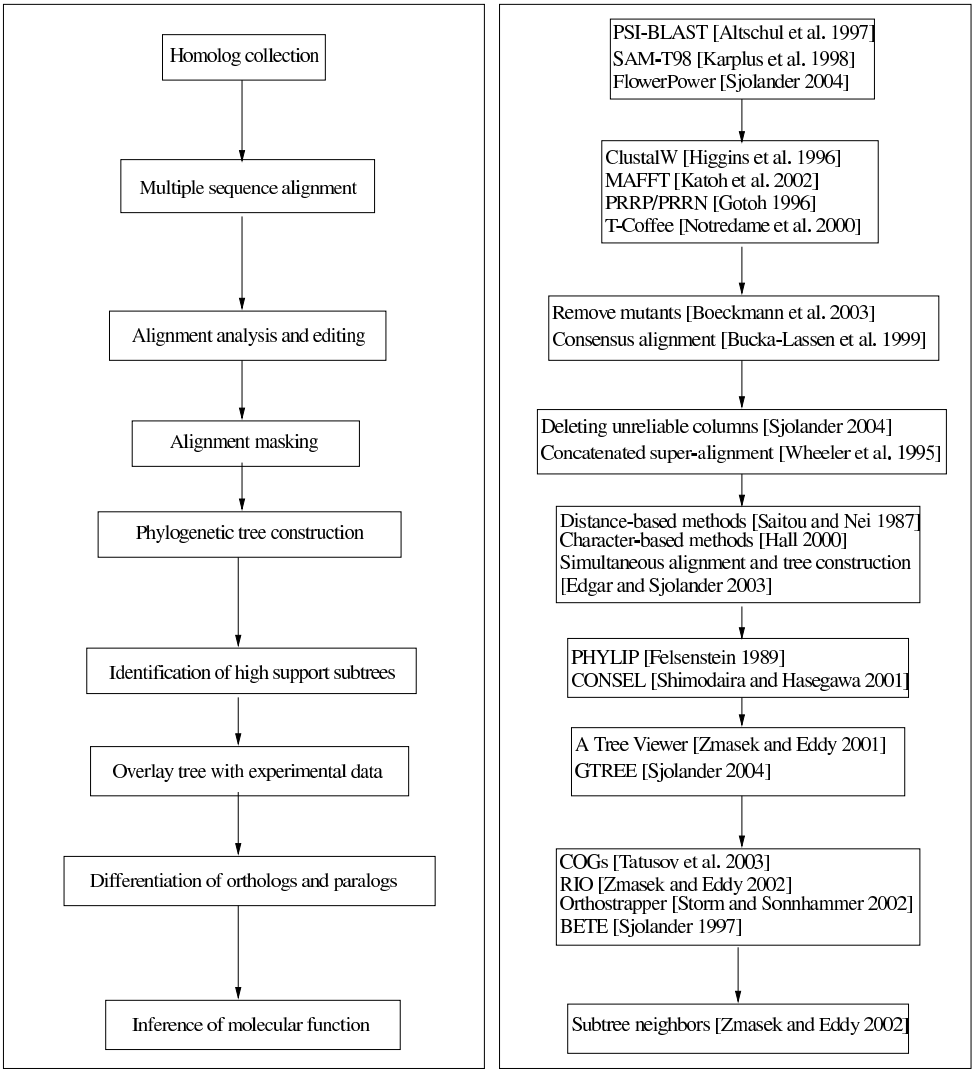


Fig. 16. (Left flowchart) The approach adopted by Sjolander [2004] (Right flowchart) The techniques used to accomplish the corresponding task in the left flowchart (Figure adapted from [Sjolander 2004])

with the techniques used to accomplish these steps. This complete system is a testimony of the strengths of phylogenomic data, particulary phylogenetic trees, when applied to the protein function prediction task.

By far, the best results in the function prediction problem via phylogenetic analysis have been reported by Engelhardt et al. [2005], and subsequently improved in [Engelhardt et al. 2006]. In their papers, they describe SIFTER (Statistical Inference of Function Through Evolutionary Relationships), which is based on the general formalism of probabilistic graph models. SIFTER defines a transition probability function for the transfer of molecular function from a parent to a child node in a single phylogenetic tree and uses standard probabilistic propagation algorithms for computing the posterior probability of a node having a certain molecular function. Various types of validation of SIFTER's re-

sults, such as ROC analysis, suggested that this approach was more accurate than the most significant sequence similarity-based algorithms, and was superior to its closest counterpart. In particular, using 100 Pfam [Sonnhammer et al. 1997] families supplemented with GO annotations as the test and training data, SIFTER achieved a very high precision with complete coverage. Thus, this work comprehensively showed the benefits of incorporating phylogenetic trees into the function prediction process.

6.2.3 *Hybrid Approaches.* It can be seen from the previous two sections that phylogenetic profiles and phylogenetic trees represent diverse forms of evolutionary knowledge and have differing abilities in predicting protein function. In such a situation, a promising idea, which has worked for several other problems, is to combine these two forms of data, thus leading to a hybrid of the approaches discussed above.

The first paper which presented such a strategy was [Vert 2002]. It proposes the use of support vector machines (SVM) for learning protein functions from their phylogenetic profiles. However, instead of the common kernel functions used for SVMs, such as the linear or the radial basis kernel, a tree kernel is proposed to calculate the similarity of the profiles in the higher dimensional space used by the SVM. This high-dimensional feature space is defined on the basis of the patterns of evolution of genes among the (hypothetical) ancestors of the organisms under consideration, in a pre-specified phylogenetic tree, such as the one shown in Figure 15. A linear time algorithm in the number of organisms, based on a post-order traversal of the tree, is also derived and its correctness proved. Upon an ROC analysis of the classification performance on the genes of S. cerevisiae, it is found that the tree kernel works significantly better than the naive linear kernel, particularly for the smaller and more heterogenous classes. Thus, this approach combines the previously proposed approaches under the aegis of a very powerful mathematical framework.

The above approach is adapted in [Narra and Liao 2005] to use extended real-valued profiles. Here, all the internal nodes of the phylogenetic trees are also assigned scores equal to the average of the scores at their children. An extended profile is now constructed for each protein by a post-order traversal of the tree. An SVM with a polynomial kernel is trained with these profiles and is used for function prediction. In evaluations using three-fold cross validation on the same data, performance better than that of [Vert 2002] is reported.

6.3 Discussion

The previous sections discuss several approaches which incorporated evolutionary knowledge into the function prediction process, leading to improvements in the results. In particular, the more extensive the knowledge, the more accurate are the predictions, as shown by projects such as [Sjolander 2004] and [Engelhardt et al. 2005]. Specifically, Brown and Sjlander [2006] cite several technical challenges in phylogenomic inference of protein function, such as the inaccuracy in phylogenetic tree construction, the reliability of existing database annotations and functional inference from orthology without taking evolutionary distance into consideration. From a more evolutionary perspective, Kensche et al. [2007] argue that for effectively addressing the problem of predicting protein function using evolutionary methods, it is critical to examine the effect of evolution on the multi-functionality of proteins, higher order functional relationships between proteins, the functional context of a protein, and the modularity of functional modules. Addressing these challenges and subsequent improvements in the current state-of-the-art in this relatively new field will

lead us closer to validation of Dobzhansky's statement that *"Nothing in biology makes sense except in the light of evolution"* [Dobzhansky 1973].

7. GENE EXPRESSION DATA

7.1 Introduction

Protein synthesis from genes occurs in prokaryotic organisms in two phases [Weaver 2002], as shown in Figure 3. In the transcription phase, an mRNA is created from the original gene by converting the latter to the corresponding RNA code. The protein is then synthesized from mRNA by translating the RNA code to the corresponding amino acid sequence according to the codon translation rules.

Gene expression experiments are a method to quantitatively measure the transcription phase of protein synthesis [Nguyen et al. 2002]. The most common category of these experiments use square-shaped glass chips measuring as little as 1 inch on either side, also known as cDNA microarrays, and hence the alternate name microarray experiments. The experiment is carried out in the following stages. In the first stage, the chip is laid out with a matrix of dots of cDNAs, usually several thousands in number, one corresponding to each of the gene being measured. In parallel, mRNA is extracted from both the normal as well as the cells of the organism that have been exposed to the condition being studied. Next, these mRNA are reverse transcripted to cDNA and colored with green and red colors respectively. These colored cDNAs are then spread on the microarray chip, leading to a hybridization of the cDNA already on the chip with those produced by the genes in the two types of cells. This generates a spot of a certain color on the chip for each gene which denotes its expression level. In the final stage of the experiment, the intensity of this region is measured by a laser scanners connected to a computer, which generates a real valued measurement of the expression of each gene as the ratio of the log intensities of red and blue colors in the region. The result of the experiment thus is a measurement of the transcription activity of the genes under the specified condition. A detailed illustration of this procedure is shown in Figure 17. Recently, single-channel experimental procedures have also become popular in the microarray community. However, the nature of the data and the approaches used to analyze them are usually very similar in nature. So we do not describe this experimental process in detail for brevity.

The primary advantage of gene expression experiments are that they offer an effective method for observing the simultaneous activity of thousands of genes under a given experimental condition. Using these activity measurements, several important inferences can be drawn about the underlying biological phenomena, such as the active pathways under the given condition. This ability to observe a global pattern of activity of genes, particularly when observed over several multiple related experimental conditions, has motivated the use of microarrays for a variety of biological studies [Slonim 2002]. Also, since data generated from one experiment can be useful for several other studies, several repositories have been set up in order to make such data publicly accessible. Several important organism- and phenomenon-specific databases are listed in Table VII. A (slightly dated) comparison of several such microarray databases appears in [Gardiner-Garden and Littlejohn 2001]. This paper presents important details of these databases, such as their commercial aspects, analytical capabilities and system requirements, and is a useful resource for researchers working with microarray data.

In conclusion, a few words about the nature of gene expression data are in place. Usu-

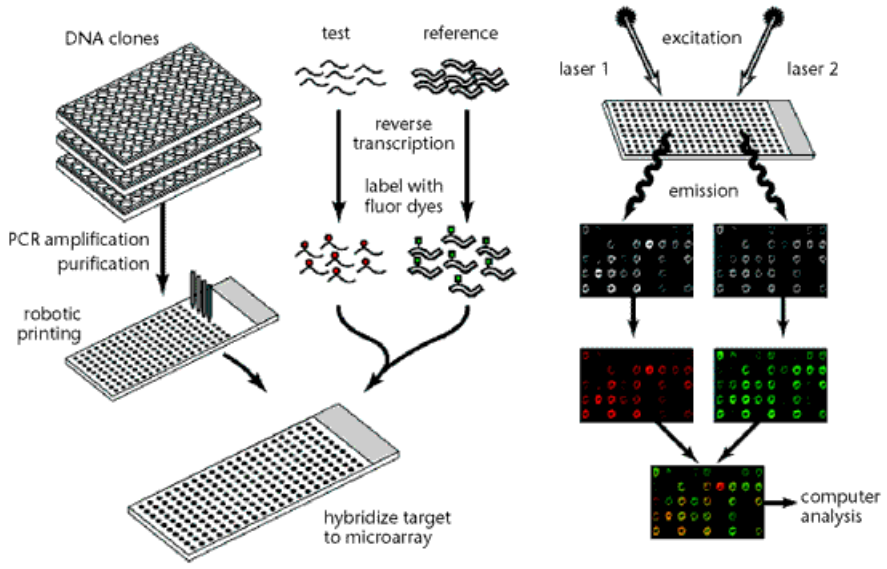


Fig. 17. An illustration of the microarray experimental procedure (Figure taken from [Duggan et al. 1999])

Name	Short Description	Reference
GEO	Largest general repository	[Barrett et al. 2005]
ArrayExpress	General repository	[Parkinson et al. 2005]
SMD	General repository	[Sherlock et al. 2001]
GeneNote	Automated human genes database	[Safran et al. 2003]
BodyMap	Several human and mouse tissues	[Sese et al. 2001]
GXD	Laboratory mouse	[Hill et al. 2004]
yMGV	S. cerevisiae and S. pombe	[Lelandais et al. 2004]
BarleyDB	Various plant species	[Shen et al. 2005]
Drosophila	Drosophila melanogaster (FruitFly)	[Neal et al. 2003]
CGED	Cancerous tissues in humans	[Kato et al. 2005]
BGED	Processes in mouse brain	[Matoba et al. 2000]
MEPD	Medaka fish	[Henrich et al. 2003]

Table VII. Some important organism- and phenomenon-specific gene expression databases. A more extensive list can be obtained from the MetaDB website (`http://www.neurotransmitter.net/metadb/index.php`).

ally, the format of gene expression data is very simple, i.e., a rectangular matrix, in which the rows correspond to genes, the columns to conditions, and the entries denote the expression measurement of a gene under a particular condition. However, since the data is generated experimentally, there may be several phenomena that may affect the quality of data produced by an experiment. Some such problems are varying degrees of hybridization across the chip, background noise in the images produces, and a difference of scale between the different experiments constituting a microarray data set. Several statistical methods have been developed for addressing these problems [Quackenbush 2002], which use the information in the experimental design, as well as the data generated, in order to

reduce the effects of these factors in the processed data set. Another important factor to consider in microarray data analysis is that the data used in research is generally of two kinds: static and temporal [Lee 2004]. The first category consists of data sets containing snapshots of the expression of certain genes in different samples under the same conditions, while the latter, also known as time-series gene expression data [Bar-Joseph 2004], consists of data sets capturing the expression of certain genes of the same organism at different instances of time. It is important to consider these characteristics of the data when it is used for computational analyses, such as running analysis algorithms to infer protein function from gene expression data.

7.2 Existing Approaches

Gene expression experiments are targeted at the simultaneous observation of the activities of thousands of genes under a certain condition. Since transcription is an intermediate step in protein synthesis, the expression measurements give an indication of which genes are active and producing proteins for a the function(s) to be performed under that condition. Thus, due to the ability to simultaneously observe thousands of genes, microarray data holds great promise for determining the function and functional associations of proteins. Also, the matrix format of the data makes it easily processed by computer algorithms. Accordingly, several computational approaches have been proposed for predicting the function of a protein from gene expression data, which will be discussed in this section.

Early approaches identified functional associations between genes by measuring the similarity between their expression profiles using statistical methods. In a study focused on identifying novel genes which may contribute to prostrate cancer in humans [Walker et al. 1999], $40,000$ genes were examined for co-expression with five genes known to be associated with prostrate cancer using the Guilt by Association (GBA) principle. As a result, eight novel genes that are significantly co-expressed with at least one known prostrate cancer causing gene are identified and are verified as being related to processes leading to the disease. However, these studies were usually very narrow in scope, and involved significant human intervention in identifying the seed or the target genes. This allowed the application of more generic techniques from data mining for this task. These can be grouped into the following three categories:

—**Clustering-based approaches**: An underlying hypothesis of gene expression analysis is that functionally similar genes have similar expression profiles, since they are expected to be activated and repressed under the same conditions. Because clustering is a natural approach for grouping similar data points, approaches in this category cluster genes on the basis of their gene expression profiles, and assign functions to the unnannotated proteins using the most dominant function for the respective clusters containing them.

—**Classification-based approaches**: A more direct solution to the problem of predicting protein function from gene expression profiles is the data mining approach of classification. Thus, approaches in this category build various types of models for the expression-function mapping using classifiers, such as neural networks, SVMs and the naive Bayes classifier, and use these models to annotate novel proteins.

—**Temporal analysis-based approaches**: As mentioned earlier, temporal gene expression experiments measure the activity of genes at different instances of time, for instance, during a disease. This behaviour can also be used to predict protein function. Thus, approaches in this category derive features from this temporal data and use classification

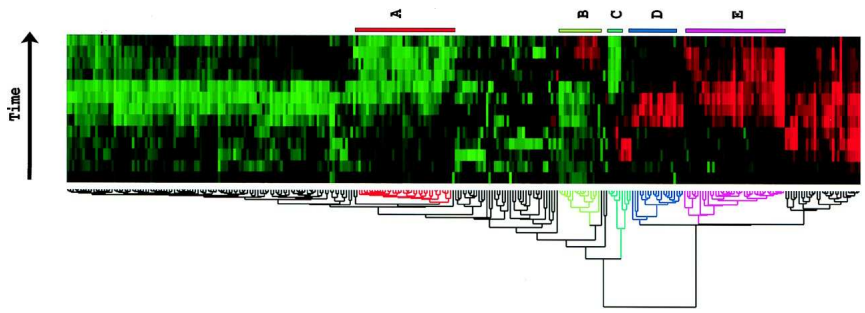


Fig. 18. Visualization of the clustering dendrogram proposed by Eisen et al. [1998]. Members of a cluster are observed to have consistent color codes

techniques to predict the functions of unannotated proteins.

7.2.1 *Clustering-based approaches.* The first category of data mining approaches applied to microarray data was that of unsupervised learning techniques, particularly clustering [Jain and Dubes 1988]. Clustering of gene expression data has been in practice for a long time [Jiang et al. 2004], and algorithms such as CAST [Ben-Dor et al. 1999], which are aimed specifically towards clustering of expression data, are widely used. However, predicting functions of genes from clusters generated by these algorithms using measures such as majority [Guthke et al. 2000], has not yielded good results [Zhou et al. 2005]. Hence, in this section, we survey several approaches that are focused towards creating clusters that can be used for function prediction.

Eisen et al. [1998] reported the first exploration of clustering for gene expression data, and laid the ground for research in this area. They clustered the budding yeast expression data using a hierarchical average-linkage clustering algorithm, with a variant of the correlation coefficient as the similarity measure. However, the main focus was not to determine the best clustering algorithm for gene expression data, but to study the following aspects of the clusters produced:

—**Visualization**: For the purpose of visualizing the cluster memberships of a large number of genes, a color coding scheme was adopted for the data, and the genes were ordered in the hierarchical clustering dendrogram, thus giving the popularly used clustering display of the type shown in Figure 18.

—**Functional analysis**: This was the main target of the study. Once the clusters could be visualized, it was clearly seen that most of the large clusters showed a strong tendency to contain genes involved in common cellular processes. For yeast, some of the most prominent patterns were seen for genes encoding ribosomal proteins, mitochondrial protein synthesis genes and genes involved in ATP synthesis and oxidative phosphorylation. Similar functional coherence was seen for a human expression data set also, though the analysis here was obscured by the limited annotation available for the human genes.

The systematically derived conclusions of this study [Eisen et al. 1998] showed that clusters of coexpressed genes are also functionally coherent. This was a landmark discovery in the field of bioinformatics, and it generated a lot of interest in the clustering of biological data.

In order to exploit the conclusions of Eisen et al. [1998], Ben-Dor et al. [1999] proposed

a heuristic clustering algorithm known as Cluster Affinity Search Technique (CAST), geared towards gene expression data. CAST has two phases, namely the *add* phase, in which elements with high affinity to the current cluster are added to the cluster, and the *remove* phase, in which elements with low affinity are removed from the cluster. The clusters are constructed incrementally one by one, and the process terminates when no more changes occur. This way, CAST is different from K-means since it constructs only one cluster at a time, while K-means updates all K clusters in each iteration. Results on both static and temporal expression data showed CAST's ability to preserve functional categories. In an interesting experiment involving the clustering of gene expression data obtained from 40 tumors and 22 normal tissues, rich clusters, both of normal ($19/22$) and tumor ($36/40$) samples, were obtained. Thus, the ability of the algorithm to extract knowledge about diseases from expression data was shown. This algorithm has been widely used in other studies [Bellaachia et al. 2002; Swift et al. 2004].

Ng et al. [2004] proposed the use of the popular latent semantic indexing technique [Letsche and Berry 1997] to eliminate noisy and redundant dimensions from the data. Only the dimensions that contribute significantly to the data are retained. The final data set is clustered using the concept of neighborhood such that the similarity between any two genes in a cluster is above a certain threshold and is significantly higher than their similarity with a gene from some other cluster. Finally, the majority annotation of the characterised genes in a cluster is predicted to be the function for the uncharacterised genes in the same cluster. A high precision and recall is reported, though this is due to the relaxed definition of recall, which is satisfied even if only one of the several functions of a gene is recovered.

The application of clustering to gene expression data is often confounded by decisions such as which clustering algorithm to use, how many clusters to find, which similarity measure to use etc. In order to increase confidence in the results derived from clustering, Zhou et al. [2005] proposed the use of the novel ontology-based pattern identification (OPI) strategy. The goal of OPI is to enable the clustering process to identify the best decisions to be made in order to identify the best cluster corresponding to a functional category. This is achieved by embedding all the decisions, such as the attribute weights, the choice of mean or median to represent the cluster and the similarity threshold, in a Euclidean space, and defining an objective function, that reflects the characteristics of the most appropriate clustering, on it. Next, a hill climbing process is used to minimize this function for all the GO functional categories and identify the best cluster for this category. Finally, the uncharacterized genes in a cluster are hypothesized to be functionally linked to the annotated ones in the same cluster, in many cases having the same function. This procedure is applied to the gene expression data describing the life cycle of the malarial parasite *Plasmodium falciparum* [Bahl et al. 2003]. As evidence of the validity of their procedure, it is noted that 12 of the 50 genes predicted in an earlier version of the study to have the Antigenic Variation function had now been verified. Also, OPI was able to identify more statistically significant clusters compared to those obtained in the an earlier study of the same data [Roch et al. 2003], where the k-means algorithm was used. Thus, OPI makes the clustering process more flexible.

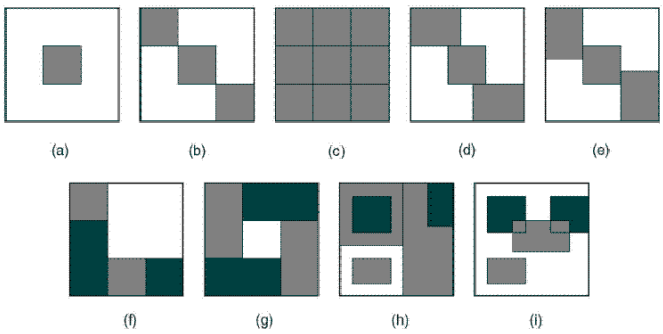
Another way in which confident results can be derived from clustering of gene expression data is by considering overlaps of clusters derived for the same data set by various algorithms. In [Wu et al. 2002], a database of clusters is constructed by applying multiple clustering algorithms, such as K-means, self-organizing maps (SOMs) and hierarchical

clustering [Tan et al. 2005]. These clusters are annotated with the class having the least p-value, which is calculated using the fractions of various functional classes in the cluster. For the purpose of function prediction, each uncharacterised gene is assigned the annotation of each cluster it belongs to, in addition to a confidence value for the prediction. The method is validated quantitatively by treating each characterised gene as uncharacterised, analyzing the overall results so obtained, and qualitatively by showing that the functions of individual genes are predicted accurately.

This strategy is carried forward by Swift et al. [2004], who propose the use of robust clustering (RC) and consensus clustering (CC) for function prediction. Robust clustering provides an incremental clustering algorithm for clustering those genes that are consistently clustered together with at least one other gene. As is apparent, RC has low coverage, though the accuracy obtained is high, thus illustrating the accuracy-coverage tradeoff. Consensus clustering relaxes the condition of complete agreement by introducing a $minimum\ agreement$ parameter. Using this parameter, an objective function for each cluster is defined, which rewards clusters containing instances having high agreement and penalizes those with low agreement. A simulated annealing procedure is then applied to minimize this function globally and thus obtain a clustering of the genes with high agreement. When evaluated in terms of the $weighted - \kappa$ measure [Altman 1997], it is found that CC improves on the performance of each of the individual clustering algorithms. It was also concluded that the clusters identified for ten functional classes were more likely to be annotated with the same classes, as compared to the individual clustering algorithms.

From the previous descriptions, it is clear that an inherent assumption of all the clustering approaches is that functionally associated genes will have similar expression profiles. This may be a necessary condition for association, but is definitely not sufficient. This problem is addressed using a graph-theoretic approach by Zhou et al. [2002]. Here, it is proposed that shortest paths between genes in a network constructed on the basis of strongly correlated expression profiles suggest a way of identifying transitively related genes. These shortest paths are constructed between genes of the same GO category (only "informative" GO categories are used here), and a test is applied to check if the genes lying on these paths are annotated with the same or a parent or child function in Saccharomyces Genome Database (SGD) [Dwight et al. 2002]. Results on the Rosetta compendium [Hughes et al. 2000] indicate a high accuracy for mitochondrial and cytoplasmic genes, but only medium accuracy for nuclear genes. Nevertheless any case, a significant result of this study was the assignment of functions to 146 yeast genes which were otherwise weakly correlated to other genes.

Finally, in another attempt to relax the coherent profiles requirement of clustering, a new form of clustering, known as *biclustering* or *coclustering*, is increasingly being applied to biological data [Madeira and Oliveira 2004]. In this form of clustering, both genes and schemes are simultaneously clustered to produce blocks of entries in the original rectangular data matrix. Several variants of biclusters in such a matrix are shown in Figure 19. It is easy to see that gene expression data, because of its rectangular format, is highly suited for the application of this technique. In addition, the underlying biological motivation for this application is that some groups of genes may be expressed only under a certain set of conditions, and the rest of the conditions act as noisy attributes when these genes are clustered using a traditional clustering algorithm. Based on this motivation, a good deal of work has been done in biclustering gene expression data [Cheng and Church 2000; Yang



Bicluster structure. (a) Single bicluster, (b) exclusive row and column biclusters, (c) checkerboard structure, (d) exclusive rows biclusters, (e) exclusive columns biclusters, (f) nonoverlapping biclusters with tree structure, (g) nonoverlapping nonexclusive biclusters, (h) overlapping biclusters with hierarchical structure, and (i) arbitrarily positioned overlapping biclusters.

Fig. 19. Different variations of biclustering (Figure taken from [Madeira and Oliveira 2004])

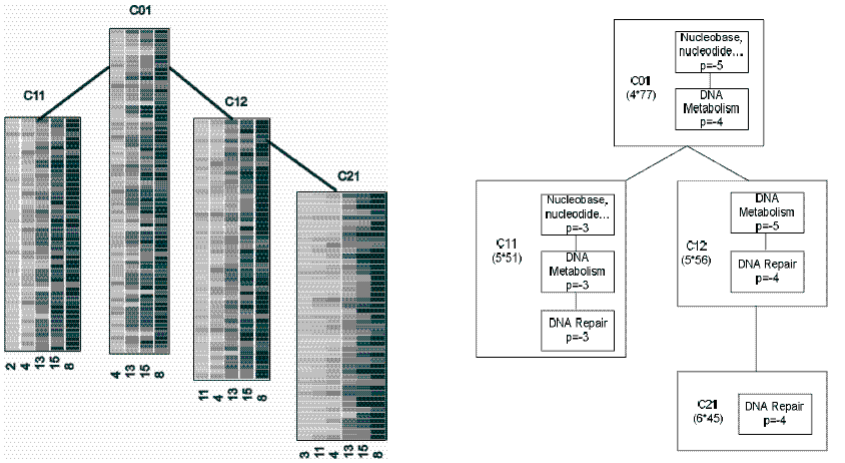
et al. 2003], though these early works did not particularly aim at discovering functionally coherent (bi)clusters. Bryan et al. [2005] developed a simulated annealing based approach for this problem and validated their results on the yeast cell cycle data set [Cho et al. 1998] using annotations from the KEGG database [Kanehisa et al. 2004]. Indeed, it was seen that the two largest clusters were enriched with members of two families, namely ribosomal proteins and nucleotide metabolism, thus demonstrating the potential of biclustering for function prediction.

The tightest coupling of biclustering and protein annotation can be seen in [Liu et al. 2004], where the structure of GO is incorporated into the hierarchical biclustering process, thus making the clusters obtained functionally enriched. In this approach, the genes are originally clustered using a hierarchical clustering algorithm, and each node in the hierarchy is annotated with the GO functional class it is most enriched with. The result of the complete process is a Smart HTP-tree (SHTP-tree), since it intelligently includes GO information in the clusters. An added attraction of this approach is that it allows overlapping clusters, which makes sense biologically. In the experiments, it is verified, both quantitatively and qualitatively, that the nodes in the SHTP-tree are indeed enriched with at least one GO class. An example of a successful mapping from TP-clusters to their Ontology SubTrees (OSTs) is shown in Figure 20.

After this detailed discussion of the various forms of clustering that have found application in microarray data analysis, one may overestimate the potential of clustering for the function prediction task. Clustering suffers from some obvious drawbacks, which have to be addressed in order to realize the full potential of this very powerful technique. Note that these issues are algorithm-independent and affect the analysis over and above the performance of the specific clustering algorithm used.

(1) Traditional algorithms find disjoint clusters of genes, which is not always the right thing to do since genes are known to be involved in multiple functional classes simultaneously. Overlapping clustering [Banerjee et al. 2005; Liu et al. 2004] offers a potential a solution to this problem.

(2) A group of genes may only be expressed in a subset of the conditions, which causes



(A) Expression profiles of the TP-Cluster subtrees (B) Corresponding OSTs of clusters in (A)

Fig. 20. A successful result from the SHTP-clustering algorithm [Liu et al. 2004]

the appearance of dense regions in the gene expression matrix. However, most algorithms do not consider this phenomenon, and feature selection algorithms are unable to perform a per-example analysis. Only biclustering [Madeira and Oliveira 2004] handles this issue to some extent.

(3) It is hard to judge the most appropriate proximity (similarity or distance) measure that should be used when clustering biological data. Euclidean distance suffers from the curse of dimensionality when the number of conditions in a data set exceeds a small number, as is usually the case. Similarly, correlation accounts for co-repression the same way as co-expression, which may not be very suitable for some applications, particularly those which transform the gene expression data into a binary format. Thus, even though correlation is the most widely used measure, the cosine measure may be more appropriate for gene expression data, because of its focus on the shape of the profile and not its magnitude [Kuramochi and Karypis 2001].

(4) The general method of transferring annotation from the majority genes in a cluster to the unannotated ones may not be useful if very few genes in a cluster have been annotated earlier.

(5) A related weakness of clustering algorithms is that they do not make use of the class label information available in many data sets. As a result, annotation can only be indirectly achieved via clustering, and in cases where this route is adopted, the results are prone to two levels of error.

Recent approaches have started addressing some of these issues. For instance, Pan [2006] proposed an approach for incorporating functional annotations into the clustering process for gene expression data. Here, the commonly used mixture model-based clustering algorithm [Duda et al. 2000] is modified so as to increase the probability of clustering together genes belonging to the same functional class. This is achieved by assigning the same prior probability to all the genes in a given functional class, during the expectation-

maximization (EM) process for computing posterior probabilities of genes belonging to different clusters. Through experiments on simulated data, as well as Hughes et al. [2000]'s large yeast microarray data set, it is shown that this incorporation of functional labels indeed improves the performance of standard clustering algorithms. This illustrates the merit of using a supervised version of clustering for inferring protein function from microarray data.

Nevertheless, the above issues are very strong points to be considered when applying clustering to gene expression data in particular. Added to them are the problems due to noise and the choice of the right algorithm and the right parameters, which are valid for many other applications. Handling these issues effectively will make the task of function prediction easier and more achievable.

7.2.2 *Classification-based approaches.* As explained in Section 7.1, the basic format of a gene expression data set is a gene-condition matrix, where each gene is denoted by a vector (row) and each condition by an attribute (column). In some applications, the functions of some of the genes in a study may be known, and act as the class labels for the corresponding vectors. The other genes remain unlabeled, and it is desired to assign a label to them. This problem can be solved using classification algorithms in data mining, such as neural networks and support vector machines [Tan et al. 2005]. This section discusses several approaches that have been proposed in this direction, and also examines the factors influencing their success or failure at the task of function prediction

One of the most cited works in the field of protein function prediction, or bioinformatics, for that matter, is [Brown et al. 2000], which describes the application of support vector machines for learning functions from yeast gene expression data. Various kernel functions are used for the SVM and these versions are compared with three other popular classifiers, namely Parzen windows, Fisher's linear discriminant [Duda et al. 2000] and two decision tree classifiers, C4.5 [Quinlan 1993] and MOC1 [Wu et al. 1999]. With respect to the consideration set of five functional classes (and another comprising the rest), it was concluded that SVMs with a radial basis kernel were the most suitable technique for the purpose. Also, it was observed that many of the false positives generated by SVM were known to be related to the functional class assigned to them by the SVM. It was argued that this may be due to various factors such as noisy data, overlaps between functional classes, and the fact that some genes are not regulated at the transcription level and hence were hard to characterize using gene expression data.

A more extensive study of the issues in the classification of gene expression data is presented in [Mateos et al. 2002]. Two major points of departure from [Brown et al. 2000] are as follows. First, Mateos et al. [2002] use multilayer perceptrons (MLP), or neural networks, for the learning and prediction tasks. Second, a wider set consisting of 96 function classes are considered for the yeast genome. For the sake of comparison with SVM, an MLP network is also trained and tested for the same five classes used in [Brown et al. 2000]. It was found that while SVM and MLP are comparable in terms of false negatives (FN), the latter performed systematically worse in terms of false positives (FP). In order to analyse the factors contributing to the poor performance of machine learning techniques, a one-against-all learning and testing procedure using MLP was followed for all the 96 classes considered. The results obtained indicated that only 8 classes were learnt with a true positive (TP) rate$\geq 40\%$, which proves that not only the learning technique, but also the nature of the data set dictates how well the relationship between the gene expression

variables and the functional class can be learnt.

The main merit of [Mateos et al. 2002] was the systematic and quantitative study of the causes for poor learning performance when gene expression data is used. They identified and quantitatively verified three main factors, besides noisy data, which determined learning performance. These factors are as follows:

—**Class size:** In general, learning performance, measured here in terms of the TP rate, tends to improve with class size. This is not surprising since the larger classes tend to act as attracters for the instances belonging to the smaller classes.

—**Class heterogeneity:** A class is homogeneous when all its members follow the same behaviour and hence act as good examples of each other during learning. Clearly, the more homogeneous a class it, the better learning performance it will give. Using a measure defined in terms of the divergence between consecutive genes in a class, Mateos et al. [2002] observed that there is a linear relationship between the TP rate and the heterogeneity measure. Also, it is suggested that larger classes tend to be more homogeneous than smaller ones, though this needs more rigorous verification.

—**Borges effect:** It is common knowledge that a protein participates in multiple functions and hence there are numerous interconnections between functional classes in terms of simultaneous memberships of genes. This heavy interaction nature of biological functions leads to poor learning rate of individual classes. For the problem of protein function prediction, this seems to be the most important factor affecting learning performance. This effect is named the *Borges effect* [Mateos et al. 2002], crediting an observation made by the philosopher Jorge Luis Borges in his widely cited work [Borges 1964]. Again, using quantitative measures of this effect, it is verified that the learning accuracy decreases with an increase of overlaps between functional classes.

The identification of these factors is significant since addressing these issues independently may be expected to lead to better classification performance using gene expression data. Mateos et al. [2002] utilize these insights to propose an iterative learning procedure for individual classes and this algorithm gives significantly better performance than a one-pass learning procedure.

In a complimentary work, Kuramochi and Karypis [2001] analyzed the feasibility of using supervised learning techniques in general, and the SVM and the k-nearest neighbor (k-NN) classifiers in particular. This study used the cosine similarity measure, which appears to be the more appropriate measure than Euclidean distance for gene expression data, as noted earlier. Similar to [Brown et al. 2000], the results for one-against-all classification were better for SVM, and a further analysis of the results led to conclusions consistent with [Mateos et al. 2002]. In addition, the problem of the prediction of m most appropriate classes for a test gene was addressed. For this, k-NN appeared to be the most appropriate solution, since the annotations of the m closest genes can be directly transferred to the gene in question. At a higher level, this study concludes that for confident prediction, it is necessary for the activity of the characterized genes to be observed under a wide variety of condition, resulting in a diverse and more informative data set.

As mentioned before, one way of increasing the confidence in the prediction of gene functions is the combination of multiple expression data sets [Hughes et al. 2000]. Moving forward from [Hughes et al. 2000], which involved significant manual interpretation, Ng et al. [2003] analyse the feasibility of combining multiple data sets for learning with

SVMs, and present a strategy to select the most informative data sets for learning individual classes. In terms of the learning cost savings measure defined in [Brown et al. 2000], it was concluded that blindly combining all the available data sets is not the most appropriate method of data preparation. Hence, a simple hill-climbing algorithm was devised, which incrementally adds the data set that provides the maximum learning cost savings, continuing this way till the maxima is achieved. Evaluation of this algorithm and its comparison with other feature selection algorithms showed the superiority of the former with respect to learning performance.

Finally, it should be observed that all the above techniques were tested for small model organisms such as *S. cerevisiae* and *C. elegans*, which have few thousand genes in their genome. The complexity of the problem increases in mammalian species, which have tens of thousands of genes. One such study of the functional classification of the mouse gene expression data set, which covers over 40,000 genes, appears in [Zhang et al. 2004]. The first part of this paper, using a methodology similar to [Eisen et al. 1998], concludes that genes in the same functional category were indeed co-expressed, and in the second part, this observation was used to classify about 10000 unannotated genes. This classification was conducted using a support vector machine, and an individual model was learnt for each of the 992 GO Biological Process categories considered. The classification accuracy was not very high, though about 1000 genes were annotated with more than 50% probability (calculated from the discriminant value of the SVM). This may indicate that mammalian gene expression data needs more sophisticated analysis in order to improve the prediction coverage and precision.

In conclusion, the nature of gene expression data has created huge interest in the machine learning community, especially researchers in classification, and there is much work being performed in the field of classification-based gene expression analysis. As of now, it would be safe to say that the state of the art in this field is support vector machines (SVMs), since they have shown the best performance with microarray data for several classification problems [Mukherjee 2003]. However, this field is still in its early stages of maturity and more advanced classification techniques such as boosting and active learning [Duda et al. 2000] are expected to be applied to this data in the future, leading to better results.

7.2.3 *Temporal analysis-based approaches.* The wide variety of approaches presented till now adopted a static view of the gene expression data. However, as was noted in the Section 7.1, this data can also be viewed as a temporal data set if each experiment tracks the activity of different genes at different points of time. Some researchers have used this temporal nature of some gene expression data sets for function prediction, and this is the topic of discussion in this section.

The first cut at analyzing time series expression data were made via clustering [Bar-Joseph 2004; Moller-Levet et al. 2003], since it is an unsupervised technique and does not require the assignment of functional labels to the genes. A detailed discussion of the issues involved in this approach can be found in [Moller-Levet et al. 2003]. In recent studies, further issues regarding the temporal clustering of gene expression have been addressed, such as the short lengths of the profiles [Ernst et al. 2005], sampling [Jiang et al. 2004], co-clustering [Heard et al. 2005] and appropriate similarity measures [Butte et al. 2001].

A more direct use of the temporal nature of gene expression profiles is made in [Hvidsten et al. 2001; Laegreid et al. 2003], in which the data is transformed into a suitable attribute-value vector format, so that a rough set based classifier could be used to extract

Gene Ontology biological process (BP) labels for unknown genes. These attributes are constructed by calculating the increase or decrease of expression values between two instances separated by an interval of three time points, and then categorizing them into three classes *high*, *medium* and *low*. This is necessary since rule-based classifiers, such as the rough set-based classifier here, can only handle nominal attributes robustly. Next, the classifier is learnt from this transformed set, in the same manner as [Hvidsten et al. 2001], and is tested on the human serum response expression data set of Iyer et al. [1999]. The coverage of the derived rule set is excellent, since it is able to predict labels for 211 of the 213 unannotated human genes in this data set, of which a small number of predictions could be verified from the literature. In addition, cross-validation tests on the training set gave an AUC score (Area Under the ROC Curve) of 0.88, which shows the robustness of the classifier function prediction from gene expression data.

A similar approach that employs inductive logic programming for learning the classification rule set from gene expression patterns which are defined in terms of sets of differentially expressed genes between individual classes is outlined in [Badea 2003]. The use of description logics (DL) [Baader et al. 2003] is also proposed for the purpose of making more fine-grained predictions than the systems are currently capable of. Finally, another rule-based classification system is presented in [Midelfart et al. 2001], which improves upon the performance of [Hvidsten et al. 2001] by including all the members of the subclass of a Gene Ontology class into its member set while learning classification rules for it. The training is more robust, since there are more representative examples available for each class.

Deng and Ali [2004] used hidden Markov models (HMMs) to model the temporal interdependence between the conditions under which the expression experiments are performed, and the dependence of the functional class on them. Analysis of the yeast gene expression data using this strategy suggested that a dual HMM modeling both expression values and experiment order was the best for this application. A more statistical approach was adopted by Gui and Li [2003], which is motivated by the observation that the temporal expression profiles of genes belonging to the same class are highly similar. To capture this observation, each individual class is modeled as a mixture of sub-classes and the parameters of the model are learnt using the EM algorithm [Duda et al. 2000]. This approach, named mixture functional discriminant analysis (MFDA) was compared against other known discriminant analysis methods on the yeast cell cycle expression data set, and MFDA was found to be marginally better than the others. These preliminary approaches suggest the need for further examination of the relationship between the functional class of a gene and its temporal expression profile.

Another interesting way of looking at the temporal behaviour of gene expression is to view it in the light of evolution. van Noort et al. [2003] exploit this view and hypothesizes that the conservation of co-expression between pairs of genes that have a common evolutionary development path can enable more confident prediction of their functional association and the pathways they are involved in. Using the test cases of *S. cerevisiae* and *C. elegans* and using correlation as measure of co-expression, two types of conservation of co-expression were defined. Paralogous conservation refers to two pairs of genes $A - B$ and $A' - B'$ in the same organism, where A and A', and B and B' are homologues of each other respectively, and both pairs have a high correlation between their gene expression profiles. The definition of orthologous conservation is similar, with the only difference

being that the two pairs belong to different organisms. When the correlation threshold for co-expression was 0.6, an accuracy of 93% and 82% are obtained for *S.* cerevisiae using orthologous and paralogous conservation respectively, thus showing the promise of the approach for the prediction of functional association between genes.

More generally, temporal gene expression analysis faces many of the same problems as the analysis of other time series [Bar-Joseph 2004]. For instance, responses for different genes may be offset in time and need to be aligned. Also, responses may be stretched or shrunk with respect to one another or may only be related to one another over various periods. Furthermore, the times series may be missing values at some times. These and other issues have been addressed by various researchers in time series analysis, although it is probably fair to say that none of these problems can be considered as being completely solved.

7.3 Discussion

The previous subsections detailed the numerous approaches that have been proposed for the inference of protein function from the microarray analysis of the coding genes for these proteins. The matrix format of the data yields itself naturally to classical data mining algorithms such as clustering and classification, with much more activity being seen the latter category in recent years. It should be noted that microarray technology has come of age very recently [van de Goor 2005], and the immense popularity it has already gathered, even in data mining publications, is an indication of the potential it possesses for inferring different forms of biological knowledge, such as function, metabolic pathways and evolutionary paths, from the information available about the simultaneous activity of thousands of genes.

However, to extend the usefulness of function prediction from gene expression data, a number of significant issues need to be addressed. One issue is the complexity of the class structure. Genes can have more than one function, and these functional classes are often organized in a hierarchy. In addition, the size of functional classes is often very different, with some being common, while others are rare. Other important issues are data preprocessing and data quality. Extracting information from gene expression data requires the proper choice of normalization and other preprocessing steps. Furthermore, the amount of information available from gene expression data is often limited by the large amount of noise and the fact that the conditions present in a data set may not be the ones needed to identify the functional similarity for some groups of genes.

8. PROTEIN INTERACTION NETWORKS

8.1 Introduction

A protein almost never performs its function in isolation. Rather, it usually interacts with other proteins in order to accomplish a certain function. However, in keeping with the complexity of the biological machinery, these interactions are of various kinds. At the highest level, they can be categorized into genetic and physical interactions. Genetic interactions occur when the mutations in one gene cause a modifications in the behavior of another gene, which implies that these interactions are only conceptual and do not occur physically in a genome. These interactions are mostly detected using computational techniques, that are discussed in detail in Section 5. Of particular interest in this section are the physical interactions between proteins, since they are more directly related to the pro-

cess through which a protein accomplishes its functions. These interactions are of various kinds, such as the simultaneous membership of two proteins in the following biological systems [Xenarios and Eisenberg 2001]:

—A metabolic and/or signaling pathway.

—A morphogenic pathway in order to perform a developmental function.

—A protein complex and other such molecular machines.

Since a protein generally interacts with more than one other protein, these interactions can be structured to form a network, and hence the name *protein interaction networks*. An example of such a protein network, the PX domain protein interaction network in yeast [Voller and Uetz 2004], is shown in Figure 22. A very common way of visualizing these networks is as undirected graphs, with the proteins acting as the nodes and the pairwise interactions acting as the edges of the graph. Such a representation can enable researchers to infer characteristics of proteins from those of proteins not even directly interacting with it.

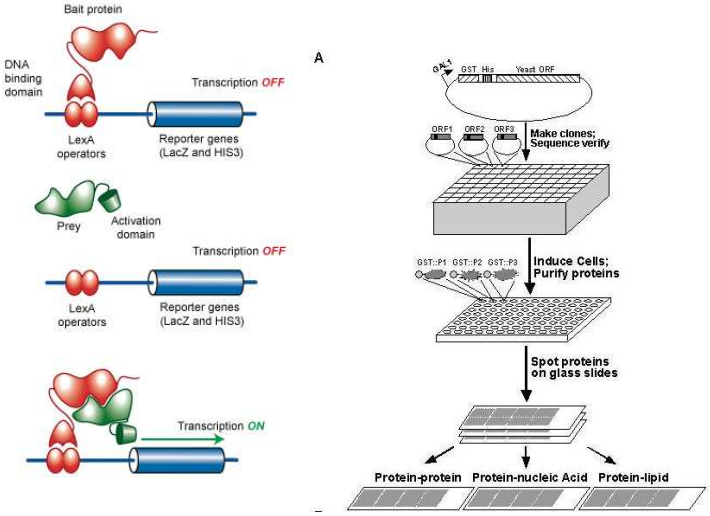
Due to the importance of the knowledge of these interactions, several high-throughput methods have been proposed for discovering them [Legrain et al. 2001]. Again, depending on the final output, these methods can be categorized into two types [Chen and Xu 2003; Droit et al. 2005], namely the discovery of pairwise interactions and extraction of protein complexes. While two-hybrid systems, protein chips and phage display are the most commonly known methods in the former category, the Tandem Affinity Purification (TAP) approach is commonly used for extracting complexes. Figure 21 illustrates some of these methods diagrammatically.

The huge number of interactions discovered by the various experimental techniques discussed above have been organized into numerous databases that have been placed in the public domain. Table VIII presents a summary of some of these databases commonly used by the function prediction approaches discussed below. Xenarios and Eisenberg [2001] present a more detailed description of some of the earlier databases, such as GRID, BIND and DIP.

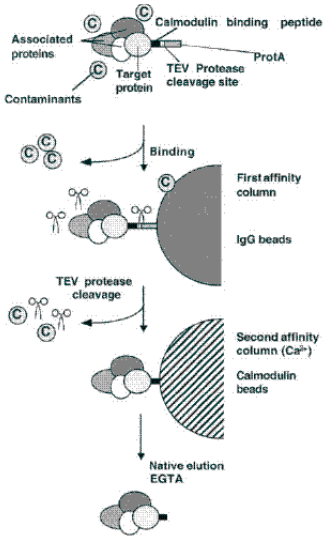
Though most of the data sets in Table VIII are commonly used, there are several issues associated with them, as discussed by Salwinski and Eisenberg [2003]. The most important of these is the large amount of noise present in high-throughput interaction data. Deng et al. [2003] showed quantitatively that the efficacy at the task of function prediction varies between different data sets, primarily due to the presence of different levels of noise. This consideration should be kept in mind during any analysis of interaction data, and techniques such as [Deane et al. 2002] may be used for retrieving the true interactions from this data. Deane et al. [2002]'s technique relies on the use of external data about the proteins, such as their expression profiles and amino acid sequences to determine the reliability of the input set of interactions, and is available for use at the DIP database's website[12].

Recently, some data mining techniques have been proposed for estimating the reliability of a given interaction using the inter-connectivity structure of the complete network [Chen et al. 2007; Pandey et al. 2007]. These techniques utilize the concept of shared neighbors, which is the set of neighbors shared by two proteins, in order to estimate the reliability of an interaction between them. [Chen et al. 2007] discuss several measures based on this

[12] http://dip.doe-mbi.ucla.edu/dip/Services.cgi?SM=1



(a) The two-hybrid method (b) The protein chip method



(c) The Tandem Affinity Purification (TAP) method.

Fig. 21. Popular experimental methods for the discovery of protein-protein interactions. The first two figures have been taken from `http://www.dualsystems.com/technologies/yeast.asp` and `http://bioinfo.mbb.yale.edu/proteinchip/db/FIG1.html`, respectively , while the third is from [Puig et al. 2001].

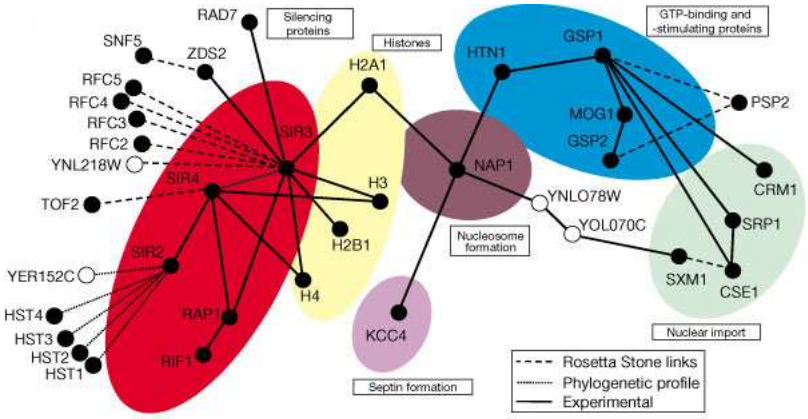


Fig. 22. Network of protein interactions (solid links) and predicted functional links (dashed links) involving silencing information regulator (SIR) proteins (Figure adapted from [Eisenberg et al. 2000]). Note that the network contains several functionally coherent clusters

Name	Ref.	Organism	Type	#Interactions	#Proteins
Ito et al	[Ito et al. 2001]	*S. cerevisiae*	Physical	4549	3278
CuraGen	[Uetz et al. 2000]	*S. cerevisiae*	Physical	957	1004
CYGD	[Guldener et al. 2005]	*S. cerevisiae*	Physical	9103	7000
		S. cerevisiae	Genetic	6385	7000
		S. cerevisiae	Complexes	2876+	7000
MIPS	[Pagel et al. 2005]	10 mammals	Physical	1800+	900+
YPD	[Costanzo et al. 2000]	*S. cerevisiae*	Physical	N.A.	N.A.
DIP	[Xenarios et al. 2002]	*S. cerevisiae*	Physical	18224	4919
		D. melanogaster	Physical	20988	7052
		E. coli	Physical	7100	1631
PIM	[Rain et al. 2001]	*H. Pylori*	Physical	1200+	261
LIGAND	[Goto et al. 2002; Vert 2002]	*S. cerevisiae*	Pathways	16650	774
TAP	[Gavin et al. 2002]	*S. cerevisiae*	Complexes	232	1739
Lehner et al	[Lehner et al. 2004]	*H. Sapiens*	Physical	91	15
GRID	[Breitkreutz et al. 2003]	*S. cerevisiae*	Physical	19791	6713
		D. melanogaster	Physical	28406	26148
		C. elegans	Physical	4453	22268
BIND	[Alfarano et al. 2005]	Several	Interactions	201882	51087
		Several	Complexes	3703	51087
HMS-PCI	[Ho et al. 2002]	*S. cerevisiae*	Complexes	3617	725
Giot et al	[Giot et al. 2003]	*D. melanogaster*	Physical	4780	4679

Table VIII. Popular protein interaction datasets used for the function prediction task and associated features

concept for estimating the reliability of a given interaction in a network, and illustrate that the functional content of the network containing the edges rated as highly confident is significantly higher than that of the original network. In fact, in subsequent studies [Chua et al. 2006; 2007], the same measures were used to extract accurate functional information from interaction networks.

Pandey et al. [2007] took a slightly different approach to the problem of identifying noisy interactions in a given network. They used the h-confidence measure [Xiong et al. 2006] from the field of association analysis in data mining [Tan et al. 2005], which when viewed in the context of a graph, can be used to estimate the similarity between two proteins on the basis of the number of their shared neighbors. Thus, h-confidence is used to estimate the likelihood of an interaction between all pairs of proteins in the network. However, it is not necessary that a an interaction is already known between a pair of proteins whose h-confidence score is high, which indicates that an interaction is probable between these proteins. Thus, the computation of this measure between all pairs of proteins provides a way of indicating the reliability of the edges already present in the network, and the likelihood of including an edge that is currently absent from it. Thus, in addition to addresing the problem of noise, Pandey et al. [2007]'s approach also addresses the problem of *incompleteness* in interaction data, which has been recently suggested to be another important problem for interaction data [Hart et al. 2006]. Indeed, in experiments on several well-known interaction datasets, significantly more accurate inferences about protein function could be obtained from the transformed interaction network, where interactions are weighted by their likelihood of occurrence, as compared to the original network.

These algorithms are of great value to function prediction approaches, given the importance of the problem of false positive interactions in interaction data sets, even the most popular ones.

8.2 The Promise of Protein Interaction Networks

It has long been known that a protein does not perform its function in isolation, but as a part of a group of proteins that cooperate to perform that function. Thus, the arrangement of the known protein-protein interactions to construct a protein interaction network provides a global view of the functions of the proteins, and how they cooperate to achieve higher goals in an organism. Fraser and Marcotte [2004] suggested that the strongest point in favour of interaction networks is that the function of a gene (protein) can be defined precisely by the topological features of the network it is a part of. In addition, protein interaction networks provide several other benefits, which may be useful for the study of protein function:

—Experimental data directly determines these networks.

—The noise inherent in experiments can be modeled by assigning weights to the edges corresponding to the reliability of the experimental method used to extract them.

—These networks represent an integrative approach, in the sense that different types of interaction data, whether genetic or physical, can be imported with equal ease to construct an all-encompassing interaction network. This integration iteratively improves the overall quality of the network.

—Networks have "features" such as areas of high connectivity separated by more sparsely connected regions. These features can be utilized for clustering to discover cellular machines such as functional modules [Snel et al. 2002].

—At a more abstract level, these networks, by their very nature, reflect the interconnected nature of biological processes.

—Interaction networks can be studied under a mathematical framework to discover precise functions for each gene/protein.

Due to the above benefits, several approaches have adopted the route of predicting function by observing the patterns of interaction of each protein in a network, as discussed in Section 8.3.

The above description shows qualitatively that protein interaction networks are a rich source of information about the *context* of a protein, i.e., the position of an individual protein in a larger view of the biological processes in a organism. For instance, in an interaction network, the context of a protein is the set of proteins that it interacts with. Huynen et al. [2003] suggested that exploiting this context information is more effective for predicting functional associations and specific functions than pairwise comparison-based approaches such as mRNA co-expression. This hypothesis is validated by the analysis of thirteen cases of functional associations predicted by genomic context methods (Section 5), that were later verified in the laboratory. A prominent example of this analysis was the discovery that the *frataxin* protein is involved in iron-sulfur cluster assembly on a protein, which is a well-known fact in genetics [Duby et al. 2002]. The authors also postulate that the construction of networks from these associations may also help in discovering functional modules, which is only a step away from protein-specific function prediction. Overall, the authors strongly suggest that if such interaction networks can be analyzed in depth, then the functional modules discovered from them can have the same potential as functional domains for protein sequences [Huynen et al. 2003].

Finally, [Fraser and Marcotte 2004] can also be interpreted as an invitation to the more mathematically-oriented research communities to address the problem of protein function prediction from protein interaction networks. However, even before the publication of this informal invitation, the data mining community had identified the great potential of these networks for solving this problem. The KDD Cup 2001 [Cheng et al. 2002], the most well known research contest in the data mining community, specifically included the task of accurately predicting the functions of several yeast genes/proteins from a set of interactions from the MIPS repository [Mewes et al. 2002]. Several innovative data mining solutions were submitted for this task, thus showing that there is considerable interest in the community for this problem of protein function prediction from interaction networks. A discussion of the submission with the best performance can be found in Section 8.3.1.

These developments show that protein interaction networks hold great potential for accurately discovering functions of proteins and this viewpoint was adopted by several researchers who have proposed very innovative solutions to this problem. Section 8.3 discusses these approaches in detail. Other reviews discussing some of these approaches and issues relate to them have also been published Sharan et al. [2007].

8.3 Existing approaches

Approaches that attempt to predict function from a protein interaction network can be broadly categorized into the following four categories:

—**Neighborhood-based approaches**: These approaches utilize the neighborhood of the query protein in the interaction network and the most “dominant” annotations among these neighbors to predict its function.

—**Global optimization-based approaches**: In many cases, the neighborhood of the query protein may not contain enough information, such as annotated proteins, for determining the function of the query protein robustly. Under these conditions, it may be advantageous to consider the structure of the entire network and use the annotations of the proteins indirectly connected to the query protein also. The approaches in this category are based on this idea, and in most cases, are based on the optimization of an objective function based on the annotations of the proteins in the network.

—**Clustering-based approaches**: The approaches in this category were based on the hypothesis that dense regions in the interaction network represented functional modules, which are natural units in which proteins perform their function. Thus, these approaches apply graph clustering algorithms to these networks and then determine the functions of unannotated proteins in the extracted modules using measures such as majority.

—**Association-based approaches**: Recently, several computationally efficient algorithms have been proposed for finding frequently occurring patterns in data, in the field of association analysis in data mining [Tan et al. 2005]. The approaches in this category use these algorithms to detect frequently occurring sets of interactions in interaction networks of protein complexes, and hypothesize that these subgraphs denote function modules. Function prediction from these modules is performed as in the clustering-based approaches.

It should be noted that, despite the above categorization, the underlying theme of all approaches in this field is that a graph-based representation enable the analysis of several topological features of an interaction network, that can be further analyzed for studying various characteristics of proteins. Hence, the accuracy of any approach is determined by the biological relevance and coverage of the features used. This is the aspect most of the following approaches differ in, and also explains the varying levels of precision and coverage reported by them. Another important factor contributing to the accuracy of these approaches is the amount of noise present in the data, which is a well-known side effect of the experimental approaches used for collecting these data. Several function prediction approaches in this field thus make it a point to demonstrate the robustness of their results when some noise, in the form of spurious interactions, is added to the original network. These issues will be clarified in the following sections that discuss the approaches in the above categories in detail.

8.3.1 *Neighborhood-based approaches.* Given a set of interconnections among a set of entities, the most intuitively straightforward approach for inferring the characteristics of these entities is to extrapolate the characteristics of their neighbors. This idea directly addresses the problem of protein function prediction from protein interaction networks and was used by a very early paper which addressed this problem [Schwikowski et al. 2000]. In this study, a network of 2709 interactions among 2039 yeast proteins were assembled from various sources such as MIPS [Mewes et al. 2002] and DIP [Salwinski et al. 2004]. Even though the prediction method was simple: the functions of a protein are assigned as the (at most) three most frequent functions among its neighbors, an accuracy of 72% was achieved for 1393 characterized proteins. Another interesting discovery made in this study was that 35% of the interactions were between proteins with no common functional annotation, some of which were shown to connected related functional classes, such as protein folding and protein translocation. This illustrates the well-known concept of *crosstalk* be-

tween biological processes [Kunkel and Brooks 2002; Poyton and McEwen 1996]. Overall, this exploratory study established the utility of protein interaction networks for making biological inferences, particularly protein function prediction.

A strategy to improve the statistical significance of these predications was proposed by Hishigaki et al. [2001]. First, instead of just the immediate neighbors, a set of n-neighboring proteins consisting of proteins reached via n links is considered for prediction. Second, the frequencies of all the functions in this neighborhood is recorded. Finally, the most "significant" function in this set is assigned to the protein of interest. This significance is tested using a χ^2-test, that compares the frequency of the function in this neighborhood with that expected according to its occurrence probability across the whole interaction network. Thus, the functions assigned by this approach are more significant than those by [Schwikowski et al. 2000], where some of the assignments may be spurious due to noise in the data. This claim is also validated using a set of 2112 physical interactions assembled in a manner similar to Schwikowski et al. [2000], and three categories of functional classification from YPD [Costanzo et al. 2000], namely subcellular localization, cellular role and biochemical function. From this validation procedure, it was found that the highest accuracies were obtained either for $n = 1$ or 2, depending on the functional class under consideration. This suggests that there might be noise and redundancies in this approach, if too many neighbors are considered. This is intuitively true, since the functions of proteins very far in the network are expected to only indirectly influence the function of the query protein, and thus they should not be weighted the same as the functions of immediate neighbors in the frequency calculations. Global optimization-based approaches, discussed in the next section, implement this observation more robustly.

Another approach for increasing the confidence in predictions made using the annotations in the neighborhood of a proteins has been presented by Kirac et al. [2006]. Here, instead of looking at only the immediate neighbors, a model is built for the sequence of annotations on the paths in the network that lead to the target protein. This model is capable of predicting the possible annotations of the target protein using the sequence of annotations of the proteins lying on paths that terminate at the target protein. This model is implemented for a set of GO functional classes using the probabilistic suffix tree data structure [Ron et al. 1996], which enables the efficient computation of the probability of a certain protein having a certain function. The algorithm is evaluated on a variety of protein interaction datasets, and results better than other neighborhood-based methods are obtained. Thus, this approach provides a robust method for implementing the extended neighborhood-based inference of protein function.

Another useful way of defining neighborhood in a network is through the concept of *shared neighborhood*, which denotes the set of neighboring proteins that are common to two proteins. The use of this concept helps identify the confidence in an interaction between two proteins, as has been done for identifying noisy edges in an interaction network [Pandey et al. 2007; Chen et al. 2007]. An approach for protein function prediction using shared neighbors is adopted in [Samanta and Liang 2003]. Here, the most significant protein pairs are identified in the order of increasing p-values of their association. This p-value is calculated using a formula derived for the probability of the two proteins having the specified number of shared neighbors, assuming this association follows a bionomial distribution. Considering these p-values as similarities between instances, the proteins are then clustered using the hierarchical clustering technique. When this algorithm is applied

to the budding yeast interaction data set taken from DIP [Xenarios et al. 2002], 163 clusters are discovered. Of these, 149 clusters are found to be subsets of some functional complex or pathway according to the *Saccharomyces* Genome Database (SGD) [Dwight et al. 2002]. Novel functions were also assigned for 81 previously unannotated yeast proteins.

A very similar strategy has been adopted in PRODISTIN [Brun et al. 2003], which uses the Czekanovski-Dice distance for calculating the distance between two proteins, and the BioNJ algorithm [Gascuel 1997] for clustering them. This strategy was able to cluster proteins more effectively according to their cellular function, which is the most relevant for function prediction. Using yet another strategy, the same group extended PRODISTIN by replacing the BioNJ algorithm with a density based clustering algorithm [Brun et al. 2004]. Using these new clusters, new functions were predicted for 37 proteins, of which 12 were novel predictions that could potentially be tested in the laboratory.

Both the above approaches utilize the concept of shared neighbors for clustering proteins in interaction networks, and derive functional modules from these clusters. A more direct approach has been adopted by Lin et al. [2006], who proposed a supervised learning approach for inferring protein function from interaction networks using the share neighbor idea. They show that the likelihood of two protein sharing a function becomes significantly higher if they have a high number of shared neighbors, as compared to just having a direct interaction between them. Motivated by this observation, they developed a probabilistic model for estimating the probability of a given protein being annotated with a certain function, by looking at the annotation of all proteins in the network that are known to be annotated with that function. The model estimates this probability conditional on at the number of shared neighbors between the target protein and each of these already annotated proteins, by modeling this as a set of independent normal distributions. In a training step, the parameters of these distributions are estimated for each function, and in the testing step, a probability is computed for each proteins carrying, as well as not carrying, the given function. A ratio of these two probabilities are then treated as likelihood scores for the protein-function pair, and are used as the final prediction. By evaluating this algorithm on an integration of several standard yeast interaction datasets for a set of FunCat functional classes, it is shown that the performance is better than the methods based on direct neighborhood-based annotation transfer. Thus, the use of common neighbors appears to be more beneficial than direct neighbors, since the former provides an effective method for incorporating the reliability of the interactions in the function prediction algorithm.

Finally, before moving on, it is important to note that the previous approaches [Samanta and Liang 2003; Brun et al. 2003; Lin et al. 2006] are very close in spirit to the shared nearest neighbor clustering algorithms in data mining [Ertoz et al. 2003; Jarvis and Patrick 1973]. These algorithms are based on a different similarity measure for the data points: the similarity between two points is the number of neighbors that they share in an interconnection network. These algorithms are known to be robust with respect to noise, regions of varying densities and clusters of varying sizes in the data [Tan et al. 2005]. Also, they are based on the paradigm of graph-based clustering, and thus may be useful for clustering protein interaction networks and making functional inferences from them.

McDermott et al. [2005] discuss a very innovative approach to function prediction that addresses a practical problem with protein interaction networks. It is known that the experimental procedure used to construct these networks are significantly labor-intensive, and thus, interaction networks are available only for some model organisms, such as yeast, fruit

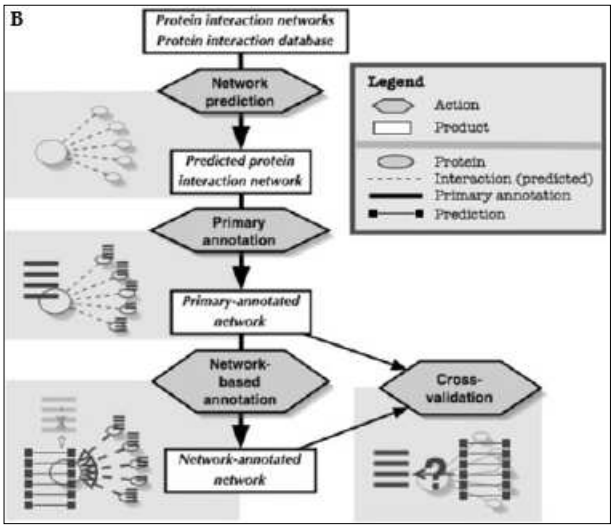


Fig. 23. Flowchart of the algorithm used to derive functional annotations from predicted protein interaction networks [McDermott et al. 2005]

fly and worm. As a result of this paucity of protein interaction data, it may become hard for research to exploit the rich knowledge they encode. Thus, McDermott et al. [2005] propose the use of *in silico* predicted protein interaction networks for function prediction using neighborhood-based approaches. Figure 23 shows a schematic flowchart of this approach. Initially, an interaction network is generated for the target organism using interaction information compiled from databases such as DIP [Xenarios et al. 2002], GRID [Breitkreutz et al. 2003] and PDB [Berman et al. 2000], through the *interolog* approach [Matthews et al. 2001; Yu et al. 2004]. The majority-in-the-neighborhood method [Schwikowski et al. 2000], and a more sensitive weighted version thereof that takes into account the link and functional annotation quality of a neighbor, were then applied to this network. The performance of the algorithm was evaluated using the precision and recall measures formulated in [Deng et al. 2004]. In the experiments, interaction networks for *D. melanogaster* (fly) and 50 other organisms were generated and annotated under this setup. Evaluation showed that the functions could be predicted with an average precision of about 70% with the weighted method, though the coverage was only 13% considering just the top-ranked prediction. However, even with this low coverage, GO primary annotations could be assigned for 60 and 132 previously unannotated fly and human proteins with an estimated precision of at least 65%. This showed that for organisms, such as fly and human, where not much interaction information is available, robust annotation can be derived from predicted networks. The authors have also set up the Bioverse web server [McDermott and Samudrala 2005] containing such predictions for a variety of organisms. In these respects, this was a path breaking piece of research.

It was mentioned earlier that one of the tasks in KDD Cup 2001 was the prediction of protein function from protein-protein interaction data [Cheng et al. 2002]. Since this was a contest, several groups submitted possible approaches for solving this problem. The entry that achieved the best performance adopted a two-step approach for this task. In

the first step, the original data was pre-processed using the RELAGGS system [Krogel and Wrobel 2001] to retrieve various attribute-value pairs for each gene, thus producing a new data set consisting of a real-valued vector for all the genes in the original data set. Since the original data was a relational data set, with each interaction being an individual entry, these new attributes were based on the immediate neighborhood of each gene, thus making this approach appropriate for the current category. This new data set was then learnt using an SVM classifier, and tested on a separate set of genes, which produced results as high as 93.6% accuracy for function prediction. However, more significant than the success of this strategy was the fact that 41 distinct data mining solutions were submitted for this task [Cheng et al. 2002], thus showing that data mining techniques have potential for solving the problem of predicting protein function from interaction networks.

Finally, the richness of protein interaction networks and the simplicity of the neighborhood approach has also motivated their use as an intermediate step in the function prediction process. In a study that uses machine learning techniques [Vert and Kanehisa 2002], these networks were used to extract meaningful features from gene expression data. The hypothesis adopted is that genes close to each other in a network are more likely to exhibit similar expression patterns, and hence have similar functions. Using rigorous kernel-based techniques, the problem is formulated mathematically, and solved to extract a similarity kernel from the network. In experiments, which employed networks derived from the LIGAND yeast pathway database [Goto et al. 2002], successful classification using SVM was achieved for classes earlier thought to be hard to identify through expression data, such as fermentation and nucleus organization [Brown et al. 2000]. In a similar study [Altaf-Ul-Amin et al. 2003], proteins identified as belonging to the same k-core of the interaction graph [West 2001], as well as to the same cluster derived from phylogenetic profiles, are assigned the same function. Finally, some recent studies have also tried to combine interaction networks with homology-based approaches to identify functionally related proteins [Espadaler et al. 2005; Okada et al. 2005]. The basic hypothesis in these studies is that if two proteins are interaction partners in a network of interactions, and also show sequence or domain homology, then there is substantial evidence that these proteins interact functionally. Thus, it can be seen from this discussion that the intuitively simple neighborhood approach has significantly impacted both the direct and indirect use of protein interaction networks for function prediction.

8.3.2 *Global optimization-based approaches.* Though the neighborhood approach is very attractive because of its simplicity, it suffers from some obvious limitations. For example, if a protein has an insufficient number of neighbors in the network, or its neighbors are not annotated, then it is difficult to make significant predictions about its function. The presence of contradictory annotations among neighbors also makes it difficult to arrive at a coherent prediction. To address these issues, several *global* approaches have been proposed. These approaches try to optimize, either directly or indirectly, an objective function defined on the whole network, which measures some global property that the network should possess once all its nodes have been annotated. Further details of these approaches follow.

One of the first papers that approached the problem from this viewpoint [Deng et al. 2003] used the theory of Markov random fields (MRF) to determine the probability of a protein having a certain function. This theory is used to determine the joint probability of the entire network with respect to a certain function. This formulation is transformed to that

of the conditional probability of a protein having a certain function given the annotations of its interaction partners. Finally, the Gibbs sampling technique is used iteratively to determine the stable values of this probability for each protein. As expected, this strategy outperforms the neighborhood-based approaches [Schwikowski et al. 2000; Hishigaki et al. 2001] in the functional annotation task for the MIPS interaction data for yeast. In one of the extension of this work, the same strategy was applied for the mapping of GO codes to proteins, with similar results [Deng et al. 2004]. In another extension [Lee et al. 2006], the MRF approach was generalized by using a diffusion kernel-based similarity between proteins in the network. This enabled the approach to transfer annotations from farther away proteins, in addition to only the neighboring proteins, weighted by their diffusion kernel-based similarity with the query protein. This generalized produced a non-trivial improvement in the accuracy of performance over several GO functional classes.

Another strategy based on MRFs is presented in [Letovsky and Kasif 2003]. This solution departs from [Deng et al. 2003] in the following two ideas:

(1) The probability of having a certain number of neighbors with a certain function is determined by a binomial probability distribution.
(2) Instead of Gibbs sampling, a heuristic version of the belief propagation algorithm is used to find stable values of these probabilities.

Notably, the first idea implies that the assignment of a label to a protein is a random process which satisfies the neighborhood constraints imposed by the network. This is unlike [Deng et al. 2003], which explicitly uses the number of interaction partners with the same and different labels to define the same probability. Hence, the results from [Letovsky and Kasif 2003] were expected to be inferior to those of [Deng et al. 2003], though a direct comparison is not possible since the former reported its results on the GRID interaction data [Breitkreutz et al. 2003]. On this data set, the algorithm achieved a high precision of 98.6% but a low recall of 21%.

Probably the most widely cited approach in this category is [Vazquez et al. 2003], which was also covered briefly above. Here, an objective function is defined for the whole network, which is a sum of the following variables:

(1) The number of neighbors of a protein having the same function as itself.
(2) The number of neighbors of a protein having the function under consideration.

Thus, this function estimates the number of pairs of interacting proteins with no common functional annotation. Since a high value of this function is biologically undesirable, it is minimized using a simulated annealing procedure. As expected, this approach outperformed the majority rule-based strategy on the *S. cerevisiae* interaction data from [Schwikowski et al. 2000], since the latter tries to optimize only the second factor above. An additional advantage of this approach was that multiple annotations of all proteins were obtained in one shot, unlike earlier approaches which ran independent optimization procedures for different functions.

In a recent paper, Sun et al. [2006] have described the MFGO (modified and faster global optimization) approach, that tries to reduce the computation requirements of [Vazquez et al. 2003]. The idea here is to redefine the objective function such that a protein is assigned multiple functions in one optimization run, as against a separate run for each function, as proposed in [Vazquez et al. 2003]. From experiments on four datasets, significant savings in computational time are observed by using MFGO, though the accuracy remains nearly

the same. Yet another approach [Leone and Pagnani 2005], motivated by the principle of Gibbs potential from physics, uses the same objective function as [Vazquez et al. 2003]. Here, belief propagation algorithm [Mezard and Parisi 2001] is employed to assign the probabilities of annotation of the proteins in the network with a certain function. However, due to the requirement of many more iterations to reach a stable solution, this strategy could not outperform [Vazquez et al. 2003] in a comparable number of iterations.

Another approach which attempts to achieve agreement between annotations of neighbors in interaction networks is presented in [Karaoz et al. 2004]. This study models interaction graphs as Hopfield networks, which are neural architectures often used in computational neuroscience [Hopfield and Tank 1986]. Under this model, an energy function is defined for each function in the GO hierarchy in terms of the weights of the edges in the network and the functional annotations of various proteins. Minimizing this function using an iterative gradient descent procedure leads to a maximally consistent assignment of the function to the proteins. The fundamental issue here was the method used to assign weights to the edges. Two such methods were implemented, namely a default weight of one and a weight equal to the absolute value of the correlation coefficient between the expression profiles of the two interacting genes. The whole procedure was run for each function in GO and results were evaluated in terms of the F-measure. This evaluation showed that, for 168 functions, more accurate predictions were made using the integrated network compared to those made using just the protein-protein interaction network taken from GRID [Breitkreutz et al. 2003]. New plausible annotations were also suggested for some proteins, thus illustrating the merit in integrating multiple information sources.

An abstract problem that has been proposed in the literature is the estimation of distances between proteins in Euclidean space from their relative positioning the interconnection network. This clearly is a hard problem and can not be solved using an ad-hoc technique. Hence, Tsuda and Noble [2004] learn this inter-protein distance matrix K by maximizing its von Neumann entropy $-tr(K \log K)$ [Nielsen and Chuang 2000]. Under pre-defined local and global constraints, this problem is transformed into its dual and is solved using standard convex optimization methods. Feeding this kernel matrix as input to an SVM program, function prediction is carried out using two yeast networks, namely the biological pathway-based network constructed by Vert and Kanehisa [2002] and the protein-protein interaction data reported by von Mering et al. [2002]. Similar results with an ROC score of about 0.80 are obtained, which justifies the accuracy of the derived distances, which can be utilized fruitfully by data mining techniques such as clustering and outlier detection.

It has been mentioned several times in this discussion that a very natural representation of interaction networks is a graph consisting of proteins as nodes and pairwise interactions as edges. Hence, it is expected that several approaches would apply graph-theoretic techniques to deduce functions from these graphs. However, reality is far from expectation. An innovative, and rare approach for this problem is described by Nabieva et al. [2005]. In this approach, the traditional max-flow min-cut algorithm for directed graphs [West 2001] is modified into an iterative flow algorithm for undirected graphs, such as interaction networks. In this algorithm, named FunctionalFlow, the sinks are proteins which are annotated with the function under consideration, while the others are sinks. Capacities of the edges are determined by the reliabilities of the experimental or computational techniques used to detect the corresponding interaction. Simulating this flow for a certain number of iterations allows the function to flow into all target nodes. A contribution of this

study is its comparative evaluation with respect to three other popular algorithms, namely Majority [Schwikowski et al. 2000], Neighborhood [Hishigaki et al. 2001] and GenMultiCut [Vazquez et al. 2003; Karaoz et al. 2004]. This evaluation is carried out on the GRID yeast interactions data set [Breitkreutz et al. 2003] with respect to 72 functional categories at depth 2 of the biological process hierarchy in GO. In terms of the ROC measure, FunctionalFlow shows the best overall performance. In particular, it outperforms Majority (and the others) for proteins with very few annotated interaction partners. Hence, this study overall has several significant contributions, namely a new function prediction method based on interaction networks, a novel implementation for an existing approach [Vazquez et al. 2003; Karaoz et al. 2004] and a comprehensive evaluation strategy.

Finally, the representation of protein networks as graphs has also resulted in the application of techniques from other fields, particularly from social network analysis [Wasserman and Faust 1994]. A social network is constructed by observing the entities interacting in a given environment and their patterns of interaction. Owing to the similarities in the structure of social networks and protein interaction networks, techniques from social network mining [Staab et al. 2005] have now found their way into the prediction of function from protein interaction networks. One such simple approach, namely the estimation of the role similarity of two individuals [Wasserman and Faust 1994], which correspond to proteins in this case, has already been applied for this task [Holme and Huss 2005]. In this approach, given two proteins, a simple iterative procedure is used to estimate how many pairs of proteins to which they are connected, either directly or indirectly, share a common annotation. Though this approach suffers from a quadratic time complexity in the number of nodes in the graph, and its results on the MIPS interaction data set [Mewes et al. 2002] are only slightly better than the very simple neighborhood counting method [Schwikowski et al. 2000], this study still makes an important contribution of demonstrating how techniques from distant fields such as social network analysis could be conveniently mapped to the field of protein function prediction from interaction networks.

Another field of computer science research that this field could benefit from is that of web search algorithms. Currently, an assumption underlying all the approaches for protein network analysis is that all the proteins have the same reliability for the information they provide. This may not always be appropriate, since in some cases, some neighbors of a protein may be providing redundant information, and it may make more sense to give a higher weight to the suggestions of the more informative neighbors. Algorithms designed for searching the web, such as finding *hubs* and *authorities* [Kleinberg 1999] and PageRank [Brin and Page 1998], that drive most of the successful search engines such as Google[13], handle this problem very well by techniques such as weighting heavily highly hyperlinked pages and adjusting redundant information coming from the same network domain. Though studies have now started applying such techniques for analysing various properties of protein sequences, such as remote homologs and motifs [Noble et al. 2005; Kuang et al. 2005], they have not yet been applied for function prediction. Given the immense success of search engines such as Google, both in accuracy and scalability, the application of appropriate modifications of their underlying algorithms may lead to useful solutions for the problem of function prediction.

The above discussion shows that a wide variety of approaches based on principles of global optimization have been proposed in the literature and many more are in the pipeline.

[13] `http://www.google.com`

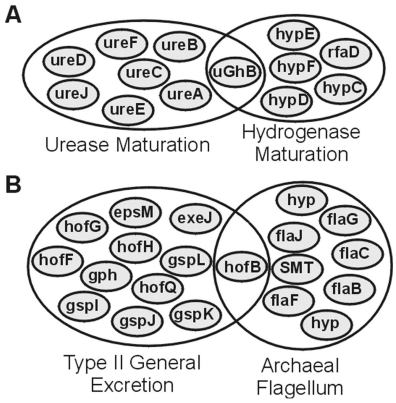


Fig. 24. Two examples of *linkers* identified in the clustering study of [Snel et al. 2002]. In these examples, *uGhB* and *hofB* connect two orthologous groups each shown by ellipses. (Figure taken from [Snel et al. 2002])

The most accurate results in the field of function prediction from interaction networks have also been achieved by these approaches, which is intuitively acceptable since they extract the maximum benefit from the knowledge of the structure of the entire network.

8.3.3 *Clustering-based approaches.* It has been reported by earlier studies that protein interaction networks often contain dense regions that contain large number of connections between the constituent proteins [Schwikowski et al. 2000]. These regions are often hypothesized to represent functional modules, which are natural units in which proteins perform their function. Clustering is a very effective approach for finding groups of similar points in a given data set, corresponding to a suitable definition of similarity. When the original data consists of points laid out as a graph, these clusters correspond to sets of highly interconnected connected points, and several algorithms have been designed for extracting these clusters from graph-based data [Brandes et al. 2003]. Since this form of clustering can be directly applied to protein interaction networks, some approaches have been proposed in the literature, that model the interaction network as a graph and apply graph clustering algorithms to it to group functionally similar proteins into modules. These modules can then be used for the annotation of uncharacterized proteins in them. The following discussion describes these approach in detail.

Snel et al. [2002] constructed a network from genomic associations detected on the basis of conserved gene order in genomes [Huynen et al. 2000], motivated by the conjecture that these associations reflect functional association between proteins [Dandekar et al. 1998]. Using a previously devised methodology [Snel et al. 2000], the genes in this network were clustered into orthologous groups. Again, when these groups were clustered using single linkage, a giant cluster of 1611 groups emerged. In the process of analyzing this large cluster, several genes named as *linkers* were identified. These linkers essentially are low-degree nodes in the interaction graph, as shown in Figure 24, that are members of two clusters (orthologous groups in this study) and thus lead to the merging of two sub-clusters in the giant cluster when the single linkage clustering algorithm is used. Thus, when 425 such linkers were eliminated, 265 sub-clusters of the large cluster were discovered, about 70% of which had a uniform functional composition according to COGs categories [Tatusov et al. 2003]. These subclusters can now be used for functional classification.

In another application of clustering [Dunn et al. 2005], the Edge-Betweenness algorithm [Girvan and Newman 2002] was used to cluster the Lehner dataset of human protein interactions [Lehner et al. 2004] to derive 21 clusters, that differed reasonably well in the GO annotations assigned to them, such as transcription, cell cycle regulation and mRNA processing. It is found that the method is robust to false-positive interactions, since multiple interactions are used to identify a cluster in the Edge-Betweenness algorithm. Hence, this study successfully extended the clustering-based functional analysis methods to human protein interaction networks, which is significant from a medical point of view. However, a significant disadvantage of this scheme is that the deletion of edges implies the loss of experimental knowledge, which is not advisable in this domain.

Besides prediction of individual protein functions, clustering can also be used for finding functional modules, which are groups of proteins that function together. Corresponding to this idea, Rives and Galitski [2003] present one such strategy for discovering functional modules in yeast protein networks. In this paper, a single-link hierarchical clustering algorithm was used to cluster the proteins in a network, with the length of the shortest path being the distance between two proteins. This algorithm is expected to resist the effect of false positives since spurious interactions are more likely to lie on spurious longer paths between proteins rather than the shortest ones. This algorithm was applied to three different types of yeast networks and successful results were obtained:

—Protein signaling network: Discovered clusters correspond to modules of signaling pathways such as the *Ras*-pathway.

—High-throughput interaction network: Some known modules such as *Lsm8* also emerged. Also, in conjunction with cellular localization data, *hub* proteins, which are highly connected nodes in the interaction graph, were also discovered

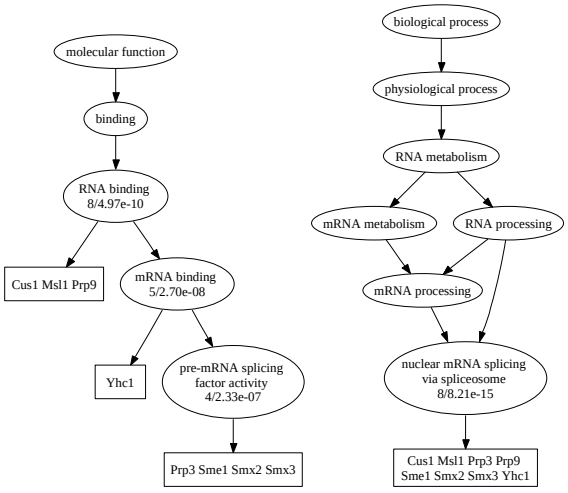
—Filamentation network: Known modules such as *Snf* and *Cdc28* emerged, and new proteins were associated with known clusters, such as *Yer124C* with *fMAPK*.

Thus, the versatility of the algorithm was shown through its effectiveness on multiple types of protein interaction networks.

In another clustering-based analysis of the yeast interaction network [Pereira-Leal et al. 2003], a different perspective was adopted. Here, the original interaction graph was transformed to its line graph [West 2001], which was clustered using the flow simulation-based TribeMCL algorithm [Enright et al. 2002]. Upon transformation back into the original graph, overlapping protein clusters were obtained, corresponding to the general biological understanding that one protein can be involved in multiple functions. Using interactions from DIP [Xenarios et al. 2002] and a weighted entropy metric, it was concluded that the clusters were homogeneous according to several functional schemes.

Overall, the success of these strategies in a wide range of networks, and an equally varied set of classification schemes reinforces the utility of clustering for interaction networks.

8.3.4 *Association Analysis-based Approaches.* Clustering is the most widely used member of the general category of unsupervised data mining algorithms. Another significant member of this category is association analysis, which comprises of techniques for the identification of frequently occurring patterns in the given data set, where the definition of a *pattern* depends on the type of the data being considered [Tan et al. 2005]. In the context of graphs, these patterns correspond to frequently occurring subgraphs in a set of graphs [Kuramochi and Karypis 2004]. Extending this concept further to a set of pro-



(a) Annotation using the GO molecular function ontology

(b) Annotation using the GO biological process ontology

Fig. 25. GO annotations of the hyperclique pattern {Cus1, Msl1, Prp3, Prp9, Sme1, Smx2, Smx3, Yhc1} found via the methodology of [Xiong et al. 2005]. The annotations in both the cases are fairly coherent, particularly in the biological process, which is the most relevant for function prediction from complexes.

tein interaction networks leads to the idea that these patterns may correspond to functional modules and may be used for functional discovery.

This idea was adopted by [Hu et al. 2005], who proposed CODENSE, an algorithm to discover coherent, dense and possibly overlapping subgraphs which occur frequently in a set of graphs. CODENSE constructs a summary graph for the set and a subset of frequent dense subgraphs from this graph, and then, the MODES algorithm [Hartuv and Shamir 2000] is used to extract the true frequent subgraphs in the original set. In order to ensure sufficient data, an indirect method of protein network construction was used in which 39 microarray datasets for yeast were transformed to networks using the correlation between expression profiles of pairs of proteins. For evaluation of performance, functions were assigned to 448 genes in the derived subgraphs using the majority rule, and the resulting accuracy was 50%

In another recent application of association analysis for finding coherent functional modules, Xiong et al. [2005] extract functional modules from protein complex data using the concept of hypercliques [Xiong et al. 2003]. A *hyperclique* is a set of frequently occurring objects (proteins in this case), such that the confidence of every rule formed using these objects exceeds a certain threshold. Also, efficient algorithms exist to extract these hypercliques from a large binary data sets. These binary data sets can be easily generated for protein complex data by treating each protein in the complex as an attribute with value 1, and each absent protein as 0. Upon running the hyperclique algorithm on binary versions of benchmark protein complex data sets [Ho et al. 2002; Gavin et al. 2002], several accurate functional modules were extracted. The functional coherence of these modules was

shown via their annotation with GO codes, both from the molecular function and biological process ontologies (see Figure 25).

The above approaches focused on identifying sub-networks of proteins that are highly conserved between a set of interactions networks. Another perspective on the frequent sub-network problem that can be considered is that of finding frequent sub-networks within a given interaction network, which may represent structurally important groups of interactions. A brute force solution to this problem is exponential in computational complexity. Hence, Chen et al. [2006] proposed an efficient algorithm *NeMoFinder*, inspired by the Apriori algorithm [Tan et al. 2005], for this problem, and showed that these motifs were generally constituted of highly reliable interactions. A uniqueness property was also defined for these motifs, in order to ensure that they were not randomly produced, which may be a problem for motifs of small size. However, this algorithm is not directly usable for inferring protein function, since it only identifies frequent subgraphs in the interaction graph whose nodes are not labeled by the proteins they represent. Thus, in order to use these motifs for function prediction Chen et al. [2007] presented a modified version of the NeMoFinder algorithm, where each motif was labeled by the proteins the nodes represent, and these proteins were annotated by their annotations. Then, these motifs we clustered on the basis of the similarity of the GO terms that they were enriched with, thus producing a set of groups of proteins, each of which carried similar GO annotations, and occurred frequently in the input network. These groups were generally enriched with related GO terms, thus showing the ability of this idea to extract functionally important components of an interaction network, even if the components are disjoint in the network. Thus, this algorithm marked a difference from the rest of the algorithms, particularly those based on clustering, which are heavily dependent on the connectivity of an interaction network.

Overall, these studies have highlighted the potential of the application of association pattern discovery algorithms to the problem of functional discovery from protein interaction networks, and more such studies can be expected in the future.

8.4 Discussion

From the above discussion, it can be observed that numerous innovative approaches have been proposed for the computational analysis of protein interactions networks, particularly for the prediction of protein function, and more are being published by the day. In particular, the results obtained from global optimization-based approaches for this problem have been impressive. These are expected to improve further as more techniques from computer science for the analysis of such networks, such as social network mining and web search techniques, are adapted into this domain.

9. LITERATURE AND TEXT

9.1 Introduction

As in all other research communities, researchers in the fields of biology and medicine publish the results of their research in various journals and conferences. As a result, over the past, a huge repository of knowledge has been created in the form of papers, books, reports, theses and other such texts. Clearly, these repositories contain a huge amount of information about important biological concepts such as protein structure and function, cancer-causing genes and several others. Thus, there is great utility in the mining of these repositories and retrieval of useful information.

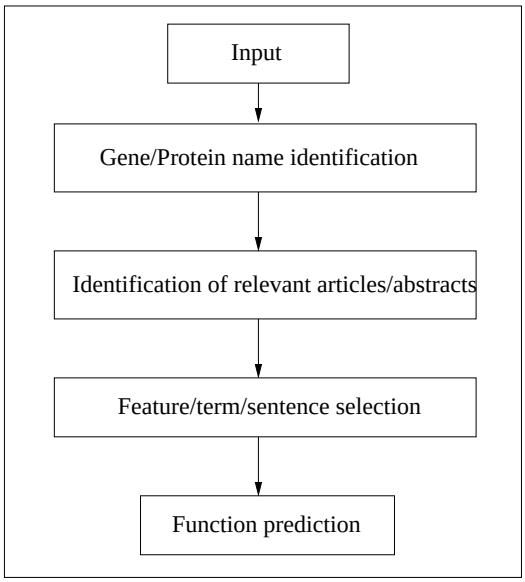


Fig. 26. The general architecture of a literature mining-based function prediction system

The most widely used database of research articles is MEDLINE[14], which currently contains about4 twelve million articles published in various journals and conferences of biomedical research, and also provides a web-based interface named PubMed[15]. Each document in MEDLINE is assigned a unique ID and prominent databases, such as SWISS-PROT [Boeckmann et al. 2003], SGD [Dwight et al. 2002] and FlyBase [FlyBase Consortium 2003], are linked to it through these id's. In addition, each document is annotated with representative MeSH[16] terms that are derived from a manually-designed ontology-based thesaurus. Owing to its extensive coverage and well-connected structure, MEDLINE is by far the most widely used source of articles describing biomedical research.

9.2 Existing Approaches

The primary target of the field of biomedical literature mining is to extract useful and valid biological information from the vast repositories of biomedical research literature such as PubMed. The form of information we are most interested in are the functions of various genes and proteins. Function prediction via literature mining aims to uncover those functions that have not been reported in the literature, and several approaches have been proposed in this direction [Nair and Rost 2004]. These approaches are the subject of discussion in this section.

The general architecture of a literature mining-based function prediction system is shown in Figure 26. The various "identification" modules in the architecture represent the prime challenge for such systems, namely the variation in the use of language by different authors. This variation is observed not only in the organization of the text in the articles,

[14] http://www.nlm.nih.gov/pubs/factsheets/medline.html
[15] http://www.pubmed.gov
[16] http://www.nlm.nih.gov/mesh/

but also in the nomenclature. For instance, $EphB2$, a protein involved in signaling in the brain, was known by various names such as $Cek5$, Nuk, Erk, $Qek5$, $Tyro6$, $Sek3$, $Hek5$ and Drt in the literature, before its official name was adopted [Nature Editorial 1999]. It is widely recognized that such variations make the task of uncovering useful information from literature very challenging [Blaschke and Valencia 2003; Nair and Rost 2004].

Despite these challenges, several powerful text analysis approaches exist in data mining, that can be used to extract useful knowledge about protein function from research articles and other texts. This analysis can be performed at various levels:

—**Shallow:** Only the abstract or introductory text is used in the analysis.

—**Deep:** The complete document or text is utilized.

Text analysis approaches for biomedical literature exist at both levels. Another classification is based on the extent of English vocabulary considered for this analysis. The two classes here are the following:

—**Free:** The analysis is performed on the complete unstructured text from the articles/abstracts.

—**Restricted:** Only a fixed-size vocabulary, such as the set of GO codes, is analyzed.

Though the above categorizations are useful, a more useful categorization of the approaches that predict function from biological literature is based on the underlying techniques and the fields these techniques are adopted from. The most dominant fields that have influenced function prediction from literate are as follows:

—**Information Retrieval (IR)**: The field of information retrieval deals with the automatic extraction of information from a large source of data, a repository of research articles in our case, starting with a natural language query issued by a human [Rijsbergen 1979]. Given this data, IR approaches essentially involve the estimation of the relevance of documents in this data set to the given query, mostly at a syntactic level, and then ranking them to find the most relevant results.

—**Text Mining**: Text mining is defined as the process of extraction of semantically interesting and non-trivial knowledge from unstructured text [Weiss et al. 2004]. The techniques in this field involve the use of intelligent data analysis techniques, such as clustering and classification, for analyzing text data. Thus, these techniques are more robustly able to handle large variations in data, which is a significant problem for research literature because of the different writing styles of the authors of different articles.

—**Natural Language Processing (NLP)**: In many cases, the use of text mining techniques is not enough, and it is pertinent to incorporate natural language understanding [Allen 1995] into the analysis algorithm. The techniques in this field deals with the modeling and analysis of natural language use by humans, so as to retrieve semantically interesting knowledge from text.

We have categorized the approaches proposed for predicting protein function from literature into the above three categories, which are discussed in the following sections.

9.2.1 *IR-based approaches.* As mentioned above, information retrieval deals with the retrieval of the most relevant documents in response to a query. This paradigm can be easily applied to the problem of function prediction from literature, where the query is the name of the gene/protein, and the data source is a repository of research articles.

The earliest solution to this problem was proposed by Tamames et al. [1998]. In this approach, which is mostly statistical, the frequency of various keywords are recorded for proteins, which were categorized into three functional classes: Energy, Communication and Information. These keywords are obtained from the SWISS-PROT database, in which the entry for each protein also contains some keywords manually identified from the literature. Using these frequencies, unclassified proteins are assigned to one of the three classes. Next, the keyword frequencies are updated using this new set of classified proteins. This process is carried on till there is no significant change in the frequencies. Using this approach, a coverage of 52% and a precision of 82% was estimated for a data set consisting of the proteins in the *Mycoplasma genitalium* genome. Though the coverage appears to be low on this data set, it becomes significant when compared to the manual classification of these proteins, which resulted in a moderate coverage of 63%.

The next step in this direction was the ProFAL (PROtein Functional Annotation through Literature) system [Couto et al. 2003]. In the retrieval phase, the documents in PubMed linked to the entries of GenBank [Benson et al. 2004], SWISS-PROT [Boeckmann et al. 2003] and PDB [Berman et al. 2000] databases are retrieved. In the extraction phase, the enzyme under consideration is annotated with all the GO terms occuring in these documents. The validation of these annotations are mostly manual, and one such evaluation of 173 annotations for enzymes in the CAZy database [Coutinho and Henrissat] by an expert curator reported 55% precision. The recall of the system was also low (40%), mostly because of the lack of bibliographic references for some enzymes.

MILANO [Rubinstein and Simon 2005] is another system which links multiple databases to annotate genes. It takes as input a set of gene identifiers and a set of custom terms. In order to handle the variation in the gene names, the gene names are expanded using the LocusLink database[17] , and the co-occurrences of these new names with the custom terms are counted in the PubMed and GeneRIF [Mitchell et al. 2003] document databases. The resulting associations are ranked according to the frequency of co-occurrence and presented to the user on a web-page. In a case study, it is verified that MILANO can indeed identify the genes most affected by the over-expression of the $p53$ gene. Besides being a sucessful application of a software engineering architecture in literature mining, the two biggest strengths of MILANO are its ability to use aliases for gene names, and its use of the GeneRIF database [Mitchell et al. 2003], which contains about $90,000$ summaries of articles about known genes.

As can be seen from the above descriptions, IR-based approaches mostly use the syntactic information in documents to extract useful information from text. The need to use semantics in these approaches led to the application of solutions based on text mining to the problem of automated function prediction from the literature. These techniques are reviewed next.

9.2.2 *Text mining-based approaches.* As mentioned before, text mining holds great potential for the analysis of biomedical literature because of its ability to utilize the semantic content of a document, and robustness to variations in writing styles and nomenclature. This section discusses several text mining-based approaches that enable the prediction of the functions of proteins from literature.

For a long time, the most popular technique for predicting protein function was the trans-

[17] www.ncbi.nlm.nih.gov/LocusLink

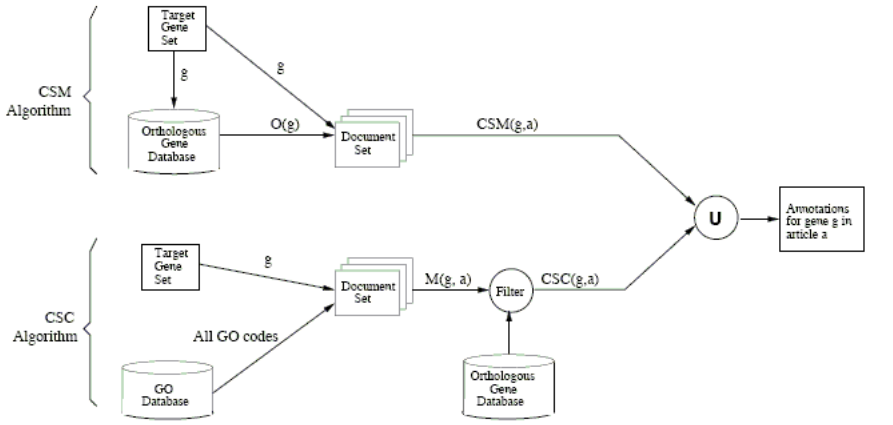


Fig. 27. Cross-species annotation procedure adopted in [Stoica and Hearst 2006]

fer of annotation from a homologous protein to the query protein (Sections 3.2 and 3.3.1). Hence, it was natural for early text mining approaches to utilize this notion for making confident predictions. One such approach is presented by Renner and Aszodi [2000], who propose a pipeline for the prediction of functions of novel gene products. The portion of this pipeline of interest to us are its final two steps. In the first step, a novel protein is input to a similarity search across multiple databases, such as SWISS-PROT [Boeckmann et al. 2003], PIR [Wu et al. 2003] and PROSITE [Hulo et al. 2006]. The results of these searches are stored as HTML documents. Next, the representative terms from all the documents are extracted and clustered on the basis of frequency of co-occurrence. Guided by the hypothesis that two documents are similar if they contain terms that are related to the same biomedical concept, the documents related to each protein are clustered using the term clusters derived earlier. An inspection of the clusters can then be used to assign function(s) to a protein. It was observed that, for most of the proteins, all the documents clustered into a single cluster, illustrating their conceptual coherence. Thus, even though the final step needed user intervention, this approach was a successful application of document clustering to the bioinformatics domain.

The idea of orthology has been used for cross-species annotations in an interesting recent paper [Stoica and Hearst 2006]. Here, the final GO annotations of a gene are assigned on the basis of the annotations of its orthologs. Two methods for doing this are proposed here, namely Cross Species Match (CSM) and Cross Species Correlation (CSC). CSM uses the GO annotations of an ortholog also as annotations of the query gene, CSC only uses those annotations that are significantly correlated with the dominant annotations of the query gene's orthologs. The complete algorithm is depicted in Figure 27. It is expected that CSM will have a high precision, while CSC will improve the recall at the cost of precision. This is indeed the pattern observed for the annotation of EBI Human [Camon et al. 2003] and MGI Mouse [Blake et al. 2003] databases, where the final set of annotations is the union of CSM and CSC annotations. This work could be especially valuable for genes of organisms that are not as well documented.

Raychaudhari et al. [2002] apply the data mining technique of classification to the pre-

diction of functions of genes on the basis of the documents they are associated with. In the first part of this study, three document classifiers, namely, the maximum entropy, naive Bayes and nearest neighbor classifiers, were constructed for 21 classes using training abstracts extracted from PubMed. Words co-occurring with the GO codes denoting the classes were identified using a χ^2-test and were used as features. After classifying a held-out test set, it was concluded that the maximum entropy classifier, with an accuracy of 72.8% is most appropriate for the functional classification task. In addition, it was verified that the probability of each classification being correct, as assigned by the maximum entropy classifier, can be used to rank the predictions. These conclusions are in agreement with results reported earlier in the statistical NLP community [Nigam et al. 1999]. Thus, this study presents a strong case of the use of text classification for function prediction.

In another simple approach, the nearest neighbor classification algorithm is applied for functional classification of proteins [Keck and Wetter 2003]. Upon receiving a query, a simple text-based similarity is calculated between its documents and those of proteins identified in a BLAST [Altschul et al. 1997] search, and the top functional keywords are transferred. The results obtained for a variety of databases such as GenProtEC [Serres et al. 2004] and MIPS [Mewes et al. 2002] are not impressive, with a precision over 0.9 being obtained only for a low recall of about 0.4. The suggested reasons for these results are the presence of homologues and inconsistent examples. However, the basic weakness of the approach is the crude similarity function, which is the backbone of any instance-based learning algorithm such as k-nearest neighbor.

Finally, a problem related to function prediction is the prediction of functional relationships between genes. Shatkay et al. [2000] addressed this problem by representing each gene by a set of abstracts and then comparing the representative sets of all pairs of genes to determine which ones are functionally related. Each set of documents, which are in turn assembled using a document similarity algorithm is converted into a binary vector, and the similarity between these vectors is calculated using cosine similarity. If the similarity is higher than a certain threshold, the two corresponding genes are declared to be functionally related. Upon testing on a set of cell cycle-regulated genes in yeast, it was found that the clusters corresponded directly to those experimentally determined by Spellman et al. [1998] for the same set. An additional advantage of this algorithm is that it yields an executive summary in the form of the most relevant terms for every gene.

9.2.3 *NLP-based approaches.* In some cases, text mining techniques are not effective for extracting protien functoin information from documents, since they are essentially unguided, and are thus unable to assign specific meanings to individual components of the content of a document. Thus, in order to perform a more guided analysis of the documents, it is necessary to incorporate natural language understanding into such an analysis system [Allen 1995]. The approaches discussed next attempt to achieve this goal.

Koike et al. [2005] try to capture the protein-functional term relationship as an ACTOR-OBJECT one, thus eliminating the false positives generated by hypothesising just on the basis of the co-occurrence of the protein and functional term in a number of documents. This strategy essentially consists of two steps. In the first step, same and similar meaning terms for each GO class are extracted from the available text using various techniques. This component thus handles the variability in function name in the literature and hence increases the recall. In the next step, gene names are identified in the text, and the sentences in which they appear are subjected to shallow parsing. If the gene appears in the ACTOR

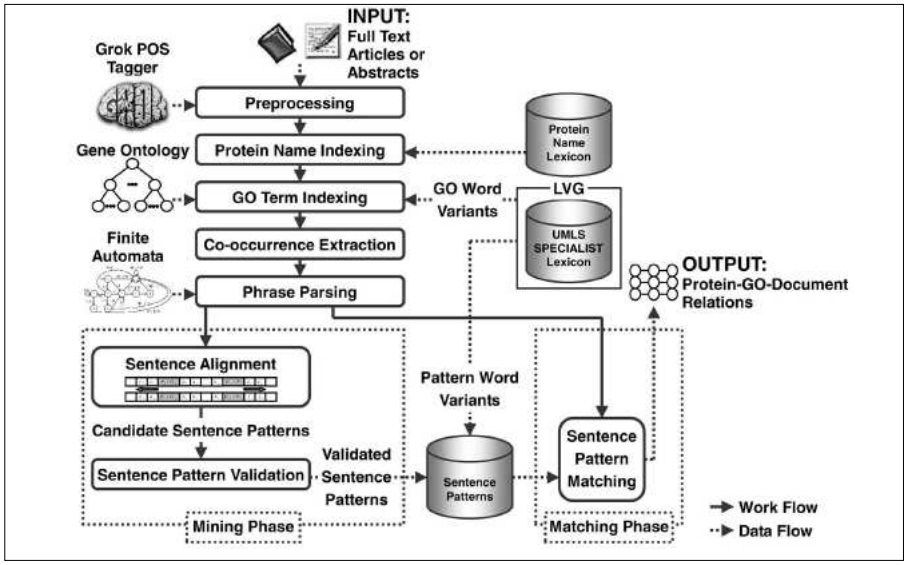


Fig. 28. NLP-based approach followed by Chiang and Yu [2005]

position and the functional term appears as the OBJECT, then a gene-function hypothesis is presented. In order to handle complicated sentence structures, some rules are defined for identifying the ACTOR(S) and OBJECT(S) in the parsed structure. Besides generating reasonably accurate hypotheses for the yeast and human genes, which are verified manually, the main reasons for low recall (less than 50%) are discussed. Some such factors are parser errors, gene name recognition errors and function and gene name not occurring in the same sentence. Addressing these problems will lead to an improvement in the system's recall.

Some of these issues are addressed by Chiang and Yu [2005], whose strategy for solving the function prediction problem is shown in Figure 28. Here, three types of variants of GO terms are defined, namely *morphological*, *syntactic* and *semantic* variants, and are extracted using various tools and techniques such as lexicons, mining of variation rules from literature and mapping of other classification systems to GO, respectively. Next, sentences in which a protein name and GO term co-occur are extracted and subjected to shallow parsing. It is observed that the descriptions of protein functions in natural language patterns follow certain patterns such as *<protein> participates in <GO>* and *<protein> is localized to <GO>*. These sentence patterns are mined using sentence alignment [Barzilay and Elhadad 2003], and patterns with a high support and confidence are used for finding the functions of uncharacterized proteins described in a separate set of test sentences. These patterns, like association patterns in data mining [Tan et al. 2005], capture the consistency in variability feature of language use by authors. In a quantitative evaluation of the system, it is concluded that processing morphological, syntactic and semantic variants indeed assists in achieving the best performance. At certain values of minimum support and confidence, the precision, recall and F-measure values obtained are better than the best submission in the BioCreAtIvE competition [Blaschke et al. 2005], thus illustrating the power of this approach.

The discussion above details well-designed systems that are able to handle greater variation in the content of the texts analysed than their counterparts in the text mining and IR domains since they are explicitly equipped with knowledge about natural language use, and thus have a greater precision at the task of function prediction. However, recall is a major issue for these systems as it is very difficult to capture all possible variations in grammatical rules. Future research in this field is expected to concentrate primarily on improving the recall of such systems.

9.2.4 *Keyword search.* In addition to the approaches discussed in the previous sections, which perform a *free* analysis of the text of research articles to prediction function, there exist keyword extraction/annotation systems that examine a restricted vocabulary used in certain databases such as SWISS-PROT [Boeckmann et al. 2003] and SGD [Spellman et al. 1998], unlike the vast vocabulary of unstructured text targeted by the former approaches. These keywords usually contain information about the function of a protein, and thus the results of these keyword extraction systems can be treated as hints for protein function prediction. Some such systems are discussed in this section.

The earliest study in this area is described in [Andrade and Valencia 1997], where a simple strategy for ranking keywords for a set of disjoint protein families is proposed. Sets of abstracts for these families are obtained from MEDLINE, and a z-score is calculated for all the words (except the stop words) using their normalized frequency of appearance in the abstracts correponding to each family. For each family, words and sentences are ranked and extracted according to the z-score and the average z-score of the constituent words, respectively. Examination of the *plastocyanin* and *xylose isomerase* classes indicates that the algorithm performs well in terms of the number of keywords and the relevance of the sentences extracted. The results are analyzed further in [Andrade and Valencia 1998], and it was found that, for the 71 classes considered, more informative keywords than the SWISS-PROT keywords are extracted for 16 classes, and z-score is able to discriminated between words characteristic of some classes and those that occur generally in the literature. However, the test set considered for this algorithm is too limited, in terms of ambiguity, for any reliable conclusion to be drawn.

Fleischmann et al. [1999] restrict themselves to SWISS-PROT keywords and attempt to annotate uncharacterized proteins in the TrEMBL database with relevant keywords. In the first step of this approach, annotated proteins in the SWISS-PROT database are grouped on the basis of which PROSITE patterns [Hulo et al. 2006] they match, and the common annotations from each group are retrieved. In the annotation step, the common annotations are transferred to the unannotated proteins that satisfy the condition of a rule. Since pattern matching is the integral step in this approach, some methods, such as ensuring high statistical significance of the match, are adopted to eliminate false positive matches. Using this approach, the annotated fraction of TrEMBL [Boeckmann et al. 2003] increased from 32% to over 51%, which is a significant improvement in coverage.

Kretschmann et al. [2001] explore the use of the C4.5 decision tree construction algorithm to automate the assignment of keywords to protein entries in the SWISS-PROT database. Once again, the features used are patterns from the Pfam [Sonnhammer et al. 1997] and PROSITE [Hulo et al. 2006] databases. In order to provide substantial training data to the learner, proteins from SWISS-PROT are categorized into ten classes on the basis of the signatures in the InterPro database [Apweiler et al. 2000]. For each class, decision trees are constructed for all keywords, if possible, and rules are derived from these trees,

and their confidence is estimated using stastical methods. The validity of these rules is tested by a ten-fold cross-validation experiment. Not surprisingly, it was concluded that rules with a higher confidence have a high precision and low recall, and vice-versa. More significant is the fact that almost 100% precision is obtained for all classes. However, this may be a result of the fact that the ten classes used for testing are disjoint, and hence, there is not much ambiguity in the assignment of keywords.

Another application of keyword classification to function prediction is presented by King et al. [2003]. In this approach, a binary matrix is constructed from the training data, indicating which genes (rows) are annotated with which GO codes (columns). The hierarchical structure of GO was also taken into account by using the idea that a gene annotated with a certain GO node should also be annotated with all ancestors of the latter. Using this matrix, a decision tree classifier M_{DT} and a bayesian network classifier M_{BN} were learnt. In order to test these classifiers, a ten-fold cross validation experiment was conducted on the SGD [Dwight et al. 2002] and FlyBase [FlyBase Consortium 2003] annotation databases. ROC analysis indicated that M_{DT} outperforms M_{BN} at low false positive rates, and vice versa at high false positive rates. Manual assessment also indicated that in many cases, the predictions made led the assessor to related literature. This is a significant step towards automated database curation [Seki and Mostafa 2004].

The latest study in this direction appears in [Perez et al. 2004]. The basic methodology behind this approach is the establishment of links between literature databases such as MEDLINE and protein databases such as SWISS-PROT. Particular use of MeSH terms, which are ontology-based annotations of abstracts in MEDLINE, is made. Using the fuzzy thesaurus model [Miyamoto 1990], three mappings are derived: between MeSH terms and GO annotations of the SWISS-PROT entries (obtained from the GOA project [Camon et al. 2003]), between MeSH terms and SWISS-PROT keywords, and between keywords derived from the abstracts to the SWISS-PROT keywords. Once these mappings are available, it is easy to annotate uncharacterized protein sequences. It is specifically noted that, though precision and recall results are inferior to those reported by Kretschmann et al. [2001], the results are still substantial since only literature and no supporting databases are used to annotate proteins. This is an advantage of this approach.

From the detailed description of the various keyword annotation systems, it can be observed that the underlying methodology here is significantly simpler than that adopted for function prediction. This is a consequence of the use of a restricted vocabulary that eliminates the need to handle variations of different kinds. However, this research is still important in order to disseminate and assimilate research results quickly, a task that becomes substantially harder if every researcher adopts a different terminology.

9.3 Standardization Initiatives

In the previous few sections, we saw how different approaches for mining the functions of proteins try to handle the problems arising out of the free use of language in biomedical literature, usually. These approaches are based on different hypotheses, and are also evaluated on different test data. This makes it difficult to perform a comparative evaluation of the various literature mining techniques for protein function prediction. To make this comparison possible, it is necessary that all the algorithms be altered to address the same task and the evaluation be carried out on a common test data set. This necessity led to the organization of two literature mining contests, namely BioCreAtIvE [Blaschke et al. 2005] and the genomics track of the Text REcognition Conference (TREC) held in 2003 [Hersh

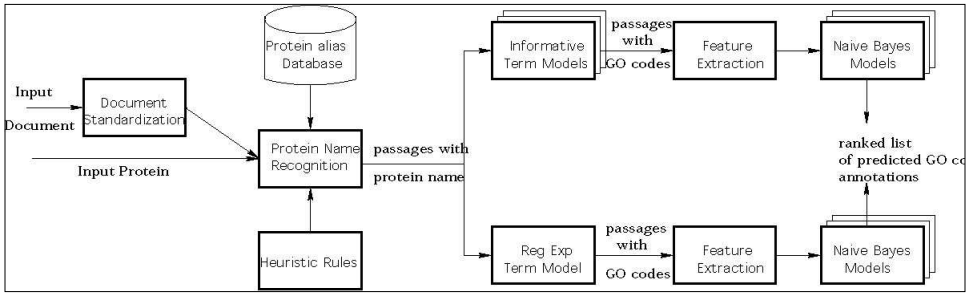


Fig. 29. Statistics-based approach followed by Ray and Craven [2005]. This was the general scheme followed by all submissions to BioCreAtIvE, with variations in individual modules

2004]. These contests are discussed in detail in this section.

9.3.1 *BioCreAtIvE.* The Critical Assessment for Information Extration in Biology (BioCreAtIvE) contest [Blaschke et al. 2005] was organized by a team of curators of biological databases and text mining researchers working on the problem of automatic information extraction from biomedical literature. The main motivation behind BioCreAtIvE was the development of common standards and shared evaluation criteria to enable comparison across various literature mining techniques. Two tasks were defined as a part of the contest, the first of which was the identification of gene (or protein) names in a given data set. Owing to the popularity of this task, and the high number of submissions [Yeh et al. 2005], a more detailed task further consisting of three sub-tasks was defined for the contest. These subtasks were:

Task 2.1. : Find pieces of relevant in article text that support a given GO annotation for a certain protein.

Task 2.2. : Given an article, find the most appropriate GO annotation for a certain protein and relevant text supporting it.

Task 2.3. : From the whole set of articles, find the most relevant articles and the most appropriate GO annotation for a given protein.

Thus, essentially, the solution for each sub-task required an extension of the solution for the previous sub-task. The training data for the whole task was essentially the entire GOA database [Camon et al. 2003], which provides GO annotations for various proteins as well as the MEDLINE id's of the documents(s) which support this annotation. The test set was a set of 212 full text articles from the *Journal of Biological Chemistry* (JBC)[18]. A common training and test set ensured that the best approach wins. The most successful of these approaches are dicussed below.

A fully statistical approach to the problem is taken by Ray and Craven [2005]. A schematic diagram of the strategy followed in this work is shown in Figure 29. To initialize, the original training set is supplemented by incorporating similar data from the SGD (yeast) [Dwight et al. 2002], FlyBase (Drasophila) [FlyBase Consortium 2003], WormBase (C. elegans) [Harris et al. 2004] and TAIR (Arabidopsis) [Huala et al. 2001] databases. Using this augmented training set, the most *informative terms* are identified for each GO code

[18] `http://www.jbc.org/`

using a χ^2-test on the contigency table recording the co-occurrence and non-co-occurrence of all the unigrams, bigrams and trigrams across documents associated with a certain GO code. When given a novel protein, the most appropriate GO assignment is assigned by summing the χ^2 values of all the informative terms for a GO code. This model is further enhanced by combining it with a naive Bayes classifier based on features extracted from raw text that capture the notion of protein name-GO code association. The resultant system is shown to be very competitive compared to the other submission, with respect to precision and recall. The most significant conclusion from this study is that supervised learning methods can be effective for the task of function prediction from textual data if sufficient data is provided at the learning stage.

Rice et al. [2005] approached Task 2 from a document classification perspective and used support vector machines (SVM) for the assignment of GO terms to proteins. For this purpose, features in the form of significant terms were extracted from each document using the C-value method [Frantzi et al. 2000]. This method combines linguistic patterns with statistical analysis to suggest the most significant terms in a document. Next, a feature-value vector for each document is then constructed using inverse document frequency (*idf*) weighting [Rijsbergen 1979] for each synterm. Finally, one SVM was trained for each GO term using the initial training data provided by the organizers. During the testing phase, these classifiers were used to rank the entity considered in the correponding sub-task. Unfortunately, performance was poor, with low precion for tasks 2.1 and 2.2, and only average precision for task 2.3, which was evaluated for a very small number of test cases. Still, a significant conclusion from this study was that relevant knowledge can be extracted from a set of documents even if it has not be explicity stated in a single document, using intelligent text indexing schemes such as *idf*.

Some other groups only presented solutions to the first two subtasks of Task 2, i.e., they addressed only the problem of assigning a function to a gene given an article of interest about it. [Ehrler et al. 2005] present the solution that achieved the best results for Task 2.2. This solution carries out both a vector space- as well a regular expression-based retrieval of the best GO terms corresponding to a document, and then merges the two lists of recommendations. Another solution is FiGO (Finding Gene Ontology) [Couto et al. 2005], which is based on the idea that the higher the frequency of a term in a document, the lower its significance. Correspondingly, all possible variations of a GO term are found, their confidence is calculated, and those annotations are transferred to the gene whose confidence exceeds a certain threshold α. However a very low precision of 10% is obtained in experiments on the benchmark data set.

Another algorithm known as the sentence sliding window algorithm [Krallinger et al. 2005] was designed to address only Task 2.1, i.e., extraction of the most relevant passage for a given protein annotation. The strength of the algorithm was in the foundation, i.e., the set of synonyms constructed for protein names and GO labels. These synonyms, known as sub-tags, were collected both from standard databases such as LocusLink, SWISS-PROT and GOA, in addition to rule-based natural language variants. Next, a score is calculated for each sub-tag in a manually chosen abstracts, using a sentence sliding approach. Finally, an overall score is calculated for a set of sentences by multiplying the score of a protein name and a GO label, which are in turn obtained by summing the respective sub-tag scores. The highest scoring set is returned as the best passage. Owing to this context-based estimation, this submission received the highest number of correct predictions for Task 2.1.

The above descriptions show the complexity of the task of annotation from biomedical literature. More such issues were brought forth by the TREC genomics track held in 2003, which is discussed next.

9.3.2 *TREC 2003 Genomics Track.* The Text REtrieval Conference (TREC) is a well-regarded series of conferences discussion issues related to information extraction from unstructured text. Looking at the exponential growth in research towards information extraction from biomedical literature, the conference managers were motivated to establish a genomics track at the conference in 2003 [Hersh 2004], and it has been conducted ever since. The tasks in this track were similar to that of BioCreAtIvE. Specifically, the primary task of the track was to find all MEDLINE references that discuss the basic biology of a give gene X, which includes isolation, structure, genetics and function of genes/proteins in various conditions. The second task was to construct GeneRIF [Mitchell et al. 2003] entries for the given genes, which provide a certain piece of text describing the function of a gene and reference of the source of this text. Training data consisted of a file describing naming variations for 50 genes from four organisms, a large chunk of MEDLINE records, and a set of GeneRIFs corresponding to these genes for validation. Testing was performed with respect to a similar set of 50 genes and their GeneRIF entries. The track attracted several submissions and inspired many other approaches, some of which are discussed here.

One of the most successful approaches was proposed by the BioText team [Bhalotia et al. 2003]. The two tasks were targeted separately. The first module was a systematically weighted combination of judgements whether a certain document references a given gene. For making these judgements, rule-based expansions were applied to the gene names, and using a character-based n-gram model [Allen 1995] and the Dice coefficient [Rijsbergen 1979], it is estimated if the abstract, the title and the MeSH descriptor of a document contain the gene name. This estimation is combined with the probability that the document has been assigned to a GeneRIF, and the documents are finally ranked to produce results for the first task. For the second task, it is observed that GeneRIF texts generally come either from the title of a document or the last sentence of its abstract. A naive Bayes classifier was used for this classification, and the final results produced were ranked second among all the submissions to the track.

Another very successful approach, which was ranked third at the track, was presented by de Bruijn and Martin [2003]. This system consisted of seven modules, most of which were similar in spirit to corresponding components of [Bhalotia et al. 2003]'s solution. In addition, two specific ideas were also implemented, namely the use of the $tf - idf$ weighting scheme to represent the documents, and an iterative procedure to identify the documents most similar to a given text query. The overall system is found to be very successful and was ranked third among all submissions. In addition, some failure conditions were identified, such as overly complex gene names, over-representation of a certain gene in the training set, gene names close to English words and failure of the evaluation metrics.

Yet another approach inspired by the track, though not published in TREC's proceedings, appears in [Seki and Mostafa 2004]. Here, an algorithm proposed by the same authors [Seki and Mostafa 2003] was used to identify gene/protein names in the texts. The documents selected were broken up into sentences and using a probabilistic score called the g-score, the likelihood of each sentence being included in a GeneRIF record is estimated. This score is motivated by the observation that GeneRIF and non-GeneRIF sentences follow distinct g-score distributions. Though initially the precision was very low, linking of

this algorithm with LocusLink entries resulted in a better precision-recall curve for the overall system.

9.4 Discussion

The wide variety and innovativeness of solutions submitted to the BioCreAtIvE and TREC contests demonstrate the potential that literature mining holds for gene/protein function prediction in particular, and for biological knowledge discovery in general. In addition, these events also emphasized the most significant problems in this task. For instance, the maximum number of submissions to BioCreAtIvE focused on the most basic task, i.e., gene name identification in text [Blaschke et al. 2005], which indicates that there is still a long way to go before perfection is achieved in this field. As this is recognized, further research and consequently better results are expected in this field.

10. MUTIPLE DATA TYPES

10.1 Introduction

With a plethora of data being generated by a wide spectrum of proteomics experiments, it may be hypothesized that sometimes what can't be discovered from one source of information may become obvious when multiple sources are analysed simultaneously. This intuition has been concretized by Kemmeren and Holstege [2003], who have suggested the following distinct advantages achieved by integrating functional genomics data:

(1) Usually, individual biological data sets provide information about complimentary biological processes, such as gene expression and protein interaction networks. Thus, combining them provides a global picture of the biological phenomena a set of genes is involved in.

(2) Often, data quality varies between different types of data, as well as within different sources of data of the same type. For instance, studies have shown significant variations between the quality of different protein interaction data sets [Deng et al. 2003]. Thus, the combination of several data sources/types improves the quality of the overall data set, since the errors in one data set may be corrected in another.

(3) The most important advantage of the integrative approach is that since only conclusions valid over a set of data types are accepted, the predictions made by this approach are usually more confident than those made on the basis of individual data sets.

Corresponding to these prominent advantages, several computational approaches have been proposed, that address the problem of protein function prediction by an integrated analysis of a variety of data types discussed in previous sections (Sections 3-9). In addition, some approaches also utilize other forms of data such as cellular localization, protein fusion and trancription factor binding sites. However, irrespective of the types and formats of the data used, it has been generally observed in this field that the results obtained by these approaches are better than those obtained using individual sources of data, thus making them very important in the landscape of protein function prediction techniques. The following section discusses several approaches based on this data integration idea.

10.2 Existing Approaches

It was discussed in the previous section that it is beneficial to combine different forms of biological data in different ways, as may be appropriate for the particular study and

its goal. This section discusses several approaches that adopt this strategy in order to predict protein function effectively. However, for the purpose of a better understanding of these approaches, we categorize them into the following categories, on the basis of the underlying ideas in each of these approaches:

(1) **Common format**: This category consists of those approaches in which the original data types are transformed into a single format by using appropriate pre-processing strategies. This pool of data can then be analyzed in one of two ways:
   —**Same technique**: Application of the same technique to the whole data set to derive a single set of predictions.
   —**Different techniques**: Application of different techniques to different data types to derive multiple sets of predictions, from which a consensus is generated.

(2) **Independent formats**: The approaches in this category retain the original format of the contributing data sets, which implies the use of separate techniques for handling each of these data sets. However, the two sub-categories handle this analysis procedure in the following different ways:
   —**Simple Combination of results**: This may also be referred to as *post-processing* of results, where the results of each individual analyses are combined to derive a consensus set of predictions.
   —**Intelligent data fusion**: Approaches in this category adopt intelligent machine learning techniques such as Bayesian networks and suitable kernel methods for modeling the dependence between the results derived from the individual analyses.
   Owing to the possibility of applying different techniques for different data types, techniques that belong to the second category are much more popular than the ones in the first category. Another advantage is that this approach avoids the loss incurred in transforming all the data types into a common format, for which well-established approaches are not available yet. Further difference between the two will become evident during the course of their discussion.

In subsequent sections, we review several approaches proposed in each of the above categories.

10.2.1 *Approaches Using a Common Data Format.* This section covers both of the above the types of approaches which apply the same technique to this data, and those that apply a variety of techniques, and then derive a consensus of the individual results. However, a downside of this category of approaches is that methods of transforming different data types to a common type, say protein networks, are not well-established, and this introduces a source of error in the analysis. Thus, not many approaches follow this strategy.

One of the first instances of the first type of approach, i.e. same format-same technique, is detailed by Schlitt et al. [2003]. Here, three novel forms of biological data were converted into gene networks. These original data sources were as follows:

—**Mutant network**: A network of genes where an interaction means that a deletion of a gene from a mutant leads to a significant change in the expression level of the interacting gene [Rung et al. 2002].

—**In-silico network**: A network of functional associations, where a gene is a transcription factor whose binding site is found in the promoter region of the interacting gene [Pilpel et al. 2001].

—**ChIP network**: Four networks constructed from genome-wide transcription factor localization data based on ChIP experiments [Lee et al. 2002].

The hypothesis in this study was that interacting genes in the resultant combined network will be functionally related. For the validation of this hypothesis, three reference networks, separate from those used as input to the prediction algorithm, were used: protein-protein physical interaction networks, protein complexes expressed as networks and a co-citation network extracted from the biomedical literature. The comparison of the generated and reference networks was conducted by calculating the p-value for the similarity of the adjacency lists of two genes in two different networks, calculated using a hypergeometric distribution. Under the best configuration of the reference networks, a true-positive rate of 82%, and a false-positive rate of 32% was obtained. Thus, this study presents a proof of concept for the same format-same technique idea.

Strong et al. [2003] provide an instance of the approaches that follow the same format-different techniques idea. Here, the authors combined their Operon method with other methods of genome comparison, namely gene neighborhood, phylogenetic profiles and gene fusion (Section 5). The Operon method consists simply of grouping all uni-directional genes on a single DNA strand, whose internucleotide distance is less than a certain threshold, into a single operon, and thus is a simplification of the gene neighborhood method [Overbeek et al. 1999b]. This is so since the latter needs multiple genomes for making an inference, as opposed to the requirement of a single genome of the former. The more interesting part of this study was validation of the combined algorithm using the *M. tuberculosis* genome. This validation showed that the signal-to-noise ratio (SNR), calculated using the keyword recovery ratio of functionally linked (signal) and random (noise) gene pairs, of the combined methods (between 10 and 13) was significantly higher than that of any individual method (maximum of 9.5 for gene fusion).

A similar strategy has been adopted for the construction of the EFICAz [Tian et al. 2004] database. However, the proposed technique here, known as CHIEFc, is more intricate than the Operon method, and is geared towards identifying functionally discriminating residues (FDRs) in enzyme sequences, and classifying them according to the FDRs discovered. CHIEFc is a multistep method, which essentially splits a pre-specified enzyme family into subfamilies using single-linkage clustering, and then builds an all-inclusive HMM-based multiple sequence alignment (MSA) for each subfamily by including those enzymes also which are not assigned to this enzyme family, but is well aligned with this subfamily. Thus, CHIEFc also accounts for functional heterogeneity. The functionally discrimanting residues (FDRs) are the ones with the lowest Z-score based on the entropy at the corresponding position in the MSA. Finally, an unannoted enzyme is assigned to one of the enzyme classes if it is perfectly aligned and has the same FDRs in the corresponding positions. In its integrative version, namely EFICAz, the results of four different techniques are combined, which are as follows (details omitted for brevity, but can be easily extrapolated from the nature of the databases used; SIT = sequence identity):

—CHIEFc family based FDR recognition [Schomburg et al. 2004]

—CHIEFc family based SIT evaluation [Schomburg et al. 2004]

—High specificity multiple Prosite pattern recognition [Hulo et al. 2006]

—Multiple Pfam family based FDR recognition [Sonnhammer et al. 1997]

While the first method shows high accuracy on the ENZYME database [Schomburg et al.

2004] on its own, requiring a consensus of at least two techniques pumps the accuracy to 100%, while accepting results from all of them leads to a much higher coverage than any of the techniques individually. Finally, an extensive prediction on the *E. coli* genome produced 132 novel predictions, few of which could be verified from the literature. This success of EFICAz, though on the comparatively simpler benchmark of enzymes, shows the potential of applying several variations of the same approach on a common data set.

Similar to Schlitt et al. [2003], Chen and Xu [2004] present another approach in which three different data types, namely protein interaction (B), protein complexes (C) and microarray data (M), are combined to build an all-inclusive network. The reliability of an edge in this network is simply calculated as the probability of functional linkage according to atleast one of these data sources. Once, this weighted network has been constructed, two techniques from Section 8.3 are used for function prediction, one function at a time:

—**Local prediction**: Only those neighbors of the protein under consideration which have been annoted with that particular function are used for calculating the reliability score, that is in turn used to rank the final GO assignments.

—**Global prediction**: This approach is similar to that adopted by Vazquez et al. [2003], where a simple objective function measuring the number of differentially annotated edges is minimized using simulated annealing.

The testing of these two techniques was conducted on the set of 4044 GO-annotated yeast genes. A ten-fold cross-validation procedure showed that global prediction is more effective than local prediction since predictions can be made even for proteins with unannotated neighbors. In addition, the global method was able to make 1802 novel predictions, of which about half had a reliability score over 0.9 (on a scale of 0 to 1). Thus, this study showed how previously well-established techniques could be used for the effective utilization of multiple data sources.

The latest paper in this category [Kemmeren et al. 2005] describes an approach that constructs a large network of protein interactions from four types of data, namely protein-protein interactions, phenotype data, cellular localization, and mRNA expression. Function predictions are made on the basis of relationships in this network which are supplemented by a reliability value calculated from the information content of each contributing data set. Though some cases of success were reported, the overall results were not very impressive because of the simple analysis technique used.

Finally, the Protein Information Resource (PIR) [Wu et al. 2003] is one of the most popular source of various data about proteins, particularly protein sequences, via its Protein Sequence Database (PSD). A very important part of PIR is its iProClass protein functional classification system [Wu et al. 2004], which integrates standard databases of the various biological data types discussed in this survey, via database links, as well as some others such as post-translational modification and ontologies (Figure 30). The classification is rule-based [Wu et al. 2003], where the rules are based on site identification, protein name checking, keyword checking and protein classification. Wu et al. [2003] cite several cases of misannotation just on the basis of sequence identity, while Wu et al. [2004] discuss several cases of how such misannotations could be corrected and enhanced information be extracted about them through the integrated databases. Thus, PIR presents a very high-profile case in favour of integrative bioinformatics.

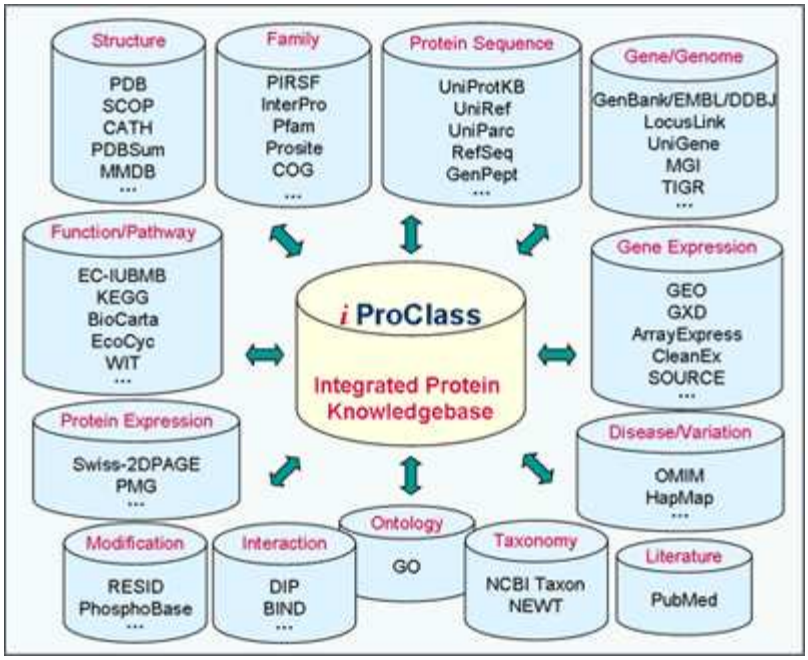


Fig. 30. Architecture of the iProClass database [Wu et al. 2003]

10.2.2 *Approaches Using Independent Data Formats.* Many more studies than those discussed in the previous section have concentrated on utilizing different forms of biological data (almost) independently, which has the advantage that different analysis techniques can be used for different types of data and the results can be combined in an appropriate way. This combination may be a simple merging of results or an intelligent modeling of the interdependencies between different sets of results derived from different sources. Both these combination strategies will be discussed in this section.

Simple Combination of Results

In a landmark paper, Marcotte et al. [1999] laid the foundations of the field of integrative functional genomics. This paper reports the first integration of the results of three independent techniques, namely phylogenetic profiles, gene fusion and correlated mRNA expression, in order to derive functionally related pairs of proteins. In yeast, upon validation using experimental interaction and pathway data, a good percentage of the links found were of the 'highest' and 'high' quality. For human genes, the evaluation was based on the SwissProt keyword recovery rate. The signal-to-noise ratio calculated according to this ratio showed that consensus links made by at least two methods were almost as reliable as those found from experimental data. Thus, this study laid a very strong foundation for this field of integrative function prediction.

The GO Engine system [Xie et al. 2002] integrates sequence homology with text data for function prediction. Here, predictions made on the basis of sequence homology searches are combined with predictions made from a simple linguistic analysis of PubMed abstracts.

This analysis simply consists of choosing the GO-gene associations with the highest log likelihood (LOD) score. The test set consisted of $670,130$ proteins taken from multiple databases, and GO was used as the source of functional classification. A cross-validation test on this data set showed almost complete coverage, while the precision was between 65% and 80%. In an additional experiment, 500 of these predictions were manually validated, and 90% accuracy was achieved. Thus, the main point made by this study was that the coverage of an annotation technique could be significantly increased, without a huge loss in precision, by combining it with other techniques based on similar ideas.

Clare and King [2003b] built upon their previous work on yeast functional classification, mainly from protein sequence data [King et al. 2001], to explore the usefulness of other forms of data for this task. The data sets used included gene expression, phenotype, sequence homology and predicted secondary structure data, the last two of which had been used in previous work [King et al. 2001]. All the data here was in a relational format, and thus their composition involved simply concatenating the respective attribute value vectors. The same approach as in [King et al. 2001] was used for functional classification. However, suitable modifications were made for ensuring the scalability, such as a more efficient frequent itemset discovery algorithm, named PolyFARM [Clare and King 2003a]. Even with these extensions, the accuracy was only around 50% for most of the data types and reached the 70% mark only for the sequence and expression data sets. However, since the classification algorithm used was rule-based, some useful rules with biological insight, such as the structure-based rule for the Mitochondrial Carrier Family (MCF) [Kuan and Jr 1993], were derived. This overall approach was also used for a functional analysis of the *Arabidopsis Thaliana* genome [Clare et al. 2006]. Here also, the overall precision was about 50%, although some interesting rules, which are either known or viable from a functional genomics viewpoint, could be derived.

ProKnow [Pal and Eisenberg 2005] is a database that integrates sequence and structure information in order to predict protein function using a probabilistic approach. Sequence information is represented as such and as motifs, while structural data is converted into folds and 3D-motifs. ProKnow also uses functional linkages from the Database of Interacting Proteins [Xenarios et al. 2002]. These features of a protein are used to derive *clues* about its function, which are used to train a naive Bayesian classifier for GO categories. Cross-validation tests on ProKnow showed about 85% precision at level 1 and 40% precision at level 9 in GO, which are not very encouraging. The reasons for this low precision could be the incompleteness of the feature set or the inappropriateness of the clues and the features used to derive them.

Finally, the JAFA server [Friedberg et al. 2006] has been set up recently to integrate the results of a variety of protein sequence-based algorithms for predicting protein function. This server collects the results of a variety of such algorithms, such as GOFigure [Khan et al. 2003], GOtcha [Martin et al. 2004], GOblet [Hennig et al. 2003], Phydbac2 [Enault et al. 2005] and InterProScan [Zbodnov and Apweiler 2001], for all the three GO ontologies, and presents them to the user in a user-friendly manner. In particular, for each GO term predicted, the output shows which algorithms agree with the prediction. Thus, JAFA is a significant effort in enhancing the utility of available tools for protein function prediction. However, the use of straightforward consensus may not be sufficient for obtaining consensus between results, particularly when used for a diverse classification scheme as GO, and may need to the use of more sophisticated consensus methods, such as semantic

similarity measures [Lord et al. 2003] for evaluating the similarity of predictions made by different algorithms.

From the above descriptions, it is evident that simple merging of results from independent sources of data and techniques has so far not yielded very good results, except in cases such as the GO Engine [Xie et al. 2002]. The reason for this is similar to the possible reasons for the failure of the naive Bayesian classifier [Duda et al. 2000]. Just as in the latter, the conditional independence assumption for the attributes may break down for real data sets, the independence assumption underlying the approaches in this section may also be invalidated since it is known that most of the biological data are complimentary. Thus, even though it is advantageous to be able to apply a different analysis technique for each type of the data, the fusion of their results or their intermediate data should be done intelligently. The next section focusses on several techniques that adopt this strategy.

Intelligent Data Fusion

As discussed above, there is a great need for techniques for the intelligent fusion of data so as to exploit the interdependencies between them in order to predict protein function confidently. Responding to this need, a large number of approaches have been proposed to carry out this fusion effectively. In fact, among those considered, this is the most populated category of integrative functional genomics approaches. Also, as will be seen as each of these approaches is discussed, data mining approaches have been used heavily because of their ability of identifying previously unknown relationships between data sets and objects [Tan et al. 2005].

Pavlidis et al. [2002] laid the foundation of this field of intelligent data fusion for function prediction through their exploratory study of the effectiveness of SVMs for this problem. In this study, mRNA expression profiles for 2465 yeast genes [Eisen et al. 1998] were integrated with their real-valued phylogenetic profiles [Pellegrini et al. 1999]. Three SVM-based kernel methods were used for this integration:

—**Early integration**: Both the attribute-value vectors were concatenated and a single SVM constructed from them for each functional class.

—**Intermediate integration**: The overall kernel function is a sum of the kernels functions calculated separately for each of the two sets of vectors. A single SVM is constructed for each class using this global kernel.

—**Late integration**: Two separate SVMs are constructed for gene expression and phylogenetic profiles respectively, and the overall discriminant value is simply the sum of the value produced by these two SVMs.

Using a cost measure which penalizes false negatives [Brown et al. 2000], it was found that the intermediate integration method gave the best results for the 108 functional classes considered. This makes sense since such an integration creates local features that are polynomial relationships between attributes within a single type of data, and the global features are formed by a linear combination of these local features. This is the strategy adopted by several feature generation techniques such as SVD and PCA [Tan et al. 2005]. Another interesting finding of this study was that it may be beneficial for the classification process if the algorithm can identify the best data set for a class before training. This systematic study thus made several useful conclusions, besides raising several important questions, to which the other approaches attempted to propose a solution.

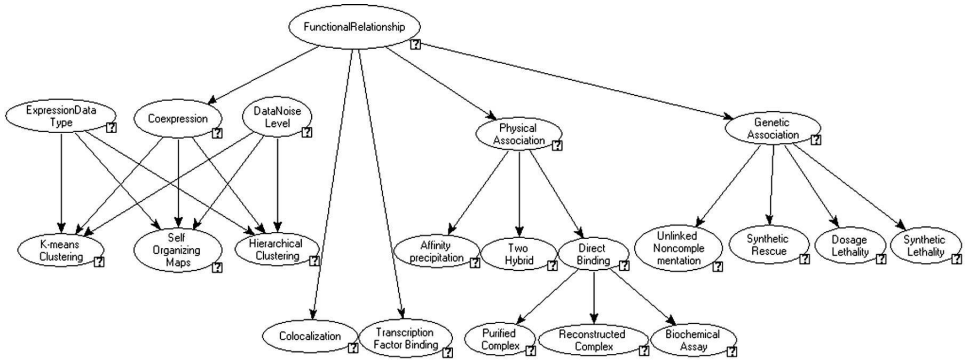


Fig. 31. General architecture of the MAGIC Bayesian network [Troyanskaya et al. 2003]

Another widely cited work in this category is the MAGIC (Multisource Association of Genes by Integration of Clusters) system [Troyanskaya et al. 2003], which employs a Bayesian belief network [Heckerman 1995] to combine the various types of data considered. These data types include clusters derived from diverse forms of data, namely gene expression, colocalization, transcription factor binding, physical associations and gene associations. The general architecture of the MAGIC Bayesian network is shown in Figure 31. The prior probabilities and the conditional probability tables of this network are provided by experts on the basis of their knowledge about the dependence of protein function on each of the data types. Thus, MAGIC combines the results of each of the intermediate analysis techniques with expert knowledge to make its predictions of functional associations between yeast genes. Indeed, using yeast genomics data and considering the hierarchical nature of GO for validation, it is shown that MAGIC outperforms all the major gene expression clustering algorithms as well as the results for using any one of the types of data as input. The effectiveness of this approach is shown by the case of the biosynthesis cluster constructed from functional associations identified by MAGIC, where 49 of the 58 known genes are annotated with protein biosynthesis, indicating a 100% recall for this cluster. Another desirable characteristic of this approach is that it can include any form of data since it takes as input gene-gene relationship matrices, where the relationship may be either binary or of some real strength.

Using a similar probabilistic approach, Huttenhower et al. [2006] proposed a scalable method for integrating several microarray datasets, in order to extract knowledge about protein function from them. For this, forty microarray datasets for *S. cerevisiae* were pre-processed, and correlations between the expression profiles of different genes were obtained from them. Next, a Bayesian network is trained for each functional class, using the correlations of gene pairs in which both genes are known to be involved in this function. Thus, a Bayesian model is obtained for predicting functional relationships between genes, based on their coexpression across a variety of microarray datasets. Indeed, this approach is able to uncover several known relationships very accurately, and also predicts other such relationships. The success of MAGIC and this approach shows the potential of Bayesian networks as a strategy for the large-scale combination of different sources of data for function prediction.

Syed and Yona [2003] report a detailed analysis of another popular classification al-

gorithm, the decision tree [Tan et al. 2005], for the problem of function prediction from multiple sources of data. However, the fusion here is in terms of individual attributes instead of complete data sets, as in other approaches. This is accomplished by learning, for each functional class, an ensemble of *probabilistic decision trees* (PDT), which are constructed by choosing a splitting attribute with a probability proportional to the information gain it provides. Thus, each tree provides a probabilistic classification for each sample, and the overall probability of assignment is calculated by weighting each probability by the performance of the corresponding tree on a separate validation data set. Furthermore, many minor yet important issues concerning decision trees were addressed carefully in this study, such as missing values, binary splitting and post-pruning using the MDL principle [Hansen and Yu 2001]. Upon evaluation, the approach was able to outperform BLAST on several enzyme families that are weakly related by homology. Also, the detailed analysis of decision trees could be useful to other function prediction approaches based on this model [Clare and King 2003b; Wang et al. 2003; Kretschmann et al. 2001].

The kernel summation idea proposed by Pavlidis et al. [2002] is heuristically generalized by Li et al. [2003] using a procedure inspired by the co-training algorithm [Blum and Mitchell 1998]. The objective of this procedure is to minimize the disagreement between classifications assigned by two classifiers. In this study, two SVMs were constructed, one each from the gene expression and phylogenetic profile data sets used by Pavlidis et al. [2002]. Then, an iterative randomized procedure was used to reduce the disagreement between the two sets of labels generated for both labeled and unlabeled examples by randomly modifying the labels in case of a mismatch. Though there is no explicit comparison in the evaluation against any other approach, the strongest advantage of this approach is that it is able to use both labeled and unlabeled examples for training. This is very useful for function prediction since even in the most well studied genomes, a significant fraction of the proteins are still unannotated.

Deng et al. [2004] extended their earlier work on probabilistic function inference from protein interaction networks [Deng et al. 2003] in order to incorporate other types of networks and features for this task. Though the same Markov random field (MRF) model is still used for propagating functional annotation throughout the networks, three specific modifications are made to the approach in [Deng et al. 2003]:

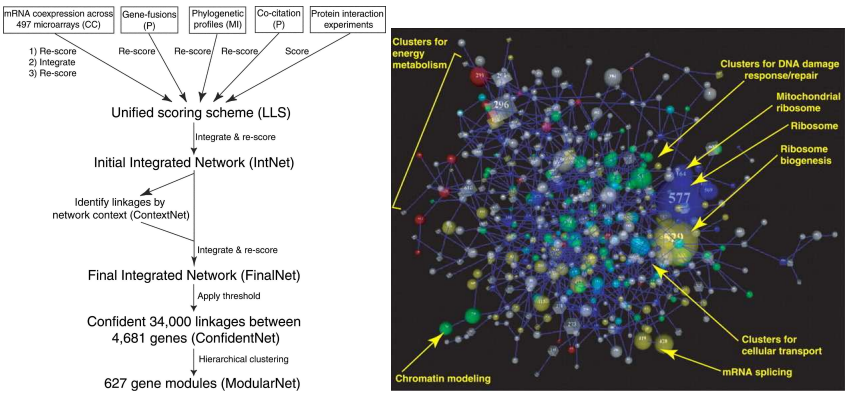
—The prior probabilities of functional classes are now computed from the protein complex data [Gavin et al. 2002] and not from frequencies of annotation.

—Three types of networks are used: physical interactions, protein complexes and mRNA coexpression. Separate MRF models are constructed for these networks.

—Pfam domains [Sonnhammer et al. 1997] are also used as protein features and probabilities conditional on them are also incorporate into the final formula for probability.

The same Gibbs sampling procedure as in [Deng et al. 2003] is still used for estimating the annotation probabilities. Indeed, the results on yeast genes improved with increasing amount of information, with the precision-recall equality point appearing at about 76%.

Another approach which integrates as many as eight different categories of data into a global network and then analyzes them is presented by Lee et al. [2004]. This approach is focused on deriving functionally coherent clusters of genes from this network using the algorithm shown in Figure 32(a). Here, the edges in the original networks constructed from individual data sets are weighted according to a log-likelihood score, and these edges are



(a) Integration algorithm (b) Strongest gene (sub)network generated

Fig. 32. Probabilistic functional network of yeast genes [Lee et al. 2004]: Algorithm and results

merged into a global network. Further incorporation of context information produced more compact versions of the network called *ContextNet* and *FinalNet*. Figure 32(b) shows the strong edges in the FinalNet generated from yeast genomics data. Functionally coherent clusters of genes, such as DNA damage response/repair, energy metabolism and mRNA splicing, can clearly be identified in this network.

[Lee et al. 2004] and other studies showed the potential of using networks of genes or proteins as the basic representation for integrating different types of biological data for functional inference. An effective approach in this direction is to supplement the edges and/or nodes in a given physical interaction network with other information about the constituent proteins, such as their amino acid sequences and expression profiles. This perspective was first adopted for the development of the PathBLAST algorithm [Kelley et al. 2003], which performed a BLAST-like pairwise alignment of input interaction networks using sequence similarity information. Here, a new graph is constructed from the two input interaction networks using the following transformation:

—A new vertex, say A/a, is created in the new graph if two proteins in the two networks, say A and a respectively, have a significant sequence similarity.

—An edge is created between two nodes A/a and B/b in the transformed graph if the nodes (A, B) and (a, b) are connected by a path of length atmost 2 in their original networks, and the edge is labeled as follows:

  —*Direct* if (A, B) and (a, b) are interacting nodes in their original networks.

  —*Gap* if (A, B) is a direct interaction and (a, b) are connected by a path of length 2 in their original networks, or vice versa.

  —*Mismath* if both (A, B) and (a, b) are connected by a path of length 2 in their original networks.

This transformation prepares the input networks in a format similar to sequence alignment used by BLAST. Next, a scoring function is defined for a path, which is computed as a combination of the likelihoods of observing the sequence similarities in nodes, and the

constituent edges, as compared to randomly generated data. Finally, interesting paths are derived from this resultant network using a dynamic programming algorithm for finding high scoring paths. The final pathways are constructed by combining paths that overlap or a separated from each other by just one interaction in the original networks. In an evaluation study comparing the networks of *S. cerevisiae* and the bacteria *H. pylori*, several interesting pathways, such as *protein synthesis and cell rescue, cytoplasmic and nuclear membrane transport* and *protein degradation and DNA repair*, observed to be substantially conserved in the two specieis. This is very interesting from a function prediction point of view, since interesting groups of proteins, such as protein complexes, in a less studied organism (such as *H. pylori*) can be aligned against the network of a well-studied organism (such as *S. cerevisiae*) to identify well represented pathways. Indeed, in an extension of this algorithm to incorporate networks of more than two organisms [Sharan et al. 2006], the conserved groups of proteins were found to be highly enriched by several GO biological process terms. Thus, PathBLAST and its extensions provide an efficient comparative method of identifying evolutionarily conserved pathways betweend different organisms, and the functions of the constituent proteins.

Another approach for the augmentation of interaction networks for function prediction is to use an organism's interaction network and weight the edges to reflect the biological viability of the corresponding interactions. The co-expression of interacting genes, measured as the correlation of their expression profiles across several microarray datasets, provides a useful measure for this viability [Kemmeren and Holstege 2003]. This has motivated the development of function prediction methods that use interaction networks weighted according to microarray data. VIRGO [Massjouni et al. 2006] is a useful webserver that adopts this approach. The GAIN algorithm [Karaoz et al. 2004] is used as the underlying technique for extracting functional information from the stored interaction datasets of *S. cerevisiae* and *H. sapiens*. However, the basic purpose of VIRGO is to provide a service for biologists to be able to derive functional annotations for the protein(s) of their interest, and is not a novel technique for function prediction by itself.

Nariai et al. [2007] presented a more rigorous integration approach, wherein they integrated five different data types, namely protein interaction data, gene expression data, protein motif information, gene knock-out phenotype data and protein localization data. The basic input to the prediction algorithm consists of two network, one being the interaction network, and the second being a co-expression network containing edges between proteins that had a correlation of over 0.85 in their expression profiles. Also, the other data types were represented as categorical variables, where for each type, a vector of all possible values is created, and for each protein, a 1 is recorded for the corresponding entry if the protein is known to have that value for that type. With this mathematical representation of the data, a naive Bayesian classification model [Tan et al. 2005] that predicts the probability of a protein having a certain function, conditional on the number of its neighbors annotated with this function in the two networks, and its feature values obtained from the other data types. Using a large set of functional annotations from the GO biological process ontology for several yeast proteins, an evaluation of the predictions generated by the integrated model was performed, in comparison to those generated by the protein interaction data alone. It was shown that if recall is the primary requirement (precision=50%), there is a huge improvement in the number of true positive predictions. However, interestingly, if a high precision of predictions is required (precision=50%), the number of true

positive predictions is not significantly different from those produced by the interaction data alone, and this number increases as more and more interaction data is included in the integrated model. Thus, in addition to providing a robust method for integrating diverse types of data for function prediciton, this study showed that the use of protein interaction data is almost indispensable for obtained precise predictions from an integrated model of protein function.

Ulitsky and Shamir [2007] considered a different version of the problem of integrating interaction data with microarray data. Here, microarray data was treated as the basic data for discovering functional modules, while the input interaction network was used as a constraint graph, which ensured that only connected modules were discovered. The problem considered was one of finding modules showing strong coexpression between the constituent proteins, while maintaining connectivity as per the input interaction network. This problem is addressed by developing a probabilistic mixture model [Duda et al. 2000], which comprises of two Gaussian distributions, one for genes expected to be co-expressed, and one for those that are not. The parameters of this model are estimated using an Expectation-Maximization procedure [Duda et al. 2000] at the begining of the algorithm. Using this model, a likelihood score is defined for each module, indicating whether this module is expected to consist of co-expressed genes, or not. Now, using this score, a greedy clustering-type algorithm is executed, which starts with a seed cluster of proteins and keeps growing it until termination. The two constraints that are satisfied at each step are that the current cluster should be connected as per the topology of the interaction network, and the likelihood score of the cluster should increase at each step. The final cluster is termed as a functional module, and is evaluated for enrichment with different GO categories. Indeed, using a cell cycle-based gene expression data set for human proteins, and a large scale set of interactions between human proteins, the algorithm is able to precisely uncover a significantly large module containing several proteins known to be involved in the cell cycle operation, thus validating the algorithm's efficacy at finding biologically relevant modules. Also, significant improvements are observed over a competing clustering algorithm for microarray data, which are evidently due to the use of interaction networks as a constraint graph. This again validates the utility of protein interaction data for integrative genomics projects.

It was noted earlier in Section 8.2 that the interest in data mining for function prediction had inspired the inclusion of this problem in the KDD Cup 2001 [Cheng et al. 2002], which is an annual contest held in conjunction with the SIGKDD conference[19]. The precise tasks assigned in this version of the contest were the predicition of protein function and subcellular location from protein interaction and other forms of proteomics data provided in a relational format, with some of the proteins having being annotated by the organizers. The winning solution [Cheng et al. 2002] for the function prediction problem used the RELAGGS algorithm [Krogel and Wrobel 2001] to convert the multi-relational data into an attribute vector format, from which an SVM classification model was learnt and used for prediction on the test set. In an extension of this solution, the winning team attempted to utilize the unlabeled data in order to improve the precision of the classifier [Krogel and Scheffer 2004], as also attempted earlier by Li et al. [2003]. However, it was surprisingly found that the popular transductive SVM learning [Joachims 1999] and co-training [Blum and Mitchell 1998] algorithms failed to improve the classifier's performance for this task. It

[19] http://www.acm.org/sigs/sigkdd/

was found that the reason for failure was the dependence between any two sets of attributes extracted from the original data. Thus, it was established that if two independent sets of functional genomics data can be identified, then a co-training based approach [Blum and Mitchell 1998] could be used to design an effective classification scheme that uses unlabeled data as well.

Previous studies [Pavlidis et al. 2002; Li et al. 2003] have shown the utility of deriving separate kernel functions for different types of biological data and combining them in different ways to achieve effective genomic data fusion. In their work, Pavlidis et al. [2002] used the simple summation operation to derive the overall kernel function. This work is generalized by [Lanckriet et al. 2004], both in terms of the types of data used, the individual kernels and the combination procedure. Table IX lists the types of data used and the novel kernels used to model similarities between proteins on their basis. The motivation of using these kernels is their success in learning with the corresponding type of data. Next, these kernels are optimally weighted, to derive a global kernel function for the fused data. The results obtained with this kernel were better than those obtained from the MRF-based approach of Deng et al. [2004] on the data set used by the latter for evaluation. In a more focused application of this approach [Lanckriet et al. 2004], ribosomal and membrane proteins in *S. cerevisiae* were more accurately classified than by using each kernel individually.

Type of data	Kernels used
Protein sequences	Smith-Waterman, BLAST and Pfam HMM-based
Protein interactions	Linear and diffusion kernels [Vert and Kanehisa 2002]
Gene expression	Radial basis kernel

Table IX. Genomic data types and corresponding kernels used in [Lanckriet et al. 2004]

A similar kernel combination approach is proposed in [Borgwardt et al. 2005]. Here, a unified graph is constructed from protein sequence and structure data, with the nodes and edges being labeled by the type of data they are derived from, and the corresponding continuous attributes that can best capture the data at that node. The popular random walk kernel for graphs [Gärtner et al. 2003] is generalized by decomposing it into individual kernels for the nodes, edges and node attributes, and the hyperkernel approach [Ong et al. 2005] is used to derive an optimal kernel for the data with respect to a specified cost function. Using this approach, only the most useful node attributes are used for classification, reducing the information collection effort. This claim was validated through evaluation of the approach on the enzyme+non-enzyme data set used in [Dobson and Doig 2003], and experiments focussed on the enzyme/non-enzyme class showed that the modified graph kernel was indeed able to outperform the simple attribute vector-based SVM classification used in [Dobson and Doig 2003]. In addition, a high accuracy of 90.83% on average was achieved for the Enzyme Classification top-level class prediction problem on a data set of 600 enzymes from the BRENDA database [Schomburg et al. 2004].

The latest work in the direction of graph-based data fusion [Tsuda et al. 2005] combines four different types of biological networks, by assigning a weight to each of them. Each graph is represented by a cost equation, which is a quadratic functional representing the *smoothness* and *consistency* conditions. The first condition implies that the function assignment score should not be too different between adjacent vertices, as defined at the

attribute level by Vert and Kanehisa [2002], while the second enforces consistency between the labels of the training set and those assigned by the algorithm. This way, the approach is able to use data both from the labeled examples, i.e., functional labels, and unlabeled examples, i.e., the connectivity structure of the combined graphs. Finally, the overall cost function is obtained as a linear combination of those of each of the constituent graphs, and is minimized using an EM-like procedure [Duda et al. 2000]. The approach showed a significant improvement over using the technique with any of the networks individually. Thus, this approach showed an efficient way of incorporating unlabeled data into the learning process to improve classification performance.

Looking back, the approaches described in this section were able to achieve an improvement not only in terms of coverage as reported for the approaches in the previous sections, but also in precision and accuracy. This improvement is primarily because of the fact that the approaches in this section allow the knowledge derived from one source, such as gene expression data, to compliment knowledge from other sources such as phylogenetic profiles, thus allowing the flow of information from one source to another. With improvement in data fusion algorithms [Hall 2001], much better results are expected from this category of approaches.

10.3 Discussion

As was discussed earlier, this fusion of data gives us advantages such as more reliable predictions and less disruption of the results due to the low quality of individual sources of data. Now, after a detailed discussion of the numerous approaches which have been proposed in this direction, it can be said confidently that nearly all of them were able to achieve their target, though different approaches in different categories achieve this to different extents. Several other conclusions can be drawn from the descriptions of approaches above, a consideration of which could lead to better results:

(1) The combination of multiple sources of data in a manner that exploits the dependencies between individual data types is more effective for function prediction than the simple merging of data types or the inferences drawn individually from them. In particular, Bayesian networks and kernel-based techniques have gained popularity for this task, because of these reasons:
   —Bayesian networks [Heckerman 1995] have a graph structure which makes it possible to model various types of dependencies such as hierarchies and independence. Also, these networks allow the flow of information between nodes, which is important for fusion-based prediction.
   —Kernel methods [Shawe-Taylor and Cristianini 2004] capture the similarity between two proteins on the basis of the evidence provided. Thus, in the domain of multiple data sources, this can be extended to the construction of separate similarity matrices on the basis of the different data types, and their consequent merger to derive an overall similarity matrix. This procedure can be seen as first capturing the local patterns in individual data sets and their subsequent generalization to derive global patterns, which can be used for function prediction.

   Of late, there has been a greater interest in kernel-based methods [Lanckriet et al. 2004; Borgwardt et al. 2005] because they allow the use of any classifier based on the similarity between examples, though some approaches have tried to combine the two frameworks for more accurate prediction of protein function [Barutcuoglu et al. 2006].

(2) By far, protein networks have been established as a unifying framework for all forms of functional genomics data, as was advocated by [Fraser and Marcotte 2004]. Some recent approaches have focussed on fusion methods based on protein networks [Borgwardt et al. 2005; Deng et al. 2004; Tsuda et al. 2005], while others have used them simply as a visual tool for diverse forms of data [Lee et al. 2004]. Irrespective of the mode of their use, networks can be used effectively for representing functional genomics data and combining them if appropriate.

(3) Several recent studies [Nariai et al. 2007; Ulitsky and Shamir 2007] have shown that the use of protein interaction data within an integration model leads to very significant improvement in the precision of predictions made from a model that aims to integrate different genomic data sets for function prediction. Thus, wherever possible, such data should be considered in studies on integration of diverse genomics data.

(4) Many of the approaches discussed in this section were simple extensions or generalizations of studies that have been conducted by authors on individual data types, such as [Marcotte et al. 1999; Deng et al. 2004; Clare and King 2003b; Krogel and Scheffer 2004]. This indicates that the fusion of multiple sources of data does not require a completely new methodology, but can be achieved by simple extensions of approaches focussing on a single form of data.

The conclusions, in combination with the progress being made in machine learning-based data fusion techniques will ensure that more accurate and extensive results are obtained from approaches targeting a fusion of genomics data in order to make predicitions about protein function.

11. CONCLUSIONS

In the previous sections, numerous approaches for the computational prediction of protein function from various types of biomedical data were discussed. Even though this is an immensely diverse field, as can be seen from the wide spectrum of data types, as well as algorithms covered in this survey, the following general conclusions can be made:

(1) In most categories, the best results were obtained from approaches that employed techniques from the fields of **data mining** and **machine learning**. Nonetheless, the analysis of biological data involves handling a number of challenges, many of which have only been partly addressed. Some of the most prominent challenges are as follows:
—Identification of the most relevant subset of the data for the functional classes being addressed.
—Possibility of a protein performing multiple functions, and thus having multiple functional labels.
—Widely varying sizes of functional classes, most classes being very small.
—Hierarchical arrangement of functional labels, as in Gene Ontology.
—Incompleteness and various types and extents of noise in biological data.
—High dimensionality of the data.
Advances in data mining to address these challenges are necessary to exploit the available biological data to predict protein function more accurately and effectively.

(2) Several approaches covered in this survey obtained very good results on a wide variety of functional classes. However, several other approaches that did not achieve comparable results still deserve discussion because of several reasons, such as the following:

—Their results may improve when more data is available for training their models.
—Improvements made to the underlying technique may lead to better results.
—The approach may have a specific area of application, for instance classification of proteins belonging to the G-protein coupled receptors superfamily [Qian et al. 2003; Cheng et al. 2005].
—A combination of these approaches with more powerful one may lead to better results than those obtained by any one of them individually, as illustrated in Section 10.

(3) The incorporation of **domain knowledge** is the most promising approach to make the algorithms for function prediction biologically robust, as shown by several successful studies based on this idea [Jensen et al. 2002; Barutcuoglu et al. 2006]. Such knowledge may come in the form of mechanisms underlying the accomplishment of protein functions, such as post-translational modifications [Mann and Jensen 2003], or a hierarchy of functional classes, such as the Gene Ontology [Ashburner et al. 2000], or for that matter, any relevant experimentally or computationally determined knowledge. A particularly effective method of integrating domain knowledge for protein function prediction is the intelligent fusion of diverse types of genomic data (Chapter 10), which has enabled the enhancement of the precision as well as coverage of the predictions, as compared to those produced by any single data type.

(4) Each **type of biological data** usually has a **strong correspondence** with a certain **type of function** that can be best predicted using data sets of that type. For instance, individual protein-level data, such as protein structure, are best capable of finding molecular function [Laskowski et al. 2003], while genome-level data, such as protein interaction networks and gene expression data, offer good insights into the biological process a protein participates in [Nabieva et al. 2005]. Further still, phylogenetic profiles have been shown to be appropriate for the task of reconstructing preserved working units of proteins, such as metabolic pathways [Pellegrini et al. 1999]. Thus, a knowledge of the nature of the available biological data may aid the identification of the form of protein function that can be predicted from it, and vice versa.

(5) The **Gene Ontology** is increasingly being established as the most appropriate functional classification scheme for protein function prediction research because of its several desirable properties (Section 2.2), and the forward looking attitude of its curators who are keeping it up-to-date with latest research. In particular, several GO-friendly approaches have recently been proposed, which incorporate the hierarchical structure of GO in the prediction technique so as to exploit the parent-child relationships between various functional classes [Engelhardt et al. 2005; Barutcuoglu et al. 2006; Liu et al. 2004; Eisner et al. 2005]. The above discussion of the correspondence between data and function type also lends support to the utility of GO. This is so since GO contains separate ontologies for three different types of protein function, namely *cellular component*, *molecular function* and *biological process*, thus making it easier to identify the most appropriate functional hierarchy to be used for making predictions from biological data of a certain type.

(6) Even though many advances have been made in the field of protein function prediction, there is still a lack of understanding of the most appropriate prediction technique for any particular category of proteins. Some attempts have been made to perform an evaluation of the current available prediction methods, such as the function prediction

component at CASP6 [Pellegrini-Calace et al. 2006], CASP7 [Lpez et al. 2007] and the Automated Function Prediction meeting in 2005 [Godzik et al. 2007]. Although a great deal was learnt from these evaluations, they were conducted on a small selected set of target proteins, which may not reflect the generalization ability of a certain function prediction technique. Thus, there is a great need for the creation of **benchmark datasets** and the adoption of a consistent **evaluation methodology**, as has occurred in the field of remote homology prediction [Rangwala and Karypis 2005; Kuang et al. 2005]. This standardization will help in the identification of both the most appropriate function predictions strategy in a certain context, and the current weaknesses and needs of the field. Datasets such as [Tetko et al. 2005] and rigorous evaluations methodologies such as those adopted by Nabieva et al. [2005] are positive steps in this direction.

Last but not the least, we firmly believe that an efficient scientific workflow can be established, in which, first, hypotheses are generated by executing the appropriate function prediction prediction algorithm on the available biological data, and then, these hypotheses are validated experimentally, thus leading to confident predictions of a protein's function. Table I, presented earlier, lists several examples where this approach has produced both useful and valid results. We hope that this survey aids both computational and experimental practitioners in molecular biology in accomplishing this task more effectively.

REFERENCES

ABASCAL, F. AND VALENCIA, A. 2003. Automatic annotation of protein function based on family identification. *Proteins 53*, 3, 683–692.

ADACHI, N. AND LIEBER, M. R. 2002. Bidirectional gene organization: a common architectural feature of the human genome. *Cell 109*, 7, 807–809.

AL-SHAHIB, A., BREITLING, R., AND GILBERT, D. 2005. Franksum: new feature selection method for protein function prediction. *Int. J. Neural Syst. 15*, 4, 259–275.

ALBERTS, B., JOHNSON, A., LEWIS, J., RAFF, M., BRAY, D., HOPKIN, K., ROBERTS, K., AND WALTER, P. 2003. *Essential Cell Biology*. Garland Science.

ALFARANO, C., ANDRADE, C. E., ANTHONY, K., BAHROOS, N., BAJEC, M., BANTOFT, K., BETEL, D., BOBECHKO, B., BOUTILIER, K., BURGESS, E., BUZADZIJA, K., CAVERO, R., D'ABREO, C., DONALDSON, I., DORAIRAJOO, D., DUMONTIER, M. J., DUMONTIER, M. R., EARLES, V., FARRALL, R., FELDMAN, H., GARDERMAN, E., GONG, Y., GONZAGA, R., GRYTSAN, V., GRYZ, E., GU, V., HALDORSEN, E., HALUPA, A., HAW, R., HRVOJIC, A., HURRELL, L., ISSERLIN, R., JACK, F., JUMA, F., KHAN, A., KON, T., KONOPINSKY, S., LE, V., LEE, E., LING, S., MAGIDIN, M., MONIAKIS, J., MONTOJO, J., MOORE, S., MUSKAT, B., NG, I., PARAISO, J. P., PARKER, B., PINTILIE, G., PIRONE, R., SALAMA, J. J., SGRO, S., SHAN, T., SHU, Y., SIEW, J., SKINNER, D., SNYDER, K., STASIUK, R., STRUMPF, D., TUEKAM, B., TAO, S., WANG, Z., WHITE, M., WILLIS, R., WOLTING, C., WONG, S., WRONG, A., XIN, C., YAO, R., YATES, B., ZHANG, S., ZHENG, K., PAWSON, T., OUELLETTE, B. F. F., AND HOGUE, C. W. V. 2005. The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Research 33*, Database issue, D418–D424.

ALLEN, J. 1995. *Natural language understanding (2nd ed.)*. Benjamin-Cummings Publishing Co. and Inc.

ALTAF-UL-AMIN, M., NISHIKATA, K., KOMA, T., MIYASATO, T., SHINBO, Y., ARIFUZZAMAN, M., WADA, C., MAEDA, M., OSHIMA, T., MORI, H., AND KANAYA, S. 2003. Prediction of protein functions based on k-cores of protein-protein interaction networks and amino acid sequences. *Genome Informatics 14*, 498–499.

ALTMAN, D. G. 1997. *Practical Statistics for Medical Research*. Chapman and Hall.

ALTSCHUL, S. F., GISH, W., MILLER, W., MEYERS, E. W., AND LIPMAN, D. J. 1990. Basic local alignment search tool. *J Mol Biol. 215*, 3, 403–410.

ALTSCHUL, S. F., MADDEN, T. L., SCHFFER, A. A., ZHANG, J., ZHANG, Z., MILLER, W., AND LIPMAN, D. J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research 25*, 17, 3389–3402.

AMARI, S.-I. 1995. Information geometry of the EM and em algorithms for neural networks. *Neural Networks 8,* 9, 1379–1408.

ANDRADE, M. A., BROWN, N. P., LEROY, C., HOERSCH, S., DE DARUVAR, A., REICH, C., FRANCHINI, A., TAMAMES, J., VALENCIA, A., OUZOUNIS, C., AND SANDER, C. 1999. Automated genome sequence analysis and annotation. *Bioinformatics 15,* 5, 391–412.

ANDRADE, M. A. AND VALENCIA, A. 1997. Automatic annotation for biological sequences by extraction of keywords from MEDLINE abstracts: Development of a prototype system. In *Proc. ISMB*. 25–32.

ANDRADE, M. A. AND VALENCIA, A. 1998. Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families. *Bioinformatics 14,* 7, 600–607.

ANDREEVA, A., HOWORTH, D., BRENNER, S. E., HUBBARD, T. J. P., CHOTHIA, C., AND MURZIN, A. G. 2004. SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Research 32,* Database issue, D226–D229.

APWEILER, R., ATTWOOD, T. K., BAIROCH, A., BATEMAN, A., BIRNEY, E., BISWAS, M., BUCHER, P., CERUTTI, L., CORPET, F., CRONING, M. D. R., DURBIN, R., FALQUET, L., FLEISCHMANN, W., GOUZY, J., HERMJAKOB, H., HULO, N., JONASSEN, I., KAHN, D., KANAPIN, A., KARAVIDOPOULOU, Y., LOPEZ, R., MARX, B., MULDER, N. J., OINN, T. M., PAGNI, M., SERVANT, F., SIGRIST, C. J. A., AND ZDOBNOV, E. M. 2000. InterPro: an integrated documentation resource for protein families and domains and functional sites. *Bioinformatics 16,* 12, 1145–1150.

ASHBURNER, M., BALL, C. A., BLAKE, J. A., BOTSTEIN, D., BUTLER, H., CHERRY, J. M., DAVIS, A. P., DOLINSKI, K., DWIGHT, S. S., EPPIG, J. T., HARRIS, M. A., HILL, D. P., ISSEL-TARVER, L., KASARSKIS, A., LEWIS, S., MATESE, J. C., RICHARDSON, J. E., RINGWALD, M., RUBIN, G. M., AND SHERLOCK, G. 2000. Gene Ontology: tool for the unification of biology. *Nature Genetics 25,* 1, 25–29.

ATTWOOD, T. K., BRADLEY, P., FLOWER, D. R., GAULTON, A., MAUDLING, N., MITCHELL, A. L., MOULTON, G., NORDLE, A., PAINE, K., TAYLOR, P., UDDIN, A., AND ZYGOURI, C. 2003. PRINTS and its automatic supplement and prePRINTS. *Nucleic Acids Research 31,* 1, 400–402.

BAADER, F., CALVANESE, D., MCGUINNESS, D. L., NARDI, D., AND PATEL-SCHNEIDER, P. F., Eds. 2003. *The description logic handbook: theory and implementation and and applications*. Cambridge University Press.

BADA, M., STEVENS, R., GOBLE, C. A., GIL, Y., ASHBURNER, M., BLAKE, J. A., CHERRY, J. M., HARRIS, M. A., AND LEWIS, S. 2004. A short study on the success of the Gene Ontology. *J. Web Sem. 1,* 2, 235–240.

BADEA, L. 2003. Functional discrimination of gene expression patterns in terms of the gene ontology. In *Proc. Pacific Symposium on Biocomputing*. 565–576.

BAHL, A., BRUNK, B., CRABTREE, J., FRAUNHOLZ, M. J., GAJRIA, B., GRANT, G. R., GINSBURG, H., GUPTA, D., KISSINGER, J. C., LABO, P., LI, L., MAILMAN, M. D., MILGRAM, A. J., PEARSON, D. S., ROOS, D. S., SCHUG, J., STOECKERT, C. J., JR, , AND WHETZEL, P. 2003. PlasmoDB: the Plasmodium genome resource. A database integrating experimental and computational data. *Nucleic Acids Research 31,* 1, 212–215.

BAILEY, T. L., BAKER, M. E., ELKAN, C. P., AND GRUNDY, W. N. 1999. MEME and MAST and MetaMEME: New tools for motif discovery in protein sequences. In *Pattern Discovery in Biomolecular Data*, J. T. L. Wang, B. A. Shapiro, and D. Shasha, Eds. Oxford University Press, 30–54.

BAIROCH, A. 2000. The ENZYME database in 2000. *Nucleic Acids Research 28,* 1, 304–305.

BALDAUF, S. L. 2003. Phylogeny for the faint of heart: a tutorial. *Trends in Genetics 19,* 6, 347–351.

BANDYOPADHYAY, D., HUAN, J., LIU, J., PRINS, J., SNOEYINK, J., WANG, W., AND TROPSHA, A. 2006. Structure-based function inference using protein family-specific fingerprints. *Protein Science 15*, 1537–1543.

BANERJEE, A., KRUMPELMAN, C., GHOSH, J., BASU, S., AND MOONEY, R. J. 2005. Model-based overlapping clustering. In *Proc Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. 532–537.

BAR-JOSEPH, Z. 2004. Analyzing time series gene expression data. *Bioinformatics 20,* 16, 2493–2503.

BARRETT, T., SUZEK, T. O., TROUP, D. B., WILHITE, S. E., NGAU, W.-C., LEDOUX, P., RUDNEV, D., LASH, A. E., FUJIBUCHI, W., AND EDGAR, R. 2005. NCBI GEO: mining millions of expression profiles–database and tools. *Nucleic Acids Research 33,* Database issue, D562–D566.

BARTLETT, G. J., TODD, A. E., AND THORNTON, J. M. 2003. Inferring protein function from structure. In *Structural Bioinformatics*, P. E. Bourne and H. Weissig, Eds. Wiley-Liss and Inc., Chapter 19, 387–407.

BARUTCUOGLU, Z., SCHAPIRE, R. E., AND TROYANSKAYA, O. G. 2006. Hierarchical multi-label prediction of gene function. *Bioinformatics 22,* 7, 830–836.

BARZILAY, R. AND ELHADAD, N. 2003. Sentence alignment for monolingual comparable corpora. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*. 25–32.

BELLAACHIA, A., PORTNOY, D., CHEN, Y., AND ELKAHLOUN, A. G. 2002. E-CAST: A data mining algorithm for gene expression data. In *Proc. 2nd Workshop on Data Mining in Bioinformatics (BIOKDD)*. 49–54.

BEN-DOR, A., SHAMIR, R., AND YAKHINI, Z. 1999. Clustering gene expression patterns. *J. Computational Biology 6*, 3-4, 281–297.

BEN-HUR, A. AND BRUTLAG, D. 2003. Remote homology detection: a motif based approach. *Bioinformatics 19*, Suppl 1, i26–i33.

BEN-HUR, A. AND BRUTLAG, D. 2005. Sequence motifs: Highly predictive features of protein function. In *Feature extraction and foundations and applications*, I. Guyon, S. Gunn, M. Nikravesh, and L. Zadeh, Eds. Springer Verlag.

BENSON, D. A., KARSCH-MIZRACHI, I., LIPMAN, D. J., OSTELL, J., AND WHEELER, D. L. 2004. Genbank: update. *Nucleic Acids Research 32,* Database issue, D23–D26.

BERMAN, H. M., WESTBROOK, J., FENG, Z., GILLILAND, G., BHAT, T. N., WEISSIG, H., SHINDYALOV, I. N., AND BOURNE, P. E. 2000. The Protein Data Bank. *Nucleic Acids Research 28,* 1, 235–242.

BHALOTIA, G., NAKOV, P., SCHWARTZ, A., AND HEARST, M. 2003. BioText team report for the TREC 2003 genomics track. In *Proc. Text Retrieval Conference (TREC)*. 93–97.

BILU, Y. AND LINIAL, M. 2002. Functional consequences in metabolic pathways from phylogenetic profiles. In *Proc. International Workshop on Algorithms in Bioinformatics (WABI)*. 263–276.

BINKOWSKI, T. A., ADAMIAN, L., AND LIANG, J. 2003. Inferring functional relationships of proteins from local sequence and spatial surface patterns. *J Mol Biol. 332,* 2, 505–526.

BITTAR, G. J. AND SONDEREGGER, B. 2004. *An Introduction to Phylogenetics and its Molecular Aspects*. EMBER Consortium.

BLAKE, J. A., RICHARDSON, J. E., BULT, C. J., KADIN, J. A., EPPIG, J. T., AND GROUP, M. G. D. 2003. MGD: the Mouse Genome Database. *Nucleic Acids Research 31,* 1, 193–195.

BLASCHKE, C., LEON, E. A., KRALLINGER, M., AND VALENCIA, A. 2005. Evaluation of biocreative assessment of task 2. *BMC Bioinformatics 6*, Suppl 1, S16.

BLASCHKE, C. AND VALENCIA, A. 2003. Automatic classification of protein functions from the literature. *Comp. Func. Genomics 4*, 75–79.

BLEKAS, K., FOTIADIS, D. I., AND LIKAS, A. 2005. Motif-based protein sequence classification using neural networks. *J Comput Biol. 12,* 1, 64–82.

BLUM, A. AND MITCHELL, T. 1998. Combining labeled and unlabeled data with co-training. In *Proc. Eleventh Annual Conference on Computational Learning Theory (COLT)*. 92–100.

BOBIK, T. A. AND RASCHE, M. E. 2000. Identification of the human methylmalonyl-coa racemase gene based on the analysis of prokaryotic gene arrangements. implications for decoding the human genome. *J Biol Chem. 276*, 40, 37194–37198.

BOECKMANN, B., BAIROCH, A., APWEILER, R., BLATTER, M.-C., ESTREICHER, A., GASTEIGER, E., MARTIN, M. J., MICHOUD, K., O'DONOVAN, C., PHAN, I., PILBOUT, S., AND SCHNEIDER, M. 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research 31,* 1, 365–370.

BORGES, J. L. 1964. *The analytical language of John Wilkins*. Other Inquisitions. University of Texas Press.

BORGWARDT, K. M., ONG, C. S., SCHNAUER, S., VISHWANATHAN, S. V. N., SMOLA, A. J., AND KRIEGEL, H.-P. 2005. Protein function prediction via graph kernels. *Bioinformatics 21,* Suppl 1, i47–i56.

BORK, P., DANDEKAR, T., DIAZ-LAZCOZ, Y., EISENHABER, F., HUYNEN, M., AND YUAN, Y. 1998. Predicting function: from genes to genomes and back. *J Mol Biol. 283*, 4, 707–725.

BORK, P. AND KOONIN, E. V. 1996. Protein sequence motifs. *Curr Opin Struct Biol. 6*, 3, 366–376.

BOURNE, P. E. 2003. CASP and CAFASP experiments and their findings. In *Structural Bioinformatics*, P. E. Bourne and H. Weissig, Eds. Wiley-Liss and Inc., Chapter 24, 499–505.

BOURNE, P. E. AND WEISSIG, H., Eds. 2003. *Structural Bioinformatics (Methods of Biochemical Analysis, V. 44)*. Wiley-Liss.

BOWERS, P. M., PELLEGRINI, M., THOMPSON, M. J., FIERRO, J., YEATES, T. O., AND EISENBERG, D. 2004. Prolinks: a database of protein functional linkages derived from coevolution. *Genome Biology 5,* 5, R35.

BRANDES, U., GAERTLER, M., AND WAGNER, D. 2003. Experiments on graph clustering algorithms. In *ESA*. 568–579.

BREITKREUTZ, B.-J., STARK, C., AND TYERS, M. 2003. The GRID: the General Repository for Interaction Datasets. *Genome Biology 4,* 3, R23.

BRENNER, S. E. 1999. Errors in genome annotation. *Trends in Genetics 15,* 4, 132–133.

BRIN, S. AND PAGE, L. 1998. The anatomy of a large-scale hypertextual Web search engine. In *Proc. Seventh international conference on World Wide Web (WWW)*. 107–117.

BROWN, D. AND SJLANDER, K. 2006. Functional classification using phylogenomic inference. *PLoS Comput Biol 2,* 6, e77.

BROWN, M. P., GRUNDY, W. N., LIN, D., CRISTINIANI, N., SUGNET, C. W., FUREY, T. S., ARES, M., JR., AND HAUSSLER, D. 2000. Knowledge based analysis of microarray gene expression data by using support vector machines. *Proc Natl Acad Sci 97,* 1, 262–267.

BRUN, C., CHEVENET, F., MARTIN, D., WOJCIK, J., GUENOCHE, A., AND JACQ, B. 2003. Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network. *Genome Biology 5,* 1, R6.

BRUN, C., HERRMANN, C., AND GUENOCHE, A. 2004. Clustering proteins from interaction networks for the prediction of cellular functions. *BMC Bioinformatics 5*, 95.

BRUNAK, S., DANCHIN, A., HATTORI, M., NAKAMURA, H., SHINOZAKI, K., MATISE, T., AND PREUSS, D. 2002. Nucleotide sequence database policies. *Science 298,* 5597, 1333.

BRYAN, K., CUNNINGHAM, P., AND BOLSHAKOVA, N. 2005. Biclustering of expression data using simulated annealing. In *Proc. 18th IEEE Symposium on Computer-Based Medical Systems (CBMS)*. 383–388.

BURGE, C. AND KARLIN, S. 1997. Prediction of complete gene structures in human genomic DNA. *J Mol Biol. 268,* 1, 78–94.

BUTTE, A. J., BAO, L., REIS, B. Y., WATKINS, T. W., AND KOHANE, I. S. 2001. Comparing the similarity of time-series gene expression using signal processing metrics. *J Biomed Inform. 34,* 6, 396–405.

CAI, C., HAN, L., JI, Z., CHEN, X., AND CHEN, Y. 2003. SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence. *Nucleic Acids Research 31,* 13, 3692–3697.

CAI, Y.-D. AND DOIG, A. J. 2004. Prediction of saccharomyces cerevisiae protein functional class from functional domain composition. *Bioinformatics 20,* 8, 1292–1300.

CALIFANO, A. 2000. SPLASH: structural pattern localization analysis by sequential histograms. *Bioinformatics 16,* 4, 341–357.

CAMON, E., MAGRANE, M., BARRELL, D., BINNS, D., FLEISCHMANN, W., KERSEY, P., MULDER, N., OINN, T., MASLEN, J., COX, A., AND APWEILER, R. 2003. The gene ontology annotation (GOA) project: Implementation of go in swiss-prot and trembl and and interpro. *Genome Research 13,* 4, 662–672.

CAMPBELL, N. A. AND REECE, J. B. 2004. *Biology*, 7 ed. Benjamin Cummings.

CAMUS, J.-C., PRYOR, M. J., MDIGUE, C., AND COLE, S. T. 2002. Re-annotation of the genome sequence of *mycobacterium tuberculosis* h37rv. *Microbiology 148,* Pt 10, 2967–2973.

CARROLL, S. AND PAVLOVIC, V. 2006. Protein classification using probabilistic chain graphs and the gene ontology structure. *Bioinformatics 22,* 15, 1871–1878.

CAVANAGH, J., FAIRBROTHER, W., PALMER III, A., AND SKELTON, N. 1996. *Protein NMR Spectroscopy: Principles and Practice*. Academic Press Inc.

CHAKRABARTI, S., VENKATRAMANAN, K., AND SOWDHAMINI, R. 2003. SMoS: a database of structural motifs of protein superfamilies. *Protein Engineering 16,* 11, 791–793.

CHALMEL, F., LARDENOIS, A., THOMPSON, J., MULLER, J., SAHEL, J.-A., LEVEILLARD, T., AND POCH, O. 2005. GOAnno: GO annotation based on multiple alignment. *Bioinformatics 21,* 9, 2095–2096.

CHEN, J., HSU, W., LEE, M. L., AND NG, S.-K. 2006. NeMoFinder: dissecting genome-wide protein-protein interactions with meso-scale network motifs. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. 106–115.

CHEN, J., HSU, W., LEE, M.-L., AND NG, S.-K. 2007. Labeling network motifs in protein interactomes for protein function prediction. In *Proc 23rd International Conference on Data Engineering (ICDE)*. 546–555.

CHEN, Y. AND XU, D. 2003. Computational analyses of high–throughput protein–protein interaction data. *Curr Protein Pept Sci. 4,* 3, 159–181.

CHEN, Y. AND XU, D. 2004. Global protein function annotation through mining genome-scale data in yeast Saccharomyces cerevisiae. *Nucleic Acids Research 32,* 21, 6414–6424.

CHENG, B. Y. M., CARBONELL, J. G., AND KLEIN-SEETHARAMAN, J. 2005. Protein classification based on text document classification techniques. *Proteins 58,* 4, 955–970.

CHENG, J., HATZIS, C., HAYASHI, H., KROGEL, M.-A., MORISHITA, S., PAGE, D., AND SESE, J. 2002. Kdd cup 2001 report. *SIGKDD Explorations 3,* 2, 47–64.

CHENG, Y. AND CHURCH, G. M. 2000. Biclustering of expression data. In *Proc. Eighth International Conference on Intelligent Systems for Molecular Biology (ISMB)*. 93–103.

CHIANG, J.-H. AND YU, H.-C. 2005. Literature extraction of protein functions using sentence pattern mining. *IEEE Transactions on Knowledge and Data Engineering 17,* 8, 1088–1098.

CHO, R., CAMPBELL, M., WINZELER, E., STEINMETZ, L., CONWAY, A., WODICKA, L., WOLFSBERG, T., GABRIELIAN, A., LANDSMAN, D., AND LOCKHART, D. 1998. A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell 2,* 1, 65–73.

CHUA, H. N., SUNG, W.-K., AND WONG, L. 2006. Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. *Bioinformatics 22,* 13, 1623–1630.

CHUA, H. N., SUNG, W.-K., AND WONG, L. 2007. Using indirect protein interactions for the prediction of Gene Ontology functions. *BMC Bioinformatics 8,* Suppl 4, S8.

CLARE, A., KARWATH, A., OUGHAM, H., AND KING, R. D. 2006. Functional bioinformatics for *Arabidopsis thaliana*. *Bioinformatics 22,* 13, 1674–1674.

CLARE, A. AND KING, R. D. 2003a. Data mining the yeast genome in a lazy functional language. In *Proc. 5th International Symposium on Practical Aspects of Declarative Languages (PADL)*. 19–36.

CLARE, A. AND KING, R. D. 2003b. Predicting gene function in Saccharomyces cerevisiae. *Bioinformatics 19,* Suppl 2, ii42–ii49.

CLARK, J. I., BROOKSBANK, C., AND LOMAX, J. 2005. It's all GO for plant scientists. *Plant Physiology 138,* 3, 1268–1279.

CONNOLLY, M. L. 1983. Solvent-accessible surfaces of proteins and nucleic acids. *Science 221,* 4612, 709–713.

COSTANZO, M. C., HOGAN, J. D., CUSICK, M. E., DAVIS, B. P., HODGES, A. M. F. P. E., KONDU, P., LENGIEZA, C., LEW-SMITH, J. E., LINGNER, C., ROBERG-PEREZ, K. J., TILLBERG, M., BROOKS, J. E., AND GARRELS, J. I. 2000. The yeast proteome database (YPD) and Caenorhabditis elegans proteome database (WormPD): comprehensive resources for the organization and comparison of model organism protein information. *Nucleic Acids Research 28,* 1, 73–76.

COUTINHO, P. M. AND HENRISSAT, B. Carbohydrate-active enzymes: an integrated database approach. In *Recent Advances in Carbohydrate Bioengineering*, H. Gilbert, G. Davies, B. Henrissat, and B. Svensson, Eds. The Royal Society of Chemistry, 3–12.

COUTO, F., SILVA, M., AND COUTINHO, P. 2003. ProFAL: Protein functional annotation through literature. In *Proc. VIII Conf. on Software Engineering and Databases (JISBD)*. 747–756.

COUTO, F. M., SILVA, M. J., AND COUTINHO, P. M. 2005. Finding genomic ontology terms in text using evidence content. *BMC Bioinformatics 6,* Suppl 1, S21.

COYNE, J. A. AND ORR, H. A. 2004. *Speciation*. Sinauer Associates, Inc.

DANDEKAR, T., SNEL, B., HUYNEN, M., AND BORK, P. 1998. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci. 23,* 9, 324–328.

DARWIN, C. R. 1909. *The Origin of Species*. The Harvard Classics. P.F. Collier & Son.

DASH, M. AND LIU, H. 1997. Feature selection for classification. *Intelligent Data Analysis 1,* 3, 131–156.

DATE, S. V. AND MARCOTTE, E. M. 2003. Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. *Nat Biotechnol. 21,* 9, 1055–1062.

DATE, S. V. AND MARCOTTE, E. M. 2005. Protein function prediction using the Protein Link EXplorer (PLEX). *Bioinformatics 21,* 10, 2558–2559.

DAVIES, J., FENSEL, D., AND VAN HARMELEN, F., Eds. 2003. *Towards the Semantic Web: Ontology-driven Knowledge Management*. John Wiley & Sons.

DAYHOFF, M. O., SCHWARTZ, R. M., AND ORCUTT, B. C. 1978. A model of evolutionary change in proteins. *Atlas of Protein Sequence and Structure 15,* Suppl 3, 345–358.

DE BRUIJN, B. AND MARTIN, J. 2003. Finding gene function using litminer. In *Proc. Text Retrieval Conference (TREC)*. 486–494.

DEANE, C. M., SALWINSKI, L., XENARIOS, I., AND EISENBERG, D. 2002. Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol Cell Proteomics 1,* 5, 349–356.

DENG, M., CHEN, T., AND SUN, F. 2004. An integrated probabilistic model for functional prediction of proteins. *J Comput Biol. 11,* 2-3, 463–475.

DENG, M., SUN, F., AND CHEN, T. 2003. Assessment of the reliability of protein–protein interactions and protein function prediction. In *Pac Symp Biocomputing*. 140–151.

DENG, M., TU, Z., SUN, F., AND CHEN, T. 2004. Mapping Gene Ontology to proteins based on proteinprotein interaction data. *Bioinformatics 20,* 6, 895–902.

DENG, M., ZHANG, K., MEHTA, S., CHEN, T., AND SUN, F. 2003. Prediction of protein function using protein–protein interaction data. *J Comput Biology 10,* 6, 947–960.

DENG, X. AND ALI, H. H. 2004. A hidden markov model for gene function prediction from sequential expression data. In *Proc. CSB*. 670–671.

DEVOS, D. AND VALENCIA, A. 2000. Practical limits of function prediction. *Proteins 41,* 1, 98–107.

DIETMANN, S., PARK, J., NOTREDAME, C., HEGER, A., LAPPE, M., AND HOLM, L. 2001. A fully automatic evolutionary classification of protein folds: Dali Domain Dictionary version 3. *Nucleic Acids Research 29,* 1, 55–57.

DOBSON, P. D. AND DOIG, A. J. 2003. Distinguishing enzyme structures from non-enzymes without alignments. *J Mol Biol. 330,* 4, 771–783.

DOBSON, P. D. AND DOIG, A. J. 2005. Predicting enzyme class from protein structure without alignments. *J Mol Biol. 345,* 1, 187–199.

DOBZHANSKY, T. 1973. Nothing in biology makes sense except in the light of evolution. *The American Biology Teacher 35,* 125–129.

DOERKS, T., BAIROCH, A., AND BORK, P. 1998. Protein annotation: detective work for function prediction. *Trends in Genetics 14,* 6, 248–250.

DRENTH, J. 1999. *Principles of protein X-ray crystallography*, 2 ed. Springer-Verlag.

DROIT, A., POIRIER, G. G., AND HUNTER, J. M. 2005. Experimental and bioinformatic approaches for interrogating protein–protein interactions to determine protein function. *Journal of Molecular Endocrinology 34,* 2, 263–280.

DUBY, G., FOURY, F., RAMAZZOTTI, A., HERRMANN, J., AND LUTZ, T. 2002. A non-essential function for yeast frataxin in ironsulfur cluster assembly. *Hum Mol Genet. 11,* 21, 2635–2643.

DUDA, R. O., HART, P. E., AND STORK, D. G. 2000. *Pattern Classification (2nd Edition)*. Wiley-Interscience.

DUGGAN, D. J., BITTNER, M., CHEN, Y., MELTZER, P., AND TRENT, J. M. 1999. Expression profiling using cDNA microarrays. *Nature Genetics 21,* 1 Suppl, 10–14.

DUNN, R., DUDBRIDGE, F., AND SANDERSON, C. M. 2005. The use of edge-betweenness clustering to investigate biological function in protein interaction networks. *BMC Bioinformatics 6,* 1, 39.

DWIGHT, S. S., HARRIS, M. A., DOLINSKI, K., BALL, C. A., BINKLEY, G., CHRISTIE, K. R., FISK, D. G., ISSEL-TARVER, L., SCHROEDER, M., SHERLOCK, G., SETHURAMAN, A., WENG, S., BOTSTEIN, D., AND CHERRY, J. M. 2002. Saccharomyces genome database (SGD) provides secondary gene annotation using the gene ontology (GO). *Nucleic Acids Research 30,* 1, 69–72.

EDDY, S. R. 1998. Profile hidden Markov models. *Bioinformatics 14,* 9, 755–763.

EDELSBRUNNER, H., FACELLO, M., AND LIANG, J. 1998. On the definition and the construction of pockets in macromolecules. *Discrete Appl. Math. 88,* 1-3, 83–102.

EDWARDS, D. 2000. *Introduction to Graphical Modelling*. Springer.

EHRLER, F., GEISSBUHLER, A., JIMENO, A., AND RUCH, P. 2005. Data-poor categorization and passage retrieval for gene ontology annotation in Swiss-Prot. *BMC Bioinformatics 6,* Suppl 1, S23.

EISEN, J. A. 1998. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Research 8,* 3, 163–167.

EISEN, M. B., SPELLMAN, P. T., BROWNDAGGER, P. O., AND BOTSTEIN, D. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U.S.A. 95,* 25, 14863–14868.

EISENBERG, D., MARCOTTE, E. M., XENARIOS, I., AND YEATES, T. O. 2000. Protein function in the post-genomic era. *Nature 405,* 6788, 823–826.

EISNER, R., POULIN, B., SZAFRON, D., LU, P., AND GREINER, R. 2005. Improving protein function prediction using the hierarchical structure of the Gene Ontology. In *Proc. IEEE CIBCB*.

ENAULT, F., SUHRE, K., ABERGEL, C., POIROT, O., AND CLAVERIE, J.-M. 2003a. Annotation of bacterial genomes using improved phylogenomic profiles. *Bioinformatics 19,* Suppl 1, i105–i107.

ENAULT, F., SUHRE, K., ABERGEL, C., POIROT, O., AND CLAVERIE, J.-M. 2003b. Phydbac (phylogenomic display of bacterial genes): An interactive resource for the annotation of bacterial genomes. *Nucleic Acids Research 31,* 13, 3720–3722.

ENAULT, F., SUHRE, K., ABERGEL, C., POIROT, O., AND CLAVERIE, J.-M. 2004. Phydbac2: improved inference of gene function using interactive phylogenomic profiling and chromosomal location analysis. *Nucleic Acids Research 32,* Web Server issue, W336–W339.

ENAULT, F., SUHRE, K., AND CLAVERIE, J.-M. 2005. Phydbac "Gene Function Predictor" : a gene annotation tool based on genomic context analysis. *BMC Bioinformatics 6,* 247.

ENGELHARDT, B. E., JORDAN, M. I., AND BRENNER, S. E. 2006. A graphical model for predicting protein molecular function. In *ICML '06: Proceedings of the 23rd international conference on Machine learning*. 297–304.

ENGELHARDT, B. E., JORDAN, M. I., MURATORE, K. E., AND BRENNER, S. E. 2005. Protein molecular function prediction by bayesian phylogenomics. *PLoS Comput Biol. 1,* 5, e45.

ENRIGHT, A. J., DONGEN, S. V., AND OUZOUNIS, C. A. 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research 30,* 7, 1575–1584.

ENRIGHT, A. J. AND OUZOUNIS, C. A. 2001. Functional associations of proteins in entire genomes by means of exhaustive detection of gene fusions. *Genome Biology 2,* 9, RESEARCH0034.

ERICKSON, H. P. 1989. Cooperativity in protein-protein association: the structure and stability of the actin filament. *J. Mol. Biol. 206,* 3, 465–474.

ERNST, J., NAU, G. J., AND BAR-JOSEPH, Z. 2005. Clustering short time series gene expression data. *Bioinformatics 21,* Suppl 1, i159–i168.

ERTOZ, L., STEINBACH, M., AND KUMAR, V. 2003. Finding clusters of different sizes and shapes and and densities in noisy and high dimensional data. In *Proc. SIAM International Conference on Data Mining*.

ESPADALER, J., ARAGUES, R., ESWAR, N., MARTI-RENOM, M. A., QUEROL, E., AVILES, F. X., SALI, A., AND OLIVA, B. 2005. Detecting remotely related proteins by their interactions and sequence similarity. *Proc Natl Acad Sci U.S.A. 102,* 20, 7151–7156.

ESPADALER, J., QUEROL, E., AVILES, F. X., AND OLIVA, B. 2006. Identification of function-associated loop motifs and application to protein function prediction. *Bioinformatics 22,* 18, 2237–2243.

FELLBAUM, C., Ed. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.

FELSENSTEIN, J. 1989. PHYLIP – phylogeny inference package (version 3.2). *Cladistics 5,* 164–166.

FERRE, F., AUSIELLO, G., ZANZONI, A., AND HELMER-CITTERICH, M. 2004. SURFACE: a database of protein surface regions for functional annotation. *Nucleic Acids Research 32,* Database issue, D240–D244.

FETROW, J. S. 1995. Omega loops: nonregular secondary structures significant in protein function and stability. *FASEB Journal 9,* 9, 708–717.

FETROW, J. S., SIEW, N., GENNARO, J. A. D., MARTINEZ-YAMOUT, M., DYSON, H. J., AND SKOLNICK, J. 2001. Genomic-scale comparison of sequence- and structure-based methods of function prediction: Does structure provide additional insight? *Protein Science 10,* 5, 1005–1014.

FETROW, J. S. AND SKOLNICK, J. 1998. Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases. *J Mol Biol. 281,* 5, 949–968.

FLEISCHMANN, W., MOLLER, S., GATEAU, A., AND APWEILER, R. 1999. A novel method for automatic functional annotation of proteins. *Bioinformatics 15,* 3, 228–233.

FLYBASE CONSORTIUM. 2003. The FlyBase database of the drasophila genome projects and community literature. *Nucleic Acids Research 31,* 172–175.

FRANTZI, K., ANANIADOU, S., AND MIMA, H. 2000. Automatic recognition of multiword terms: the c-value/nc-value method. *International Journal on Digital Libraries 3,* 2, 115–130.

FRASER, A. G. AND MARCOTTE, E. M. 2004. A probabilistic view of gene function. *Nat Genetics 36,* 6, 559–564.

FRENCH, L. 2005. Fast Protein Superfamily Classification using Principal Component Null Space Analysis. M.S. thesis, Faculty of Graduate Studies and Research and University of Windsor, Ontario and Canada. Appendix A: A Survey on Remote Homology Detection and Protein Superfamily Classification.

FRIEDBERG, I., HARDER, T., AND GODZIK, A. 2006. JAFA: a protein function annotation meta-server. *Nucleic Acids Research 34,* Web Server issue, W379–W381.

FULLER, G. M., BOUGHTER, J. M., AND MORAZZANI, M. 1974. Evidence for multiple polypeptide chains in the membrane protein spectrin. *Biochemistry 13,* 15, 3036–3041.

GAASTERLAND, T. AND RAGAN, M. A. 1998. Microbial genescapes: phyletic and functional patterns of ORF distribution among prokaryotes. *Microb Comp Genomics 3,* 4, 199–217.

GABALDON, T. 2006. Computational approaches for the prediction of protein function in the mitochondrion. *Am J Physiol Cell Physiol 291,* C1121–C1128.

GABALDON, T. AND HUYNEN, M. A. 2004. Prediction of protein function and pathways in the genome era. *Cell Mol Life Sci. 61,* 7–8, 930–944.

GARDINER-GARDEN, M. AND LITTLEJOHN, T. 2001. A comparison of microarray databases. *Briefings in Bioinformatics 2,* 2, 143–148.

GÄRTNER, T., FLACH, P. A., AND WROBEL, S. 2003. On graph kernels: Hardness results and efficient alternatives. In *COLT*. 129–143.

GASCUEL, O. 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol Biol Evol. 14,* 7, 685–695.

GAVIN, A.-C., BOSCHE, M., KRAUSE, R., GRANDI, P., MARZIOCH, M., BAUER, A., SCHULTZ, J., RICK, J. M., MICHON, A.-M., CRUCIAT, C.-M., REMOR, M., HOFERT, C., SCHELDER, M., BRAJENOVIC, M., RUFFNER, H., MERINO, A., KLEIN, K., HUDAK, M., DICKSON, D., RUDI, T., GNAU, V., BAUCH, A., BASTUCK, S., HUHSE, B., LEUTWEIN, C., HEURTIER, M.-A., COPLEY, R. R., EDELMANN, A., QUERFURTH, E., RYBIN, V., DREWES, G., RAIDA, M., BOUWMEESTER, T., BORK, P., SERAPHIN, B., KUSTER, B., NEUBAUER1, G., AND SUPERTI-FURGA, G. 2002. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature 415,* 6868, 141–147.

GENNAROA, J. A. D., SIEW, N., HOFFMANA, B. T., ZHANGB, L., SKOLNICK, J., NEILSON, L. I., AND FETROW, J. S. 2001. Enhanced functional annotation of protein sequences via the use of structural descriptors. *J Struct Biol. 134,* 2–3, 232–245.

GERICKE, N. M. 2005. Scientific models of gene function - a historical overview. A basis for teaching gene function.

GERLT, J. A. AND BABBITT, P. C. 2000. Can sequence determine function? *Genome Biology 1,* 5, REVIEWS0005.

GETHER, U. 2000. Uncovering molecular mechanisms involved in activation of G protein-coupled receptors. *Endocr Rev. 21,* 1, 90–113.

GIOT, L., BADER, J. S., BROUWER, C., CHAUDHURI, A., KUANG, B., LI, Y., HAO, Y. L., OOI, C. E., 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, C. A., FINLEY, R. L., JR., WHITE, K. P., BRAVERMAN, M., JARVIE, T., GOLD, S., LEACH, M., KNIGHT, J., SHIMKETS, R. A., MCKENNA, M. P., CHANT, J., AND ROTHBERG, J. M. 2003. A protein interaction map of drosophila melanogaster. *Science 302,* 5651, 1727–1736.

GIRVAN, M. AND NEWMAN, M. E. J. 2002. Community structure in social and biological networks. *Proc Natl Acad Sci U.S.A. 99,* 12, 7821–7826.

GO CONSORTIUM. 2006. The Gene Ontology (GO) project in 2006. *Nucleic Acids Research 34,* Database issue, D322–D326.

GODZIK, A., JAMBON, M., AND FREIDBERG, I. 2007. Computational protein function prediction: Are we making progress? *Cell Mol Life Sci.*

GOH, C.-S., BOGAN, A. A., JOACHIMIAK, M., WALTHER, D., AND COHEN, F. E. 2000. Co-evolution of proteins with their interaction partners. *J Mol Biol. 299,* 2, 283–293.

GOMEZ-PEREZ, A., CORCHO, O., AND FERNANDEZ-LOPEZ, M. 2004. *Ontological Engineering : with examples from the areas of Knowledge Management and e-Commerce and the Semantic Web. First Edition (Advanced Information and Knowledge Processing)*. Springer.

GOTO, S., OKUNO, Y., HATTORI, M., NISHIOKA, T., AND KANEHISA, M. 2002. Ligand: database of chemical compounds and reactions in biological pathways. *Nucleic Acids Research 30*, 1, 402–404.

GUI, J. AND LI, H. 2003. Mixture functional discriminant analysis for gene function classification based on time course gene expression data. In *Proc. Joint Statistical Meeting (Biometrics Section)*.

GULDENER, U., MUNSTERKOTTER, M., KASTENMULLER, G., STRACK, N., VAN HELDEN, J., LEMER, C., RICHELLES, J., WODAK, S. J., GARCIA-MARTINEZ, J., PEREZ-ORTIN, J. E., MICHAEL, H., KAPS, A., TALLA, E., DUJON, B., ANDRE, B., SOUCIET, J. L., MONTIGNY, J. D., BON, E., GAILLARDIN, C., AND MEWES, H. W. 2005. CYGD: the comprehensive yeast genome database. *Nucleic Acids Research 33*, Database issue, D364–D368.

GUMBEL, E. J. 2004. *Statistics of Extremes*. Dover Publications.

GURUPRASAD, K., PRASAD, M. S., AND KUMAR, G. R. 2000. Database of structural motifs in proteins. *Bioinformatics 16*, 4, 372–375.

GUTHKE, R., SCHMIDT-HECK, W., HAHN, D., AND PFAFF, M. 2000. Gene expression data mining for functional genomics. In *Proc. European Symposium on Intelligent Techniques*. 170–177.

GUYON, I., WESTON, J., BARNHILL, S., AND VAPNIK, V. 2002. Gene selection for cancer classification using support vector machines. *Machine Learning 46*, 1-3, 389–422.

HAFT, D. H., SELENGUT, J. D., AND WHITE, O. 2003. The TIGRFAMs database of protein families. *Nucleic Acids Research 31*, 1, 371–373.

HALL, D. L. 2001. *Handbook of Multisensor Data Fusion*. CRC Press.

HAN, L. Y., ZHENG, C. J., LIN, H. H., CUI, J., LI, H., ZHANG, H. L., TANG, Z. Q., AND CHEN, Y. Z. 2005. Prediction of functional class of novel plant proteins by a statistical learning method. *New Phytologist 168*, 1, 109.

HANNENHALLI, S. S. AND RUSSELL, R. B. 2000. Analysis and prediction of functional sub-types from protein sequence alignments. *J Mol Biol. 303*, 1, 61–76.

HANSEN, M. H. AND YU, B. 2001. Model selection and the principle of minimum description length. *Journal of the American Statistical Association 96*, 454, 746–774.

HARGER, C., SKUPSKI, M., BINGHAM, J., FARMER, A., HOISIE, S., ET AL. 1998. The Genome Sequence DataBase (GSDB): improving data quality and data access. *Nucleic Acids Research 26*, 1, 21–26.

HARRIS, T. W., CHEN, N., CUNNINGHAM, F., TELLO-RUIZ, M., ANTOSHECHKIN, I., BASTIANI, C., BIERI, T., BLASIAR, D., BRADNAM, K., CHAN, J., CHEN, C., CHEN, W., KENNY, P. D. E., KISHORE, R., LAWSON, D., LEE, R. R., MULLER, H. M., OZERSKY, C. N. P., PETCHERSKI, A., ROGERS, A., SABO, A., SCHWARZ, E. M., WANG, K. V. A. Q., DURBIN, R., SPIETH, J., STERNBERG, P. W., AND STEIN, L. D. 2004. WormBase: a multi-species resource for nematode biology and genomics. *Nucleic Acids Research 32*, D411–D417.

HART, G. T., RAMANI, A. K., AND MARCOTTE, E. M. 2006. How complete are current yeast and human protein-interaction networks? *Genome Biology 7*, 11, 120.

HARTUV, E. AND SHAMIR, R. 2000. A clustering algorithm based on graph connectivity. *Inf. Process. Lett. 76*, 4-6, 175–181.

HAYETE, B. AND BIENKOWSKA, J. R. 2005. GOTrees: Predicting GO associations from protein domain composition using decision trees. In *Pacific Symposium on Biocomputing*. 140–151.

HEARD, N. A., HOLMES, C. C., STEPHENS, D. A., HAND, D. J., AND DIMOPOULOS, G. 2005. Bayesian coclustering of anopheles gene expression time series: Study of immune defense response to multiple experimental challenges. *Proc Natl Acad Sci U.S.A. 102*, 47, 16939–16944.

HECKERMAN, D. 1995. A Tutorial on Learning Bayesian Networks. Tech. Rep. MSR-TR-95-06, Microsoft Research.

HEGYI, H. AND GERSTEIN, M. 1999. The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J Mol Biol. 288*, 1, 147–164.

HENNIG, S., GROTH, D., AND LEHRACH, H. 2003. Automated Gene Ontology annotation for anonymous sequence data. *Nucleic Acids Research 31*, 13, 3712–3715.

HENRICH, T., RAMIALISON, M., QUIRING, R., WITTBRODT, B., FURUTANI-SEIKI, M., WITTBRODT, J., AND KONDOH, H. 2003. MEPD: a Medaka gene expression pattern database. *Nucleic Acids Research 32,* 1, 72–74.

HERSH, W. 2004. Report on TREC 2003 genomics track first-year results and future plans. *SIGIR Forum 38,* 1, 69–72.

HIGGINS, D. G., THOMPSON, J. D., AND GIBBONS, T. J. 1996. Using CLUSTAL for multiple sequence alignments. *Methods Enzymol. 266,* 383–402.

HILL, D. P., BEGLEY, D. A., FINGER, J. H., HAYAMIZU, T. F., MCCRIGHT, I. J., SMITH, C. M., BEAL, J. S., CORBANI, L. E., BLAKE, J. A., EPPIG, J. T., KADIN, J. A., RICHARDSON, J. E., AND RINGWALD, M. 2004. The mouse Gene Expression Database (GXD): updates and enhancements. *Nucleic Acids Research 32,* Database issue, D568–D571.

HINRICHS, A. S., KAROLCHIK, D., BAERTSCH, R., BARBER, G. P., BEJERANO, G., ET AL. 2006. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Research 34,* Database Issue, D590–D598.

HISHIGAKI, H., NAKAI, K., ONO, T., TANIGAMI, A., AND TAKAGI, T. 2001. Assessment of prediction accuracy of protein function from protein–protein interaction data. *Yeast 18,* 6, 523–531.

HO, Y., GRUHLER, A., HEILBUT, A., BADER, G. D., MOORE, L., ADAMS, S.-L., MILLAR, A., TAYLOR, P., BENNETT, K., BOUTILIER, K., YANG, L., WOLTING, C., DONALDSON, I., SCHANDORFF, S., SHEWNARANE, J., VO, M., TAGGART, J., GOUDREAULT, M., MUSKAT, B., ALFARANO, C., DEWAR, D., LIN, Z., MICHALICKOVA, K., WILLEMS, A. R., SASSI, H., NIELSEN, P. A., RASMUSSEN, K. J., ANDERSEN, J. R., JOHANSEN, L. E., HANSEN, L. H., JESPERSEN, H., PODTELEJNIKOV, A., NIELSEN, E., CRAWFORD, J., POULSEN, V., SORENSEN, B. D., MATTHIESEN, J., HENDRICKSON, R. C., GLEESON, F., PAWSON, T., MORAN, M. F., DUROCHER, D., MANN, M., HOGUE, C. W. V., FIGEYS, D., AND TYERS, M. 2002. Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry. *Nature 415,* 6868, 180–183.

HOBOHM, U. AND SANDER, C. 1994. Enlarged representative set of protein structures. *Protein Science 3,* 3, 522–524.

HODGMAN, T. C. 2000. A historical perspective on gene/protein functional assignment. *Bioinformatics 16,* 1, 10–15.

HOLM, L. AND PARK, J. 2000. DaliLite workbench for protein structure comparison. *Bioinformatics 16,* 6, 566–567.

HOLM, L. AND SANDER, C. 1994. The FSSP database of structurally aligned protein fold families. *Nucleic Acids Research 22,* 3600–3609.

HOLME, P. AND HUSS, M. 2005. Role-similarity based functional prediction in networked systems: Application to the yeast proteome. *Journal of The Royal Society Interface 2,* 327–333.

HONIG, B. AND NICHOLLS, A. 1995. Classical electrostatics in biology and chemistry. *Science 268,* 5214, 1144–1149.

HOPFIELD, J. J. AND TANK, D. W. 1986. Computing with neural circuits: a model. *Science 233,* 4764, 625–633.

HORN, F., BETTLER, E., OLIVEIRA, L., CAMPAGNE, F., COHEN, F. E., AND VRIEND, G. 2003. Gpcrdb information system for g protein-coupled receptors. *Nucleic Acids Research 31,* 1, 294–297.

HORN, F., VRIEND, G., AND COHEN, F. E. 2001. Collecting and harvesting biological data: the GPCRDB and NucleaRDB information systems. *Nucleic Acids Research 29,* 1, 346–349.

HOU, J., JUN, S.-R., ZHANG, C., AND KIM, S.-H. 2005. Global mapping of the protein structure space and application in structure-based inference of protein function. *Proc Natl Acad Sci U.S.A. 102,* 10, 3651–3656.

HOU, J., SIMS, G. E., ZHANG, C., AND KIM, S.-H. 2003. A global representation of the protein fold space. *Proc Natl Acad Sci U.S.A. 100,* 5, 2386–2390.

HU, H., YAN, X., HUANG, Y., HAN, J., AND ZHOU, X. J. 2005. Mining coherent dense subgraphs across massive biological networks for functional discovery. *Bioinformatics 21,* Suppl. 1, i213–i221.

HU, P., BADER, G., WIGLE, D. A., AND EMILI, A. 2007. Computational prediction of cancer-gene function. *Nature Reviews Cancer 7,* 23–34.

HUALA, E., DICKERMAN, A., GARCIA-HERNANDEZ, M., WEEMS, D., REISER, L., LAFOND, F., HANLEY, D., KIPHART, D., ZHUANG, J., HUANG, W., MUELLER, L., BHATTACHARYA, D., BHAYA, D., SOBRAI, B., BEAVIS, B., SOMERVILLE, C., AND RHEE, S. 2001. The arabidopsis information resource (TAIR): A comprehensive database and web-based information retrieval and analysis and visualization system for a model plant. *Nucleic Acids Research 29,* 102–105.

HUAN, J., BANDYOPADHYAY, D., WANG, W., SNOEYINK, J., PRINS, J., AND TROPSHA, A. 2005. Comparing graph representations of protein structure for mining family-specific residue-based packing motifs. *Journal of Computational Biology 12,* 6, 657–671.

HUANG, J. Y. AND BRUTLAG, D. L. 2001. The EMOTIF database. *Nucleic Acids Research 29,* 1, 202–204.

HUBBARD, T., ANDREWS, D., CACCAMO, M., CAMERON, G., CHEN, Y., CLAMP, M., CLARKE, L., COATES, G., COX, T., CUNNINGHAM, F., CURWEN, V., CUTTS, T., DOWN, T., DURBIN, R., FERNANDEZ-SUAREZ, X. M., GILBERT, J., HAMMOND, M., HERRERO1, J., HOTZ, H., HOWE, K., IYER, V., JEKOSCH, K., KAHARI1, A., KASPRZYK, A., KEEFE1, D., KEENAN, S., KOKOCINSCI, F., LONDON1, D., LONGDEN, I., MCVICKER, G., MELSOPP, C., MEIDL, P., POTTER, S., PROCTOR, G., RAE, M., RIOS, D., SCHUSTER, M., SEARLE, S., SEVERIN, J., SLATER, G., SMEDLEY, D., SMITH, J., SPOONER, W., STABENAU, A., STALKER, J., STOREY, R., TREVANION, S., URETA-VIDAL, A., VOGEL, J., WHITE, S., WOODWARK, C., AND BIRNEY, E. 2005. Ensembl 2005. *Nucleic Acids Research 33,* Database issue, D447–D453.

HUGHES, T. R., MARTON, M. J., JONES, A. R., ROBERTS, C. J., STOUGHTON, R., ARMOUR, C. D., BENNETT, H. A., COFFEY, E., DAI, H., HE, Y. D., KIDD, M. J., KING, M. M., MEYER, M. R., SLADE, D., LUM, P. Y., STEPANIANTS, S. B., SHOEMAKER, D. D., GACHOTTE, D., CHAKRABURTTY, K., SIMON, J., BARD, M., AND FRIEND, S. H. 2000. Functional discovery via a compendium of expression profiles. *Cell 102,* 1, 109–126.

HULO, N., BAIROCH, A., BULLIARD, V., CERUTTI, L., CASTRO, E. D., LANGENDIJK-GENEVAUX, P. S., PAGNI, M., AND SIGRIST, C. J. A. 2006. The PROSITE database. *Nucleic Acids Research 34,* Database issue, D227–D230.

HUTCHISON III, C. A., PETERSON, S. N., GILL, S. R., CLINE, R. T., WHITE, O., FRASER, C. M., SMITH, H. O., AND VENTER, J. C. 1999. Global transposon mutagenesis and a minimal Mycoplasma genome. *Science 286,* 5547, 2165–2169.

HUTTENHOWER, C., HIBBS, M., MYERS, C., AND TROYANSKAYA, O. G. 2006. A scalable method for integration and functional analysis of multiple microarray datasets. *Bioinformatics 22,* 23, 2890–2897.

HUYNEN, M., SNEL, B., LATHE III, W., AND BORK, P. 2000. Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Research 10,* 8, 1204–1210.

HUYNEN, M. A., SNEL, B., BORK, P., AND GIBSON, T. J. 2001. The phylogenetic distribution of frataxin indicates a role in iron-sulfur cluster protein assembly. *Hum Mol Genet. 10,* 21, 2463–2468.

HUYNEN, M. A., SNEL, B., VON MERING, C., AND BORK, P. 2003. Function prediction and protein networks. *Curr Opin Cell Biology 15,* 2, 191–198.

HVIDSTEN, T., KOMOROWSKI, J., SANDVIK, A., AND LAEGREID, A. 2001. Predicting gene function from gene expressions and ontologies. In *Proc. Pacific Symposium on Biocomputing (PSB)*. 299–310.

ITO, T., CHIBA, T., OZAWADAGGER, R., YOSHIDA, M., HATTORIDAGGER, M., AND SAKAKI, Y. 2001. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U.S.A. 98,* 8, 4569–4574.

IYER, V. R., EISEN, M. B., ROSS, D. T., SCHULER, G., MOORE, T., LEE, J. C. F., TRENT, J. M., STAUDT, L. M., JR., J. H., BOGUSKI, M. S., LASHKARI, D., SHALON, D., BOTSTEIN, D., AND BROWN, P. O. 1999. The transcriptional program in the response of human fibroblasts to serum. *Science 283,* 5398, 83–87.

JAAKKOLA, T., DIEKHANS, M., AND HAUSSLER, D. 2000. A discriminative framework for detecting remote protein homologies. *J Comput Biol. 7,* 1–2, 95–114.

JAIN, A. K. AND DUBES, R. C. 1988. *Algorithms for Clustering Data*. Prentice-Hall and Inc.

JARVIS, R. AND PATRICK, E. 1973. Clustering using a similarity measure based on shared nearest neighbors. *IEEE Trans. Comput. 22,* 1025–1034.

JENSEN, L. J., GUPTA, R., BLOM, N., DEVOS, D., TAMAMES, J., KESMIR, C., NIELSEN, H., STAERFELDT, H. H., RAPACKI, K., WORKMAN, C., ANDERSEN, C. A. F., KNUDSEN, S., KROGH, A., VALENCIA, A., AND BRUNAK, S. 2002. Prediction of human protein function from post-translational modifications and localization features. *J Mol Biol. 319,* 5, 1257–1265.

JENSEN, L. J., GUPTA, R., STAERFELDT, H.-H., AND BRUNAK, S. 2003. Prediction of human protein function according to Gene Ontology categories. *Bioinformatics 19,* 5, 635–642.

JIANG, D., PEI, J., RAMANATHAN, M., TANG, C., AND ZHANG, A. 2004. Mining coherent gene clusters from gene-sample-time microarray data. In *Proc. Tenth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*. 430–439.

JIANG, D., TANG, C., AND ZHANG, A. 2004. Cluster analysis for gene expression data: A survey. *IEEE Transactions on Knowledge and Data Engineering 16,* 11, 1370–1386.

JOACHIMS, T. 1999. Transductive inference for text classification using support vector machines. In *Proc. Sixteenth International Conference on Machine Learning (ICML)*. 200–209.

JONES, S. AND THORNTON, J. M. 2004. Searching for functional sites in protein structures. *Curr Opin Chem Biol. 8,* 1, 3–7.

JR, E. S., GRECHKIN, Y., MIKHAILOVA, N., AND SELKOV, E. 1998. MPW: the metabolic pathways database. *Nucleic Acids Research 26,* 1, 43–45.

KANEHISA, M., GOTO, S., KAWASHIMA, S., OKUNO, Y., AND HATTORI, M. 2004. The KEGG resource for deciphering the genome. *Nucleic Acids Research 32,* Database issue, D277–D280.

KARAOZ, U., MURALI, T. M., LETOVSKY, S., ZHENG, Y., DING, C., CANTOR, C. R., AND KASIF, S. 2004. Whole-genome annotation by using evidence integration in functional-linkage networks. *Proc Natl Acad Sci U.S.A. 101,* 9, 2888–2893.

KARCHIN, R., KARPLUS, K., AND HAUSSLER, D. 2002. Classifying G-protein coupled receptors with support vector machines. *Bioinformatics 18,* 1, 147–159.

KARPLUS, K., BARRETT, C., AND HUGHEY, R. 1998. Hidden Markov models for detecting remote protein homologies. *Bioinformatics 14,* 10, 846–856.

KARZAI, A. W., SUSSKIND, M. M., AND SAUER, R. T. 1999. SmpB and a unique RNA-binding protein essential for the peptide-tagging activity of SsrA (tmRNA). *EMBO Journal 18,* 13, 3793–3799.

KATO, K., YAMASHITA, R., MATOBA, R., MONDEN, M., NOGUCHI, S., TAKAGI, T., AND NAKAI, K. 2005. Cancer gene expression database ($cged$): a database for gene expression profiling with accompanying clinical information of human cancer tissues. *Nucleic Acids Research 33,* Database issue, D533–D536.

KAWABATA, T. 2003. MATRAS: a program for protein 3D structure comparison. *Nucleic Acids Research 31,* 13, 3367–3369.

KECK, H.-P. AND WETTER, T. 2003. Functional classification of proteins using a nearest neighbour algorithm. *In Silico Biology 3,* 3, 265–275.

KELLEY, B. P., SHARAN, R., KARP, R. M., SITTLER, T., ROOT, D. E., STOCKWELL, B. R., , AND IDEKER, T. 2003. Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *Proc Natl Acad Sci U.S.A. 100,* 20, 11394–11399.

KEMMEREN, P. AND HOLSTEGE, F. 2003. Integrating functional genomics data. *Biochem Soc Trans. 31,* 6, 1484–1487.

KEMMEREN, P., KOCKELKORN, T. T. J. P., BIJMA, T., DONDERS, R., AND HOLSTEGE, F. C. P. 2005. Predicting gene function through systematic analysis and quality assessment of high-throughput data. *Bioinformatics 21,* 8, 1644–1652.

KENSCHE, P. R., VAN NOORT, V., DUTILH, B. E., AND HUYNEN, M. A. 2007. Practical and theoretical advances in predicting the function of a protein by its phylogenetic distribution. *J R Soc Interface*.

KERSEY, P. J., DUARTE, J., WILLIAMS, A., KARAVIDOPOULOU, Y., BIRNEY, E., AND APWEILER, R. 2004. The International Protein Index: An integrated database for proteomics experiments. *Proteomics 4,* 7, 1985–1988.

KESELER, I. M., COLLADO-VIDES, J., GAMA-CASTRO, S., INGRAHAM, J., PALEY, S., PAULSEN, I. T., PERALTA-GIL, M., AND KARP, P. D. 2005. EcoCyc: a comprehensive database resource for Escherichia coli. *Nucleic Acids Research 33,* Database issue, D334–D337.

KHAN, S., SITU, G., DECKER, K., AND SCHMIDT, C. J. 2003. GoFigure: automated Gene Ontology annotation. *Bioinformatics 19,* 18, 2484–2485.

KIN, T., KATO, T., AND TSUDA, K. 2004. Protein classification via kernel matrix completion. In *Kernel Methods in Computational Biology*, B. Scholkopf, K. Tsuda, and J.-P. Vert, Eds. MIT Press.

KING, O. D., FOULGER, R. E., DWIGHT, S. S., WHITE, J. V., AND ROTH, F. P. 2003. Predicting gene function from patterns of annotation. *Genome Research 13,* 5.

KING, R. D., KARWATH, A., CLARE, A., AND DEHASPE, L. 2000. Accurate prediction of protein functional class from sequence in the Mycobacterium tuberculosis and Escherichia coli genomes using data mining. *Yeast 17,* 4, 283–293.

KING, R. D., KARWATH, A., CLARE, A., AND DEHASPE, L. 2001. The utility of different representations of protein sequence for predicting functional class. *Bioinformatics 17,* 5, 445–454.

KINOSHITA, K., FURUI, J., AND NAKAMURA, H. 2001. Identification of protein functions from a molecular surface database and eF-site. *J Struct Funct Genomics 2*, 1, 9–22.

KINOSHITA, K. AND NAKAMURA, H. 2004. eF-site and PDBjViewer: database and viewer for protein functional sites. *Bioinformatics 20*, 8, 1329–1330.

KIRAC, M., OZSOYOGLU, G., AND YANG, J. 2006. Annotating proteins by mining protein interaction networks. *Bioinformatics 22*, 14, e260–e270.

KLEINBERG, J. M. 1999. Authoritative sources in a hyperlinked environment. *J. ACM 46*, 5, 604–632.

KLEYWEGT, G. J. 1996. Use of non-crystallographic symmetry in protein structure refinement. *Acta Cryst. D52*, 842–857.

KLEYWEGT, G. J. 1999. Recognition of spatial motifs in protein structures. *J Mol Biol. 285*, 4, 1887–1897.

KOIKE, A., NIWA, Y., AND TAKAGI, T. 2005. Automatic extraction of gene/protein biological functions from biomedical text. *Bioinformatics 21*, 7, 1227–1236.

KOLESOVA, G., MEWESA, H. W., AND FRISHMAN, D. 2001. SNAPping up functionally related genes based on context information: a colinearity-free approach. *J Mol Biol. 311*, 4, 639–656.

KOLESOVA, G., MEWESA, H. W., AND FRISHMAN, D. 2002. SNAPper: gene order predicts gene function. *Bioinformatics 18*, 7, 1017–1019.

KOLODNY, R., KOEHL, P., AND LEVITT, M. 2005. Comprehensive evaluation of protein structure alignment methods: Scoring by geometric measures. *J Mol Biol. 346*, 4, 1173–1188.

KOLODNY, R. AND LINIAL, N. 2004. Approximate protein structural alignment in polynomial time. *Proc Natl Acad Sci U.S.A. 101*, 33, 12201–12206.

KOONIN, E. V. AND GALPERIN, M. Y. 2002. *Sequence - Evolution - Function : Computational Approaches in Comparative Genomics*. Springer.

KORBEL, J. O., JENSEN, L. J., VON MERING, C., AND BORK, P. 2004. Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs. *Nature Biotechnology 22*, 7, 911–917.

KOSKI, L. B., GRAY, M. W., LANG, B. F., AND BURGER, G. 2005. AutoFACT: an automatic functional annotation and classification tool. *BMC Bioinformatics 16*, 6, 151.

KRALLINGER, M., PADRON, M., AND VALENCIA, A. 2005. A sentence sliding window approach to extract protein annotations from biomedical articles. *BMC Bioinformatics 6*, Suppl 1, S19.

KRETSCHMANN, E., FLEISCHMANN, W., AND APWEILER, R. 2001. Automatic rule generation for protein annotation with the C4.5 data mining algorithm applied on SWISS-PROT. *Bioinformatics 17*, 10, 920–926.

KRISSINEL, E. AND HENRICK, K. 2004. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Cryst. D60*, 2256–2268.

KROGEL, M.-A. AND SCHEFFER, T. 2004. Multi-relational learning, text mining, and semi-supervised learning for functional genomics. *Mach. Learn. 57*, 1-2, 61–81.

KROGEL, M.-A. AND WROBEL, S. 2001. Transformation-based learning using multirelational aggregation. In *Proc. 11th International Conference on Inductive Logic Programming*. 142–155.

KUAN, J. AND JR, M. H. S. 1993. The mitochondrial carrier family of transport proteins: structural, functional, and evolutionary relationships. *Crit Rev Biochem Mol Biol. 28*, 3, 209–233.

KUANG, R., IE, E., WANG, K., WANG, K., SIDDIQI, M., FREUND, Y., AND LESLIE, C. 2005. Profile-based string kernels for remote homology detection and motif extraction. *Journal of Bioinformatics and Computational Biology 3*, 3, 527–550.

KUANG, R., WESTON, J., NOBLE, W. S., AND LESLIE, C. 2005. Motif-based protein ranking by network propagation. *Bioinformatics 21*, 19, 3711–3718.

KUNIK, V., SOLAN, Z., EDELMAN, S., RUPPIN, E., AND HORN, D. 2005. Motif extraction and protein classification. In *Proc. Computational Systems Bioinformatics (CSB)*. 80–85.

KUNKEL, B. N. AND BROOKS, D. M. 2002. Cross talk between signaling pathways in pathogen defense. *Curr Opin Plant Biol. 5*, 4, 325–331.

KURAMOCHI, M. AND KARYPIS, G. 2001. Gene classification using expression profiles: A feasibility study. In *Proc. BIBE*. 191–200.

KURAMOCHI, M. AND KARYPIS, G. 2004. An efficient algorithm for discovering frequent subgraphs. *IEEE Trans. Knowl. Data Eng. 16*, 9, 1038–1051.

LAEGREID, A., HVIDSTEN, T. R., MIDELFART, H., KOMOROWSKI, J., AND SANDVIK, A. K. 2003. Predicting gene ontology biological process from temporal gene expression patterns. *Genome Research 13,* 5, 965–979.

LAN, N., JANSEN, R., AND GERSTEIN, M. 2002. Towards a systematic definition of protein function that scales to the genome level: Defining function in terms of interactions. *Proc. IEEE 90,* 12, 1848–1858.

LANCKRIET, G. R. G., BIE, T. D., CRISTIANINI, N., JORDAN, M. I., AND NOBLE, W. S. 2004. A statistical framework for genomic data fusion. *Bioinformatics 20,* 16, 2626–2635.

LANCKRIET, G. R. G., DENG, M., CRISTIANINI, N., JORDAN, M. I., AND NOBLE, W. S. 2004. Kernel-based data fusion and its application to protein function prediction in yeast. In *Proc. Pacific Symposium on Biocomputing (PSB)*. 300–311.

LASKOWSKI, R. A. 1995. SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J Mol Graph. 13,* 5, 323–330.

LASKOWSKI, R. A., WATSON, J. D., AND THORNTON, J. M. 2003. From protein structure to biochemical function? *J Struct Funct Genomics 4,* 2–3, 167–177.

LASKOWSKI, R. A., WATSON, J. D., AND THORNTON, J. M. 2005. Profunc: a server for predicting protein function from 3D structure. *Nucleic Acids Research 33,* Web Server Issue, W89–W93.

LEE, H., TU, Z., DENG, M., SUN, F., AND CHEN, T. 2006. Diffusion kernel based logistic regression models for protein function prediction. *OMICS: Integrative Biology 10,* 1, 40–55.

LEE, I., DATE, S. V., ADAI, A. T., AND MARCOTTE, E. M. 2004. A probabilistic functional network of yeast genes. *Science 306,* 5701, 1555–1558.

LEE, M.-L. T. 2004. *Analysis of Microarray Gene Expression Data*. Kluwer Academic Publishers.

LEE, T. I., RINALDI, N. J., ROBERT, F., ODOM, D. T., BAR-JOSEPH, Z., GERBER, G. K., HANNETT, N. M., HARBISON, C. T., CRAIG M. THOMPSON, D. I. S., ZEITLINGER, J., JENNINGS, E. G., MURRAY, H. L., GORDON, D. B., REN, B., WYRICK, J. J., TAGNE, J.-B., VOLKERT, T. L., FRAENKEL, E., GIFFORD, D. K., AND YOUNG, R. A. 2002. Transcriptional regulatory networks in *Saccharomyces cerevisiae. Science 298,* 5554, 799–804.

LEGRAIN, P., WOJCIK, J., AND GAUTHIER, J.-M. 2001. Protein–protein interaction maps: a lead towards cellular functions. *Trends in Genetics 17,* 6, 346–352.

LEHNER, B., SEMPLE, J. I., BROWN, S. E., COUNSELL, D., CAMPBELL, R. D., AND SANDERSON, C. M. 2004. Analysis of a high-throughput yeast two-hybrid system and its use to predict the function of intracellular proteins encoded within the human MHC class III region. *Genomics 83,* 1, 153–167.

LELANDAIS, G., CROM, S. L., DEVAUX, F., VIALETTE, S., CHURCH, G. M., JACQ, C., AND MARC, P. 2004. yMGV: a cross-species expression data mining tool. *Nucleic Acids Research 32,* Database issue, D323–D335.

LEONE, M. AND PAGNANI, A. 2005. Predicting protein functions with message passing algorithms. *Bioinformatics 21,* 2, 239–247.

LETOVSKY, S. AND KASIF, S. 2003. Predicting protein function from protein/protein interaction data: a probabilistic approach. *Bioinformatics 19,* Suppl 1, i197–i204.

LETSCHE, T. A. AND BERRY, M. W. 1997. Large-scale information retrieval with latent semantic indexing. *Inf. Sci. 100,* 1-4, 105–137.

LEVY, E. D., OUZOUNIS, C. A., GILKS, W. R., AND AUDIT, B. 2005. Probabilistic annotation of protein sequences based on functional classifications. *BMC Bioinformatics 6,* 302.

LEWIS, S. E. 2005. Gene Ontology: looking backwards and forwards. *Genome Biology 6,* 1, 103.

LI, J., HALGAMUGE, S. K., KELLS, C. I., AND TANG, S.-L. 2007. Gene function prediction based on genomic context clustering and discriminative learning: an application to bacteriophages. *BMC Bioinformatics 8,* Suppl 4, S6.

LI, T., ZHU, S., LI, Q., AND OGIHARA, M. 2003. Gene functional classification by semi-supervised learning from heterogeneous data. In *Proc. ACM Symposium on Applied computing (SAC)*. 78–82.

LIBERLES, D. A., THORN, A., VON HEIJNE, G., AND ELOFSSON, A. 2002. The use of phylogenetic profiles for gene predictions. *Current Genomics 3,* 3, 131–137.

LIN, C., JIANG, D., AND ZHANG, A. 2006. Prediction of protein function using common-neighbors in protein-protein interaction networks. In *BIBE '06: Proceedings of the Sixth IEEE Symposium on BionInformatics and BioEngineering (BIBE'06)*. 251–260.

LIN, D. 1998. An information-theoretic definition of similarity. In *Proc. International Conference on Machine Learning*. 296–304.

LIU, A. H. AND CALIFANO, A. 2001. Functional classification of proteins by pattern discovery and top-down clustering of primary sequences. *IBM Systems Journal 40,* 2, 379–393.

LIU, J., WANG, W., AND YANG, J. 2004. Gene ontology friendly biclustering of expression profiles. In *Proc. IEEE Computational Systems Bioinformatics Conference (CSB)*. 436–447.

LORD, P. W. ET AL. 2003. Investigating semantic similarity measures across the gene ontology: The relationship between sequence and annotation. *Bioinformatics 19,* 10, 1275–1283.

LORD, P. W., STEVENS, R. D., BRASS, A., AND GOBLE, C. A. 2003. Semantic similarity measures as tools for exploring the gene ontology. In *Pacific Symposium on Biocomputing*. 601–612.

LUSCOMBE, N. M., AUSTIN, S. E., BERMAN, H. M., AND THORNTON, J. M. 2000. An overview of the structures of protein-DNA complexes. *Genome Biology 1,* 1, REVIEWS001.

LPEZ, G., ROJAS, A., TRESS, M., AND VALENCIA, A. 2007. Assessment of predictions submitted for the CASP7 function prediction category. *Proteins: Structure, Function, and Bioinformatics 69,* S8, 165–174.

MACKAY, D. 1992. The evidence framework applied to classification networks. *Neural Computation 4,* 5, 720–736.

MADEIRA, S. C. AND OLIVEIRA, A. L. 2004. Biclustering algorithms for biological data analysis: A survey. *IEEE/ACM Trans. Comput. Biol. Bioinformatics 1,* 1, 24–45.

MANN, M. AND JENSEN, O. N. 2003. Proteomic analysis of post-translational modifications. *Nat Biotechnol. 21,* 3, 255–261.

MARCHLER-BAUER, A., ANDERSON, J. B., CHERUKURI, P. F., DEWEESE-SCOTT, C., GEER, L. Y., GWADZ, M., HE, S., HURWITZ, D. I., JACKSON, J. D., KE, Z., LANCZYCKI, C. J., LIEBERT, C. A., LIU, C., LU, F., MARCHLER, G. H., MULLOKANDOV, M., SHOEMAKER, B. A., SIMONYAN, V., SONG, J. S., THIESSEN, P. A., YAMASHITA, R. A., YIN, J. J., ZHANG, D., AND BRYANT, S. H. 2005. CDD: a Conserved Domain Database for protein classification. *Nucleic Acids Research 33,* Database issue, D192–D196.

MARCOTTE, C. J. V. AND MARCOTTE, E. M. 2002. Predicting functional linkages from gene fusions with confidence. *Applied Bioinformatics 1,* 2, 93–100.

MARCOTTE, E. M. 2000. Computational genetics: finding protein function by nonhomology methods. *Curr Opin Struct Biol. 10,* 3, 359–365.

MARCOTTE, E. M. 2004. Practical computational approaches to inferring protein function. *Drug Discovery Today: BIOSILICO 2,* 1, 24–29.

MARCOTTE, E. M., PELLEGRINI, M., NG, H.-L., RICE, D. W., YEATES, T. O., AND EISENBERG, D. 1999. Detecting protein function and protein-protein interactions from genome sequences. *Science 285,* 5428, 751–753.

MARCOTTE, E. M., PELLEGRINI, M., THOMPSON, M. J., YEATES, T. O., AND EISENBERG, D. 1999. A combined algorithm for genome-wide prediction of protein function. *Nature 402,* 6757, 83–86.

MARCOTTE, E. M., XENARIOS, I., VAN DER BLIEK, A. M., AND EISENBERG, D. 2000. Localizing proteins in the cell from their phylogenetic profiles. *Proc Natl Acad Sci U.S.A. 97,* 22, 12115–12120.

MARTIN, A., ORENGO, C., HUTCHINSON, E., JONES, S., KARMIRANTZOU, M., LASKOWSKI, R., MITCHELL, J., TARONI, C., AND THORNTON, J. 1998. Protein folds and functions. *Structure 6,* 7, 875–884.

MARTIN, D. M. A., BERRIMAN, M., AND BARTON, G. J. 2004. GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes. *BMC Bioinformatics 5,* 178.

MASSJOUNI, N., RIVERA, C. G., AND MURALI, T. M. 2006. VIRGO: computational prediction of gene functions. *Nucleic Acids Research 34,* Web Server issue, W340–W344.

MATEOS, A., DOPAZO, J., JANSEN, R., TU, Y., GERSTEIN, M., AND STOLOVITZKY, G. 2002. Systematic learning of gene functional classes from dna array expression data by using multilayer perceptrons. *Genome Research 12,* 11, 1703–1715.

MATOBA, R., SAITO, S., UENO, N., MARUYAMA, C., MATSUBARA, K., AND KATO, K. 2000. Gene expression profiling of mouse postnatal cerebellar development. *Physiol. Genomics 4,* 2, 155–164.

MATTHEWS, L. R., VAGLIO, P., REBOUL, J., GE, H., DAVIS, B. P., GARRELS, J., VINCENT, S., AND VIDAL, M. 2001. Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "Interologs". *Genome Research 11,* 12, 2120–2126.

MCDERMOTT, J., BUMGARNER, R., AND SAMUDRALA, R. 2005. Functional annotation from predicted protein interaction networks. *Bioinformatics 21,* 15, 3217–3226.

MCDERMOTT, J. AND SAMUDRALA, R. 2005. Bioverse: functional and structural and contextual annotation of proteins and proteomes. *Nucleic Acids Research 31,* 13, 3736–3737.

MELLOR, J. C., YANAI, I., CLODFELTER, K. H., MINTSERIS, J., AND DELISI, C. 2002. Predictome: a database of putative functional links between proteins. *Nucleic Acids Research 30,* 1, 306–309.

MEWES, H. W., FRISHMAN, D., GULDENER, U., MANNHAUPT, G., MAYER, K., MOKREJS, M., MORGENSTERN, B., MUNSTERKOTTER1, M., RUDD, S., AND WEIL, B. 2002. MIPS: a database for genomes and protein sequences. *Nucleic Acids Research 30,* 1, 31–34.

MEZARD, M. AND PARISI, G. 2001. The Bethe lattice spin glass revisited. *European Physical Journal B 20*, 217–233.

MI, H., LAZAREVA-ULITSKY, B., LOO, R., KEJARIWAL, A., VANDERGRIFF, J., RABKIN, S., GUO, N., MURUGANUJAN, A., DOREMIEUX, O., CAMPBELL, M. J., KITANO, H., AND THOMAS, P. D. 2005. The PANTHER database of protein families and subfamilies and functions and pathways. *Nucleic Acids Research 33*, Database Issue, D284–D288.

MIDELFART, H., LAEGREID, A., AND KOMOROWSKI, J. 2001. Classification of gene expression data in an ontology. In *Proc. ISMDA*. 186–194.

MITCHELL, J. A., ARONSON, A. R., MORK, J. G., FORK, L. C., HUMPHREY, S. M., AND WARD, J. M. 2003. Gene indexing: Characterization and analysis of NLM's GeneRIFs. In *Proc. AMIA Symposium*. 460–464.

MIYAMOTO, S. 1990. *Fuzzy Sets in Information and Cluster Analysis: Theory and Decision Library*. Kluwer Academic Publishers.

MOLLER-LEVET, C. S., CHO, K.-H., YIN, H., AND WOLKENHAUER, O. 2003. Clustering of gene expression time-series data. Tech. rep., Department of Computer Science, University of Rostock, Germany.

MORIYAMA, E. N. AND KIM, J. 2006. Protein family classification with discriminant function analysis. In *Genome Exploitation: Data Mining the Genome*, J. P. Gustafson, Ed. Springer.

MOSZER, I., JONES, L. M., MOREIRA, S., FABRY, C., AND DANCHIN, A. 2002. SubtiList: the reference database for the bacillus subtilis genome. *Nucleic Acids Research 30,* 1, 62–65.

MOULT, J. 2005. A decade of CASP: progress and bottlenecks and prognosis in protein structure prediction. *Curr Opin Struct Biol. 15,* 3, 285–289.

MOULT, J. AND MELAMUD, E. 2000. From fold to function. *Curr Opin Struct Biol. 10,* 3, 384–389.

MUKHERJEE, S. 2003. Classifying microarray data using support vector machines. In *A Practical Approach to Microarray Data Analysis*, D. P. Berrar, W. Dubitzky, and M. Granzow, Eds. Kluwer Academic Publishers, Chapter 9, 166–185.

MYERS, C. L. ET AL. 2006. Finding function: evaluation methods for functional genomic data. *BMC Genomics 7*, 187.

NABIEVA, E., JIM, K., AGARWAL, A., CHAZELLE, B., AND SINGH, M. 2005. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics 21,* Suppl. 1, i1–i9.

NAIR, R. AND ROST, B. 2004. Annotating protein function through lexical analysis. *AI Magazine 25,* 1, 45–56.

NAJMANOVICH, R. J., TORRANCE, J. W., AND THORNTON, J. M. 2005. Prediction of protein function from structure: insights from methods for the detection of local structural similarities. *Biotechniques 38,* 6, 847,849,851.

NAKAMURA, H. 1996. Roles of electrostatic interaction in proteins. *Q Rev Biophys. 29,* 1, 1–90.

NARIAI, N., KOLACZYK, E. D., AND KASIF, S. 2007. Probabilistic protein function prediction from heterogeneous genome-wide data. *PLoS ONE 2,* 3, e337.

NARRA, K. AND LIAO, L. 2005. Use of extended phylogenetic profiles with E-values and support vector machines for protein family classification. *International Journal of Computer and Information Sciences 6,* 1.

NATURE EDITORIAL. 1999. Wanted: a new order in protein nomenclature. *Nature 401,* 6752, 411.

NEAL, S. J., GIBSON, M. L., SO, A. K.-C., AND WESTWOOD, J. T. 2003. Construction of a cDNA-based microarray for Drosophila melanogaster: a comparison of gene transcription profiles from SL2 and Kc167 cells. *Genome 46*, 5, 879–892.

NG, S.-K., TAN, S.-H., AND SUNDARARAJAN, V. 2003. On combining multiple microarray studies for improved functional classification by whole-dataset feature selection. *Genome Informatics 14*, 44–53.

NG, S.-K., ZHU, Z., AND ONG, Y.-S. 2004. Whole-genome functional classification of genes by latent semantic analysis on microarray data. In *Proc. Second Asia-Pacific Conference on Bioinformatics*. 123–129.

NGUYEN, D. V., ARPAT, A. B., WANG, N., AND CARROLL, R. J. 2002. DNA microarray experiments: biological and technological aspects. *Biometrics 58,* 4, 701–717.

NIELSEN, M. A. AND CHUANG, I. L. 2000. *Quantum Computation and Quantum Information*. Cambridge University Press.

NIGAM, K., LAFFERTY, J., AND MCCALLUM, A. 1999. Using maximum entropy for text classification. In *IJCAI Workshop on Machine Learning for Information Filtering*. 61–67.

NOBLE, W. S., KUANG, R., LESLIE, C., AND WESTON, J. 2005. Identifying remote protein homologs by network propagation. *The FEBS Journal 272,* 20, 5119–5128.

NOVRE, N. L. AND CHANGEUX, J.-P. 2001. LGICdb: the ligand-gated ion channel database. *Nucleic Acids Research 29,* 1, 294–295.

ODA, M., FURUKAWA, K., OGATA, K., SARAI, A., AND NAKAMURA, H. 1998. Thermodynamics of specific and non-specific DNA binding by the c-Myb DNA-binding domain. *J Mol Biol. 276,* 3, 571–590.

OKADA, K., KANAYA, S., AND ASAI, K. 2005. Accurate extraction of functional associations between proteins based on common interaction partners and common domains. *Bioinformatics 21,* 9, 2043–2048.

OLIVER, S. 1996. A network approach to the systematic analysis of yeast gene function. *Trends in Genetics 12,* 7, 241–242.

ONG, C. S., SMOLA, A. J., AND WILLIAMSON, R. C. 2005. Learning the kernel with hyperkernels. *J. Mach. Learn. Res. 6,* 1043–1071.

ORENGO, C. A., BRAY, J. E., BUCHAN, D. W. A., HARRISON, A., LEE, D., PEARL, F. M. G., SILLITOE, I., TODD, A. E., AND THORNTON, J. M. 2002. The CATH protein family database: a resource for structural and functional annotation of genomes. *Proteomics 2,* 1, 11–21.

ORENGO, C. A. AND TAYLOR, W. R. 1996. SSAP: sequential structure alignment program for protein structure comparison. *Methods Enzymol. 266,* 617–635.

ORENGO, C. A., TODD, A. E., AND THORNTON, J. M. 1999. From protein structure to function. *Curr Opin Struct Biol. 9,* 3, 374–382.

O'SULLIVAN, O., SUHRE, K., ABERGEL, C., HIGGINS, D. G., AND NOTREDAME, C. 2004. 3DCoffee: combining protein sequences and structures within multiple sequence alignments. *J Mol Biol. 340,* 2, 385–395.

OUALI, M. AND KING, R. D. 2000. Cascaded multiple classifiers for secondary structure prediction. *Protein Science 9,* 6, 1162–1176.

OUZOUNIS, C. A., COULSON, R. M. R., ENRIGHT, A. J., KUNIN, V., AND PEREIRA-LEAL, J. B. 2003. Classification schemes for protein structure and function. *Nat Rev Genet. 4,* 7, 508–519.

OVERBEEK, R., FONSTEIN, M., D'SOUZA, M., PUSCH, G. D., AND MALTSEV, N. 1999a. Use of contiguity on the chromosome to predict functional coupling. *In Silico Biology 1,* 2, 93–108.

OVERBEEK, R., FONSTEIN, M., D'SOUZA, M., PUSCH, G. D., AND MALTSEV, N. 1999b. The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci U.S.A. 96,* 6, 2896–2901.

PAGEL, P., KOVAC, S., OESTERHELD, M., BRAUNER, B., DUNGER-KALTENBACH, I., FRISHMAN, G., MONTRONE, C., MARK, P., STUMPFLEN, V., MEWES, H.-W., RUEPP, A., AND FRISHMAN, D. 2005. The MIPS mammalian protein–protein interaction database. *Bioinformatics 21,* 6, 832–834.

PAL, D. AND EISENBERG, D. 2005. Inference of protein function from protein structure. *Structure 13,* 1, 121–130.

PAN, W. 2006. Incorporating gene functions as priors in model-based clustering of microarray gene expression data. *Bioinformatics 22,* 7, 795–801.

PANDEY, G. AND KUMAR, V. 2007. Incorporating functional inter-relationships into algorithms for protein function prediction. In *ISMB/ECCB Special Interest Group meeting on Automated Function Prediction*.

PANDEY, G., STEINBACH, M., GUPTA, R., GARG, T., AND KUMAR, V. 2007. Association analysis-based transformations for protein interaction networks: a function prediction case study. In *KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. 540–549.

PARKINSON, H., SARKANS, U., SHOJATALAB, M., ABEYGUNAWARDENA, N., CONTRINO, S., COULSON, R., FARNE, A., LARA, G. G., HOLLOWAY, E., KAPUSHESKY, M., LILJA, P., MUKHERJEE, G., OEZCIMEN, A., RAYNER, T., ROCCA-SERRA, P., SHARMA, A., SANSONE, S., AND BRAZMA, A. 2005. ArrayExpress–a public repository for microarray gene expression data at the EBI. *Nucleic Acids Research 33,* Database issue, D553–D555.

PASQUIER, C., PROMPONAS, V. J., AND HAMODRAKAS, S. J. 2001. PRED-CLASS: cascading neural networks for generalized protein classification and genome-wide applications. *Proteins 44,* 3, 361–369.

PAVLIDIS, P., WESTON, J., CAI, J., AND GRUNDY, W. N. 2002. Learning gene functional classifications from multiple data types. *J Comput Biol. 9,* 2, 401–411.

PAZOS, F., HELMER-CITTERICH, M., AUSIELLO, G., AND VALENCIA, A. 1997. Correlated mutations contain information about protein-protein interaction. *J Mol Biol. 271,* 4, 511–523.

PAZOS, F. AND STERNBERG, M. J. E. 2004. Automated prediction of protein function and detection of functional sites from structure. *Proc Natl Acad Sci U.S.A. 101,* 41, 14754–14759.

PAZOS, F. AND VALENCIA, A. 2001. Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Engineering 14,* 9, 609–614.

PEARSON, W. R. AND LIPMAN, D. J. 1988. Improved tools for biological sequence comparison. *Proc Natl Acad Sci U.S.A. 85,* 8, 2444–2448.

PELLEGRINI, M., MARCOTTE, E. M., THOMPSON, M. J., EISENBERG, D., AND YEATES, T. O. 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U.S.A. 96,* 8, 4285–4288.

PELLEGRINI-CALACE, M., SORO, S., AND TRAMONTANO, A. 2006. Revisiting the prediction of protein function at CASP6. *FEBS Journal 273,* 13, 2977–2983.

PEREIRA-LEAL, J. B., ENRIGHT, A. J., AND OUZOUNIS, C. A. 2003. Detection of functional modules from protein interaction networks. *Proteins 54,* 1, 49–57.

PEREZ, A., RODRIGUEZ, A., TRELLES, O., AND THODE, G. 2002. A computational strategy for protein function assignment which addresses the multidomain problem. *Comp. Funct. Gen. 3,* 5, 423–440.

PEREZ, A. J., PEREZ-IRATXETA, C., BORK, P., THODE, G., AND ANDRADE, M. A. 2004. Gene annotation from scientific literature using mappings between keyword systems. *Bioinformatics 20,* 13, 2084–2091.

PILPEL, Y., SUDARSANAM, P., AND CHURCH, G. M. 2001. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nature Genetics 29,* 2, 153–159.

POYTON, R. O. AND MCEWEN, J. E. 1996. Crosstalk between nuclear and mitochondrial genomes. *Annu Rev Biochem. 65,* 563–607.

PRLIC, A., DOMINGUES, F. S., LACKNER, P., AND SIPPL, M. J. 2004. WILMAautomated annotation of protein sequences. *Bioinformatics 20,* 1, 127–128.

PUIG, O., CASPARY, F., RIGAUT, G., RUTZ, B., BOUVERET, E., BRAGADO-NILSSON, E., WILM, M., AND SERAPHIN, B. 2001. The Tandem Affinity Purification (TAP) method: a general procedure of protein complex purification. *Methods 24,* 3, 218–229.

QIAN, B. AND GOLDSTEIN, R. A. 2003. Detecting distant homologs using phylogenetic tree-based HMMs. *Proteins 52,* 3, 446–453.

QIAN, B., SOYER, O. S., NEUBIG, R. R., AND GOLDSTEIN, R. A. 2003. Depicting a protein's two faces: GPCR classification by phylogenetic tree-based HMMs. *FEBS Letters 554,* 1, 95–99.

QUACKENBUSH, J. 2002. Microarray data normalization and transformation. *Nature Genetics 32,* 496–501.

QUINLAN, J. R. 1993. *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc.

RAIN, J.-C., SELIG, L., REUSE, H. D., BATTAGLIA, V., REVERDY, C., SIMON, S., LENZEN, G., PETEL, F., WOJCIK, J., SCHACHTER, V., CHEMAMA, Y., LABIGNE2, A., AND LEGRAIN, P. 2001. The protein–protein interaction map of *Helicobacter pylori*. *Nature 409,* 6817, 211–215.

RANGWALA, H., DERONNE, K., AND KARYPIS, G. 2006. Protein structure prediction using string kernels. Tech. Rep. 06-005, Department of Computer Science and Engineering, University of Minnesota.

RANGWALA, H. AND KARYPIS, G. 2005. Profile-based direct kernels for remote homology detection and fold recognition. *Bioinformatics 21,* 23, 4239–4247.

RASTAN, S. AND BEELEY, L. J. 1997. Functional genomics: going forwards from the databases. *Curr Opin Genet Dev. 7,* 6, 777–783.

RAWLINGS, N. D. AND BARRETT, A. J. 1999. MEROPS: the peptidase database. *Nucleic Acids Research 27,* 1, 325–331.

RAY, S. AND CRAVEN, M. 2005. Learning statistical models for annotating proteins with function information using biomedical text. *BMC Bioinformatics 6,* Suppl 1, S18.

RAYCHAUDHARI, S., CHANG, J. T., SUTPHIN, P. D., AND ALTMAN, R. B. 2002. Associating genes with gene ontology codes using a maximum entropy analysis of biomedical literature. *Genome Research 12,* 1, 203–214.

RENNER, A. AND ASZODI, A. 2000. High-throughput functional annotation of novel gene products using document clustering. In *Proc. Pac. Symp. Biocomputing*. 54–68.

RICE, S. B., NENADIC, G., AND STAPLEY, B. I. 2005. Mining protein function from text using term-based support vector machines. *BMC Bioinformatics 6,* Suppl 1, S22.

RIFKIN, R. AND KLAUTAU, A. 2004. In defense of one-vs-all classification. *J. Mach. Learn. Res. 5*, 101–141.

RIGOUTSOS, I. AND FLORATOS, A. 1998. Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm. *Bioinformatics 14,* 1, 55–67.

RIGOUTSOS, I., FLORATOS, A., OUZOUNIS, C., GAO, Y., AND PARIDA, L. 1999. Dictionary building via unsupervised hierarchical motif discovery in the sequence space of natural proteins. *Proteins 37,* 2, 264–277.

RIGOUTSOS, I., HUYNH, T., FLORATOS, A., PARIDA, L., AND PLATT, D. 2002. Dictionary-driven protein annotation. *Nucleic Acids Research 30,* 17, 3901–3916.

RIJSBERGEN, C. J. V. 1979. *Information Retrieval*. Butterworth-Heinemann.

RILEY, M. 1998. Systems for categorizing functions of gene products. *Curr Opin Struct Biol. 8*, 3, 388–392.

RILEY, M. L., SCHMIDT, T., WAGNER, C., MEWES, H.-W., AND FRISHMAN, D. 2005. The PEDANT genome database in 2005. *Nucleic Acids Research 33,* Database issue, D308–D310.

RISON, S. C. G., HODGMAN, T. C., AND THORNTON, J. M. 2000. Comparison of functional annotation schemes for genomes. *Funct Integr Genomics 1,* 1, 56–69.

RIVES, A. W. AND GALITSKI, T. 2003. Modular organization of cellular networks. *Proc Natl Acad Sci U.S.A. 100,* 3, 1128–1133.

ROBERTS, R. J. 2004. Identifying protein functiona call for community action. *PLoS Biology 2,* 3, 293–294.

ROCH, K. G. L., ZHOU, Y., BLAIR, P. L., GRAINGER, M., MOCH, J. K., HAYNES, J. D., LA VEGA, P. D., HOLDER, A. A., BATALOV, S., CARUCCI, D. J., AND WINZELER, E. A. 2003. Discovery of gene function by expression profiling of the malaria parasite life cycle. *Science 301,* 5639, 1503–1508.

RON, D., SINGER, Y., AND TISHBY, N. 1996. The power of amnesia: Learning probabilistic automata with variable memory length. *Machine Learning 25,* 2-3, 117–149.

ROST, B. 1996. PHD: predicting one-dimensional protein structure by profile-based neural networks. *Methods Enzymol. 266,* 525–539.

ROST, B., LIU, J., NAIR, R., WRZESZCZYNSKI, K. O., AND OFRAN, Y. 2003. Automatic prediction of protein function. *Cell Mol Life Sci. 60,* 12, 2637–2650.

ROST, B., YACHDAV, G., AND LIU, J. 2003. The PredictProtein server. *Nucleic Acids Research 32,* Web Server Issue, W321–W326.

RUBINSTEIN, R. AND SIMON, I. 2005. MILANO–custom annotation of microarray results using automatic literature searches. *BMC Bioinformatics 6,* 1, 12.

RUEPP, A., ZOLLNER, A., MAIER, D., ALBERMANN, K., HANI, J., MOKREJS, M., TETKO, I., GULDENER, U., MANNHAUPT, G., MUNSTERKOTTER, M., AND MEWES, H. W. 2004. The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Research 32,* 18, 5539–5545.

RUNG, J., SCHLITT, T., BRAZMA, A., FREIVALDS, K., AND VILO, J. 2002. Building and analysing genome-wide gene disruption networks. *Bioinformatics 18,* Suppl 2, S202–S210.

SAFRAN, M., CHALIFA-CASPI, V., SHMUELI, O., OLENDER, T., LAPIDOT, M., ROSEN, N., SHMOISH, M., PETER, Y., GLUSMAN, G., FELDMESSER, E., ADATO, A., PETER, I., KHEN, M., ATAROT, T., GRONER, Y., AND LANCET, D. 2003. Human gene-centric databases at the Weizmann Institute of Science: GeneCards and UDB and CroW 21 and HORDE. *Nucleic Acids Research 31,* 1, 142–146.

SAIER JR., M. H. 2000. A functional-phylogenetic classification system for transmembrane solute transporters. *Microbiol Mol Biol Rev. 64,* 2, 354–411.

SALAMOV, A. A. AND SOLOVYEV, V. V. 1995. Prediction of protein secondary structure by combinin nearest-neighbor algorithms and multiple sequence alignment. *J Mol Biol. 247*, 11–15.

SALGADO, H., MORENO-HAGELSIEB, G., SMITH, T. F., AND COLLADO-VIDESDAGGER, J. 2000. Operons in *Escherichia coli*: Genomic analyses and predictions. *PNAS 97,* 12, 6652–6657.

SALWINSKI, L. AND EISENBERG, D. 2003. Computational methods of analysis of protein-protein interactions. *Curr Opin Struct Biology 13,* 3, 377–382.

SALWINSKI, L., MILLER, C. S., SMITH, A. J., PETTIT, F. K., BOWIE, J. U., AND EISENBERG, D. 2004. The Database of Interacting Proteins: 2004 update. *Nucleic Acids Research 32,* Database issue, D449–D451.

SAMANTA, M. P. AND LIANG, S. 2003. Predicting protein functions from redundancies in large-scale protein interaction networks. *Proc Natl Acad Sci U.S.A. 100*, 22, 12579–12583.

SASSON, O., KAPLAN, N., AND LINIAL, M. 2006. Functional annotation prediction: All for one and one for all. *Protein Science 15*, 1557–1562.

SCHLITT, T., PALIN, K., RUNG, J., DIETMANN, S., LAPPE, M., UKKONEN, E., AND BRAZMA, A. 2003. From gene networks to gene function. *Genome Research 13*, 12, 2568–2576.

SCHOMBURG, I., CHANG, A., EBELING, C., GREMSE, M., HELDT, C., HUHN, G., AND SCHOMBURG, D. 2004. BRENDA and the enzyme database: updates and major new developments. *Nucleic Acids Research 32*, Database issue, D431–D433.

SCHUG, J., DISKIN, S., MAZZARELLI, J., BRUNK, B. P., AND JR., C. J. S. 2002. Predicting gene ontology functions from ProDom and CDD protein domains. *Genome Research 12*, 4, 648–655.

SCHULZ, G. E. AND SCHIRMER, R. H. 1996. Principles of protein structure.

SCHWIKOWSKI, B., UETZ, P., AND FIELDS, S. 2000. A network of protein-protein interactions in yeast. *Nature Biotechnology 18*, 12, 1257–1261.

SCORDIS, P., FLOWER, D., AND ATTWOOD, T. 1999. FingerPRINTScan: intelligent searching of the PRINTS motif database. *Bioinformatics 15*, 10, 799–806.

SECKO, D. 2007. A monk's flourishing garden: the basics of molecular biology explained. *The Science Creative Quarterly 3*.

SEKI, K. AND MOSTAFA, J. 2003. A probabilistic model for identifying protein names and their name boundaries. In *CSB '03: Proceedings of the IEEE Computer Society Conference on Bioinformatics*. 251.

SEKI, K. AND MOSTAFA, J. 2004. Toward database curation support in biology: Automated gene function identification from texts. Tech. Rep. LAIR04-02, Laboratory for Applied Informatics Research and Indiana University.

SERRES, M. H., GOSWAMI, S., AND RILEY, M. 2004. GenProtEC: an updated and improved analysis of functions of Escherichia coli K-12 proteins. *Nucleic Acids Research 32*, Database issue, D300–D302.

SERRES, M. H. AND RILEY, M. 2000. MultiFun, a multifunctional classification scheme for Escherichia coli K-12 gene products. *Microb Comp Genomics 5*, 4, 205–222.

SERVANT, F., BRU, C., CARRERE, S., COURCELLE, E., GOUZY, J., PEYRUC, D., AND KAHN, D. 2002. ProDom: automated clustering of homologous domains. *Brief Bioinform. 3*, 3, 246–251.

SESE, J., NIKAIDOU, H., KAWAMOTO, S., MINESAKI, Y., MORISHITA, S., AND OKUBO, K. 2001. BodyMap incorporated PCR-based expression profiling data and a gene ranking system. *Nucleic Acids Research 29*, 1, 156–158.

SESHASAYEE, A. S. N. AND BABU, M. M. 2005. Contextual inference of protein function. In *Encyclopaedia of Genetics and Genomics and Proteomics and Bioinformatics*, S. Subramaniam, Ed. John Wiley and Sons.

SEVILLA, J. L. ET AL. 2005. Correlation between gene expression and go semantic similarity. *IEEE/ACM Trans. Comput. Biol. Bioinformatics 2*, 4, 330–338.

SHAKHNOVICH, B. E. 2005. Improving the precision of the structure-function relationship by considering phylogenetic context. *PLoS Computational Biology 1*, 1, e9.

SHARAN, R., SUTHRAM, S., KELLEY, R. M., KUHN, T., MCCUINE, S., UETZ, P., SITTLER, T., KARP, R. M., AND IDEKER, T. 2006. Conserved patterns of protein interaction in multiple species. *Proc Natl Acad Sci U.S.A. 102*, 6, 1974–1979.

SHARAN, R., ULITSKY, I., AND SHAMIR, R. 2007. Network-based prediction of protein function. *Molecular Systems Biology 3*, 88.

SHATKAY, H., EDWARDS, S., WILBUR, W. J., AND BOGUSKI, M. 2000. Genes and themes and microarrays: Using information retrieval for large-scale gene analysis. In *Proc. ISMB*. 317–328.

SHATSKY, M., NUSSINOV, R., AND WOLFSON, H. J. 2004. A method for simultaneous alignment of multiple protein structures. *Proteins 56*, 1, 143–156.

SHAWE-TAYLOR, J. AND CRISTIANINI, N. 2004. *Kernel Methods for Pattern Analysis*. Cambridge University Press.

SHEN, L., GONG, J., CALDO, R. A., NETTLETON, D., COOK, D., WISE, R. P., AND DICKERSON, J. A. 2005. BarleyBase–an expression profiling database for plant genomics. *Nucleic Acids Research 33*, Database issue, D614–D618.

SHENDURE, J., MITRA, R. D., VARMA, C., AND CHURCH, G. M. 2004. Advanced sequencing technologies: methods and goals. *Nature Reviews Genetics 5*, 5, 335–344.

SHERLOCK, G., HERNANDEZ-BOUSSARD, T., KASARSKIS, A., BINKLEY, G., MATESE, J. C., DWIGHT, S. S., KALOPER, M., WENG, S., JIN, H., BALL, C. A., EISEN, M. B., SPELLMAN, P. T., BROWN, P. O., BOTSTEIN, D., AND CHERRY, J. M. 2001. The Stanford Microarray Database. *Nucleic Acids Research 29*, 1, 152–155.

SHINDYALOV, I. N. AND BOURNE, P. E. 1998. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Engineering 11*, 9, 739–747.

SHIRASU, K., LAHAYE, T., TAN, M.-W., ZHOU, F., AZEVEDO, C., AND SCHULZE-LEFERT, P. 1999. A novel class of eukaryotic zinc-binding proteins is required for disease resistance signaling in barley and development in c. elegans. *Cell 99*, 4, 355–366.

SICKMANN, A., MREYEN, M., AND MEYER, H. E. 2003. Mass spectrometry–a key technology in proteome research. *Adv Biochem Eng Biotechnol. 83*, 141–176.

SJOLANDER, K. 1997. Bayesian evolutionary tree estimation. In *Proc. Computing in the Genome Era*.

SJOLANDER, K. 2004. Phylogenomic inference of protein molecular function: advances and challenges. *Bioinformatics 20*, 2, 170–179.

SKOLNICK, J. AND FETROW, J. S. 2000. From genes to protein structure and function: novel applications of computational approaches in the genomic era. *Trends in Biotechnology 18*, 1, 34–39.

SKOLNICK, J., FETROW, J. S., AND KOLINSKI, A. 2000. Structural genomics and its importance for gene function analysis. *Nature Biotechnology 18*, 3, 283–287.

SLONIM, D. K. 2002. From patterns to pathways: gene expression data analysis comes of age. *Nature Genetics 32*, Suppl, 502–508.

SMITH, B. 2003. Ontology. In *Blackwell Guide to the Philosophy of Computing and Information*, L. Floridi, Ed. Blackwell Publishers.

SNEL, B., BORK, P., AND HUYNEN, M. A. 2002. The identification of functional modules from the genomic association of genes. *Proc Natl Acad Sci U.S.A. 99*, 9, 5890–5895.

SNEL, B., LEHMANN, G., BORK, P., AND HUYNEN, M. A. 2000. STRING: A web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Research 28*, 18, 3442–3444.

SNYDER, E. E. AND STORMO, G. D. 1995. Identification of protein coding regions in genomic DNA. *J Mol Biol. 248*, 1, 1–18.

SONNHAMMER, E. L., EDDY, S. R., AND DURBIN, R. 1997. Pfam: A comprehensive database of protein domain families based on seed alignments. *Proteins 28*, 3, 405–420.

SPELLMAN, P. T., SHERLOCK, G., ZHANG, D. M. Q., IYER, V. R., ANDERS, K., EISEN, M. B., BROWN, P. O., BOTSTEIN, D., AND FUTCHER, B. 1998. Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell 9*, 12, 3273–3297.

STAAB, S., DOMINGOS, P., MIKA, P., GOLBECK, J., DING, L., FININ, T. W., JOSHI, A., NOWAK, A., AND VALLACHER, R. R. 2005. Social networks applied. *IEEE Intelligent Systems 20*, 1, 80–93.

STOICA, E. AND HEARST, M. 2006. Predicting gene functions from text using a cross-species approach. In *Proc. Eleventh Pacific Symposium on Biocomputing (PSB)*. 88–99.

STRONG, M., MALLICK, P., PELLEGRINI, M., THOMPSON, M. J., AND EISENBERG, D. 2003. Inference of protein function and protein linkages in Mycobacterium tuberculosis based on prokaryotic genome organization: a combined computational approach. *Genome Biology 4*, 9, R59.

SUN, S., ZHAO, Y., JIAO, Y., YIN, Y., CAI, L., ZHANG, Y., LU, H., CHENA, R., AND BU, D. 2006. Faster and more accurate global protein function assignment from protein interaction networks using the MFGO algorithm. *FEBS Letters 580*, 7, 1891–1896.

SUZUKI, A., ANDO, T., YAMATO, I., AND MIYAZAKI, S. 2005. Fcanal: Structure based protein function prediction method. application to enzymes and binding proteins. *Chem-Bio Informatics Journal 5*, 3, 39–55.

SWIFT, S., TUCKER, A., VINCIOTTI, V., MARTIN, N., ORENGO, C., LIU, X., AND KELLAM, P. 2004. Consensus clustering and functional interpretation of gene-expression data. *Genome Biology 5*, 11, R94.

SYED, U. AND YONA, G. 2003. Using a mixture of probabilistic decision trees for direct prediction of protein function. In *Proc. Seventh annual international conference on Research in computational molecular biology (RECOMB)*. 289–300.

TAMAMES, J., OUZOUNIS, C., CASARI, G., SANDER, C., AND VALENCIA, A. 1998. EUCLID: automatic classification of proteins in functional classes by their database annotations. *Bioinformatics 14,* 6, 542–543.

TAN, P.-N., KUMAR, V., AND SRIVASTAVA, J. 2004. Selecting the right objective measure for association analysis. *Inf. Syst. 29,* 4, 293–313.

TAN, P.-N., STEINBACH, M., AND KUMAR, V. 2005. *Introduction to Data Mining.* Addison-Wesley.

TAO, Y., SAM, L., LI, J., FRIEDMAN, C., AND LUSSIER, Y. A. 2007. Information theory applied to the sparse gene ontology annotation network to predict novel gene function. *Bioinformatics 23,* 13, i529–i538.

TATUSOV, R. L., FEDOROVA, N. D., JACKSON, J. D., JACOBS, A. R., KIRYUTIN, B., KOONIN, E. V., KRYLOV, D. M., MAZUMDER, R., MEKHEDOV, S. L., NIKOLSKAYA, A. N., RAO, B. S., SMIRNOV, S., SVERDLOV, A. V., VASUDEVAN, S., WOLF, Y. I., YIN1, J. J., AND NATALE, D. A. 2003. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics 4,* 41.

TATUSOV, R. L., KOONIN, E. V., AND LIPMAN, D. J. 1997. A genomic perspective on protein families. *Science 278,* 5338, 631–637.

TATUSOVA, T. A., KARSCH-MIZRACHI, I., AND OSTELL, J. A. 1999. Complete genomes in WWW Entrez: data representation and analysis. *Bioinformatics 15,* 536–543.

TEICHMANN, S. A., CHOTHIA, C., AND GERSTEIN, M. 1999. Advances in structural genomics. *Curr Opin Struct Biol. 9,* 3, 390–399.

TEICHMANN, S. A. AND MITCHISON, G. 2000. Computing protein function. *Nature Biotechnology 18,* 1, 27.

TETKO, I. V., BRAUNER, B., DUNGER-KALTENBACH, I., FRISHMAN, G., MONTRONE, C., FOBO, G., RUEPP, A., ANTONOV, A. V., SURMELI, D., AND MEWES, H.-W. 2005. MIPS bacterial genomes functional annotation benchmark dataset. *Bioinformatics 21,* 10, 2520–2521.

THODE, G., GARCIA-RANEA, J. A., AND JIMENEZ, J. 1996. Search for ancient patterns in protein sequences. *J Mol Evol. 42,* 2, 224–233.

THOREN, A. 2000. The phylprom database – extending the use of phylogenetic profiles and their applications for membrane proteins. M.S. thesis, Stockholm Bioinformatics Center, Stockholm University.

THORNTON, J. M., ORENGO, C. A., TODD, A. E., AND PEARL, F. M. G. 1999. Protein folds and functions and evolution. *J Mol Biol. 293,* 2, 333–342.

THORNTON, J. M., TODD, A. E., MILBURN, D., BORKAKOTI, N., AND ORENGO, C. A. 2000. From structure to function: Approaches and limitations. *Nature Structural Biology 7,* Suppl, 991–994.

TIAN, W., ARAKAKI, A. K., AND SKOLNICK, J. 2004. EFICAz: a comprehensive approach for accurate genome-scale enzyme function inference. *Nucleic Acids Research 32,* 21, 6226–6239.

TROYANSKAYA, O. G., DOLINSKI, K., OWEN, A. B., ALTMAN, R. B., AND BOTSTEIN, D. 2003. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae). *Proc Natl Acad Sci U.S.A. 100,* 14, 8348–8353.

TSAI, C. J. AND NUSSINOV, R. 1996. Hydrophobic folding units at protein-protein interfaces: implications to protein folding and to protein-protein association. *Protein Science 6,* 7, 1426–1437.

TSUDA, K., KIN, T., AND ASAI, K. 2002. Marginalized kernels for biological sequences. *Bioinformatics 18,* Suppl 1, S268–S275.

TSUDA, K. AND NOBLE, W. S. 2004. Learning kernels from biological networks by maximizing entropy. *Bioinformatics 20,* Suppl. 1, i326–i333.

TSUDA, K., SHIN, H., AND SCHOLKOPF, B. 2005. Fast protein classification with multiple networks. *Bioinformatics 21,* Suppl 1, ii59–ii65.

UETZ, P., GIOT, L., CAGNEY, G., MANSFIELD, T. A., JUDSON, R. S., KNIGHT, J. R., 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., AND ROTHBERG, J. M. 2000. A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature 403,* 6770, 623–627.

ULITSKY, I. AND SHAMIR, R. 2007. Identification of functional modules using network topology and high-throughput data. *BMC Systems Biology 1,* 8.

VAN DE GOOR, T. A. 2005. A history of DNA microarrays. *Pharmaceutical Discovery.*

VAN NOORT, V., SNEL, B., AND HUYNEN, M. A. 2003. Predicting gene function by conserved co-expression. *TRENDS in Genetics 19,* 5, 238–242.

VAZQUEZ, A., FLAMMINI, A., MARITAN, A., AND VESPIGNANI, A. 2003. Global protein function prediction from protein–protein interaction networks. *Nat Biotechnology 21,* 6, 697–700.

VERT, J.-P. 2002. A tree kernel to analyze phylogenetic profiles. *Bioinformatics 18,* Suppl 1, S276–S284.

VERT, J.-P. AND KANEHISA, M. 2002. Graph-driven features extraction from microarray data using diffusion kernels and kernel cca. In *Proc. NIPS*. 1425–1432.

VINAYAGAM, A., DEL VAL, C., SCHUBERT, F., EILS, R., GLATTING, K.-H., SUHAI, S., AND KNIG, R. 2006. GOPET: A tool for automated predictions of Gene Ontology terms. *BMC Bioinformatics 7,* 161.

VLAHOVICEK, K., MURVAI, J., BARTA, E., AND PONGOR, S. 2002. The SBASE protein domain library and release 9.0: an online resource for protein domain identification. *Nucleic Acids Research 30,* 1, 273–275.

VOLLER, C. S. AND UETZ, P. 2004. The phox homology (PX) domain protein interaction network in yeast. *Mol Cell Proteomics 3,* 11, 1053–1064.

VON MERING, C., HUYNEN, M., JAEGGI, D., SCHMIDT, S., BORK, P., AND SNEL, B. 2003. STRING: a database of predicted functional associations between proteins. *Nucleic Acids Research 31,* 1, 258–261.

VON MERING, C., KRAUSE, R., SNEL, B., CORNELL, M., OLIVER, S. G., FIELDS, S., AND BORK, P. 2002. Comparative assessment of large–scale data sets of protein–protein interactions. *Nature 417,* 6887, 399–403.

WALKER, M. G., VOLKMUTH, W., SPRINZAK, E., HODGSON, D., AND KLINGLER, T. 1999. Prediction of gene function by genome-scale expression analysis: prostate cancer-associated genes. *Genome Research 9,* 12, 1198–1203.

WALLACE, A. C., BORKAKOTI, N., AND THORNTON, J. M. 1997. TESS: a geometric hashing algorithm for deriving 3d coordinate templates for searching structural databases. application to enzyme active sites. *Protein Science 6,* 11, 2308–2323.

WANG, C. AND SCOTT, S. D. 2005. New kernels for protein structural motif discovery and function classification. In *Proc. 22nd International Conference on Machine Learning (ICML)*. 940–947.

WANG, J. T. L., MA, Q., SHASHA, D., AND WU, C. H. 2001. New techniques for extracting features from protein sequences. *IBM Systems Journal 40,* 2, 426–441.

WANG, X., SCHROEDER, D., DOBBS, D., AND HONAVAR, V. G. 2003. Automated data-driven discovery of motif-based protein function classifiers. *Inf. Sci. 155,* 1-2, 1–18.

WASSERMAN, S. AND FAUST, K. 1994. *Social Network Analysis : Methods and Applications (Structural Analysis in the Social Sciences)*. Structural Analysis in the Social Sciences. Cambridge University Press.

WEAVER, R. H. 2002. *Molecular Biology*. McGraw Hill.

WEBB, E., Ed. 1992. *Enzyme Nomenclature 1992*. Academic Press.

WEISS, S. M., INDURKHYA, N., ZHANG, T., AND DAMERAU, F. J. 2004. *Text Mining: Predictive Methods for Analyzing Unstructured Information*. Springer.

WEST, D. B. 2001. *Introduction to Graph Theory*. Prentice Hall.

WHISSTOCK, J. C. AND LESK, A. M. 2003. Prediction of protein function from protein sequence and structure. *Q Rev Biophys. 36,* 3, 307–340.

WILD, D. L. AND SAQI, M. A. S. 2004. Structural proteomics: Inferring function from protein structure. *Current Proteomics 1,* 1, 59–65.

WILLIAMS, C. K. I. 2002. On a connection between kernel PCA and metric multidimensional scaling. *Mach. Learn. 46,* 1-3, 11–19.

WILSON, C. A., KREYCHMAN, J., AND GERSTEIN, M. 2000. Assessing annotation transfer for genomics: quantifying the relations between protein sequence and structure and function through traditional and probabilistic scores. *J Mol Biol. 297,* 1, 233–249.

WOLFSON, H. J., SHATSKY, M., SCHNEIDMAN-DUHOVNY, D., DROR, O., SHULMAN-PELEG, A., MA, B., AND NUSSINOV, R. 2005. From structure to function: methods and applications. *Curr Protein Pept Sci. 6,* 2, 171–183.

WU, C., BERRY, M., SHIVAKUMAR, S., AND MCLARTY, J. 1995. Neural networks for full-scale protein sequence classification: Sequence encoding with singular value decomposition. *Machine Learning 21,* 1-2, 177–193.

WU, C., WHITSON, G., MCLARTY, J., ERMONGKONCHAI, A., AND CHANG, T. C. 1992. Protein classification artificial neural system. *Protein Science 1,* 5, 667–677.

WU, C. H., HUANG, H., NIKOLSKAYA, A., HU, Z., AND BARKER, W. C. 2004. The iProClass integrated database for protein functional analysis. *Comput Biol Chem. 28,* 1, 87–96.

WU, C. H., YEH, L.-S. L., HUANG, H., ARMINSKI, L., CASTRO-ALVEAR, J., CHEN, Y., HU, Z.-Z., LEDLEY, R. S., SUZEK, P. K. B. E., VINAYAKA, C. R., ZHANG, J., AND BARKER, W. C. 2003. The protein information resource. *Nucleic Acids Research 31*, 345–347.

WU, D., BENNETT, K. P., CRISTIANINI, N., AND SHAWE-TAYLOR, J. 1999. Large margin decision trees for induction and transduction. In *Proc. International Conference on Machine Learning*. 474–483.

WU, J., KASIF, S., AND DELISI, C. 2003. Identification of functional links between genes using phylogenetic profiles. *Bioinformatics 19*, 12, 1524–1530.

WU, L. F., HUGHES, T. R., DAVIERWALA, A. P., ROBINSON, M. D., STOUGHTON, R., AND ATSCHULER, S. J. 2002. Large-scale prediction of saccharomyces cerevisiae gene function using overlapping transcriptional clusters. *Nature Genetics 31*, 3, 255–265.

WU, Q. AND MANIATIS, T. 1999. A striking organization of a large family of human neural cadherin-like cell adhesion genes. *Cell 97*, 6, 779–790.

XENARIOS, I. AND EISENBERG, D. 2001. Protein interaction databases. *Curr Opin Biotechnology 12*, 4, 334–339.

XENARIOS, I., SALWINSKI, L., DUAN, X. J., HIGNEY, P., KIM, S.-M., AND EISENBERG, D. 2002. DIP and the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Research 30*, 1, 303–305.

XIE, H., WASSERMAN, A., LEVINE, Z., NOVIK, A., GREBINSKIY, V., SHOSHAN, A., AND MINTZ, L. 2002. Large-scale protein annotation through Gene Ontology. *Genome Research 12*, 5, 785–794.

XIONG, H., HE, X., DING, C., ZHANG, Y., KUMAR, V., AND HOLBROOK, S. R. 2005. Identification of functional modules in protein complexes via hyperclique pattern discovery. In *Proc. Pacific Symposium on Biocomputing (PSB)*. 221–232.

XIONG, H., TAN, P.-N., AND KUMAR, V. 2003. Mining strong affinity association patterns in data sets with skewed support distribution. In *ICDM '03: Proceedings of the Third IEEE International Conference on Data Mining*. 387–394.

XIONG, H., TAN, P.-N., AND KUMAR, V. 2006. Hyperclique pattern discovery. *Data Min. Knowl. Discov. 13*, 2, 219–242.

YANAI, I., DERTI, A., AND DELISI, C. 2001. Genes linked by fusion events are generally of the same functional category: a systematic analysis of 30 microbial genomes. *Proc Natl Acad Sci U.S.A. 98*, 14, 7940–7945.

YANG, J., WANG, H., WANG, W., AND YU, P. 2003. Enhanced biclustering on expression data. In *Proc. Third IEEE Symposium on BioInformatics and BioEngineering (BIBE)*. 321–327.

YANG, Y. AND PEDERSEN, J. O. 1997. A comparative study on feature selection in text categorization. In *Proc. Fourteenth International Conference on Machine Learning (ICML)*. 412–420.

YANG, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci. 13*, 5, 555–556.

YEH, A., MORGAN, A., COLOSIMO, M., AND HIRSCHMAN, L. 2005. BioCreAtIvE Task 1A: gene mention finding evaluation. *BMC Bioinformatics 6*, Suppl 1, S2.

YU, H., LUSCOMBE, N. M., LU, H. X., ZHU, X., XIA, Y., HAN, J.-D. J., BERTIN, N., CHUNG, S., VIDAL, M., AND GERSTEIN, M. 2004. Annotation transfer between genomes: Proteinprotein interologs and proteinDNA regulogs. *Genome Research 14*, 6, 1107–1118.

ZBODNOV, E. M. AND APWEILER, R. 2001. InterProScan–an integration platform for the signature-recognition methods in InterPro. *Bioinformatics 17*, 9, 847–848.

ZEHETNER, G. 2003. OntoBlast function: From sequence similarities directly to potential functional annotations by ontology terms. *Nucleic Acids Research 31*, 13, 3799–3803.

ZHANG, W. AND CHAIT, B. T. 2000. ProFound: an expert system for protein identification using mass spectrometric peptide mapping information. *Anal Chem. 72*, 11, 2482–2489.

ZHANG, W., MORRIS, Q. D., CHANG, R., SHAI, O., BAKOWSKI, M. A., MITSAKAKIS, N., MOHAMMAD, N., ROBINSON, M. D., ZIRNGIBL, R., SOMOGYI, E., LAURIN, N., EFTEKHARPOUR, E., SAT, E., GRIGULL, J., PAN, Q., PENG, W.-T., KROGAN, N., , GREENBLATT, J., , FEHLINGS, M., VAN DER KOOY, D., AUBIN, J., BRUNEAU, B. G., ROSSANT, J., BLENCOWE, B. J., FREY, B. J., AND HUGHES, T. R. 2004. The functional landscape of mouse gene expression. *J. Biol. 3*, 5, 21.

ZHENG, Y., ROBERTS, R. J., AND KASIF, S. 2002. Genomic functional annotation using co-evolution profiles of gene clusters. *Genome Biology 3*, 11, research0060.1–0060.9.

ZHOU, X., KAO, M.-C. J., AND WONG, W. H. 2002. Transitive functional annotation by shortest-path analysis of gene expression data. *Proc Natl Acad Sci U.S.A. 99,* 20, 12783–12788.

ZHOU, Y., YOUNG, J. A., SANTROSYAN, A., CHEN, K., YAN, S. F., AND WINZELER, E. A. 2005. In silico gene function prediction using ontology-based pattern identification. *Bioinformatics 21,* 7, 1237–1245.