Toward a Systematic Definition of Protein Function That Scales to the Genome Level: Defining Function in Terms of Interactions

NING LAN, RONALD JANSEN, AND MARK GERSTEIN

Invited Paper

The ultimate goal of functional genomics is to elucidate the function of all the genes in the genome. However, the current notions of function are crafted for individual proteins. The degree to which they can scale to the genomic level is not clear. In this paper, we review the diverse approaches to functional classification, focusing on their ability to meet this challenge of scale. Our review emphasizes a number of key parameters of the systems: their accuracy, comprehensiveness, level of standardization, flexibility, and support for data mining. We then propose an approach that synthesizes a number of the promising features of the existing systems. Our approach, which we call a function grid, is based on the notion of defining a protein's function through molecular interactions—specifically, in terms of its probability of interaction with various ligands, the list of which can be expanded infinitely. To illustrate how our function grid can be used in genome-wide prediction of function, we construct a grid of yeast genes; combine it with other genomic information, including sequence features, structure, subcellular localization, and messenger ribonucleic acid expression; and then use decision trees and support vector machines to predict deoxyribonucleic acid binding.

Keywords*—*Function, grid, interaction, ontology, proteome.

I. Introduction

The recent flood of genomic sequence and structural data has shifted the research focus of global-scale biology from deoxyribonucleic acid (DNA) to ribonucleic acid (RNA) and proteins, and has presented the challenge for bioinformatics to turn data into knowledge, i.e., integrate the ever-growing data to ascribe functions of proteins, cells, and ultimately organisms [1]. Exactly how function is defined on a global scale is particularly important. The definition of protein function has progressed rapidly in recent years toward systematic representation, and is still under intensive study and debate. In this paper, we review the pros and cons of the currently established systems of functional representation and project a gridlike structure that defines protein function through molecular interactions.

Protein function can be determined by experimental means or through homology-based annotation transfer.

Manuscript received April 30, 2002; revised September 8, 2002.

N. Lan and R. Jansen are with the Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520 USA (e-mail: lan@bioinfo.mbb.yale.edu; Ronald.Jansen@Yale.edu).

M. Gerstein is with the Departments of Molecular Biophysics and Biochemistry and Computer Science, Yale University, New Haven, CT 06520 USA (e-mail: Mark.Gerstein@Yale.edu).

Digital Object Identifier 10.1109/JPROC.2002.805302

II. Experimental Methods for Genomic Function Determination

Experimental approaches to the analysis of gene function on a genomic scale include oligonucleotide and complementary DNA (cDNA) microarrays, gene disruption through transposon insertion [2] or deletion [3], [4], yeast two-hybrid assays [5]–[8], proteome microarrays [9]–[11], and the tandem affinity purification (TAP) tagging method [12], [13]. These methods each aim at defining gene function from different angles; therefore, each has its own strengths and weaknesses. Oligonucleotide and cDNA microarrays measure messenger RNA (mRNA) expression on a genomic scale under various conditions, and thus indirectly indicate each gene's involvement in certain biological processes, and which genes may have related cellular function. Yeast two-hybrid assays explore protein–protein interactions in a pairwise fashion, whereas TAP tagging is useful for detecting protein complexes of two or more proteins. Proteome chip can measure both the biochemical activity of proteins and the interaction of proteins with other molecules, such as other proteins, metabolites, or drugs. All these approaches aim to elucidate gene function in terms of molecular interactions, the caveat being that the experimental systems do not exactly mimic physiological conditions; therefore, the results obtained may not agree with individual *in vivo* assays. Gene disruption measures the resulting phenotype following disablement of each gene and thereby explicates

gene function in terms of physiological activity of the organism. However, this is applicable to only a subset of genes whose interruption causes discernible phenotypic changes.

Computational approaches can be used to organize the genome-scale data into clusters of functionally related genes or to indicate the involvement of genes in certain biological processes, whereas the precise function of a gene needs to be determined through either individual experimental assays or homology-based prediction.

III. Annotation Transfer for Functional Genomics

The function of proteins of known sequence has been experimentally determined for only a small fraction of them. For a larger fraction of proteins, functional annotation is transferred based on the idea that proteins of similar sequence and structure are presumably descended from a common ancestral protein, and have related functions. In practice, annotation transfer has been made possible by the exponential growth of the number of fully sequenced genomes. Initially, the best hit in database comparisons based on simple sequence similarity was used for annotation transfer [14], [15]. This straightforward approach is usually credible when the compared species are relatively close phylogenetically [16]. However, at a larger phylogenetic distance, the situation is complicated by the occurrence of gene duplications [17]. More robust algorithms became increasingly available, often focused on the existence of key motifs and patterns associated with function, followed by structure modeling [18]–[20]. Functional linkage can also be detected through phylogenetic profiling, analysis of fusion pattern of protein domains, and the gene neighbor method (for a review, see [21]). Success in this field was facilitated by the tremendous growth in the number of known three-dimensional protein structures.

However, much caution needs to be taken in annotation transfer, in that the relationship between sequence or structure similarity and functional similarity is not as straightforward as that between sequence and structure similarity. Incorrect annotation transfer could result in progressive corruption of genome databases, as the error could be carried over to other proteins when the errant proteins are used as basis for further annotations [22]–[24]. Although a clear, well-characterized relationship exists between sequence and structure similarity, the sequence-function and structure-function relationships are much more challenging to characterize explicitly. One limitation to the accuracy of functional annotation transfer is that a minimum similarity is required to reliably predict protein function. For protein pairs that share the same fold, usually 30%–40% sequence identity is required for function to be conserved precisely [25], [26]. Examples also exist where proteins of high sequence and structural similarity perform disparate functions, such as lysozyme and a-lactalbumin, or proteins with different structural folds have identical function, such as subtilisin and chymotrypsin [27].

Another limitation lies in the vague definition of "function" itself. The rapidly growing number of fully sequenced genomes calls for the development of a comprehensive system for functionally classifying proteins that support interoperation of genomic databases.

IV. Systematic Representation of Protein Function

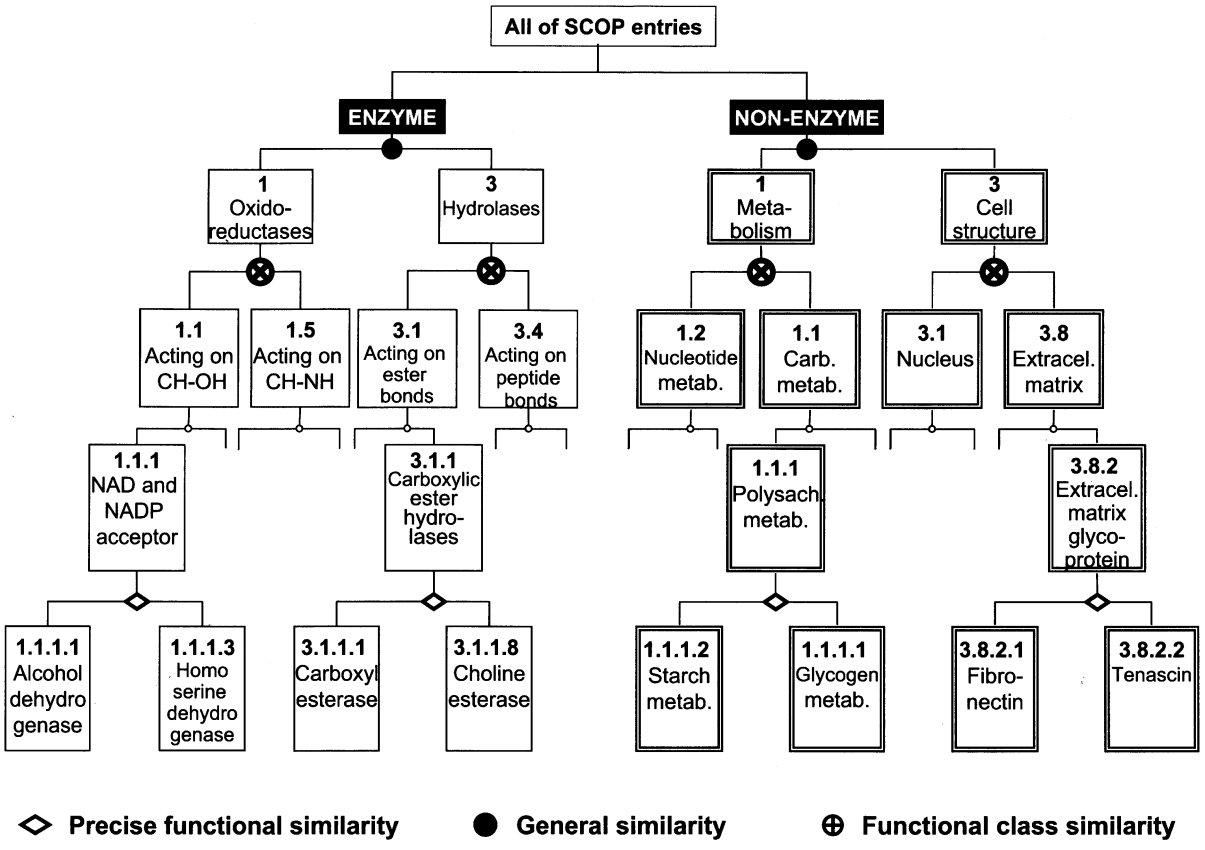
Traditionally, biologists consider protein function as a simple phrase, often indicated by the name of the gene, such as mkk1 [mitogen-activated protein (MAP) kinase kinase] or tbf1 (TTAGGG repeat binding factor). However, often the name of a gene is not directly related to its function and can be misleading. Sometimes one who discovers a gene may name it arbitrarily, such as the *Drosophila* genes *Yippee*, for the reaction of a graduate student upon cloning it, or *Starry Night*, for the swirling hair pattern resulting from mutation of the gene, which resembles the painting by Van Gogh [28]. More often, the name is based on how a gene was identified, such as the *hsl* genes (histone synthetic lethal), which may have something to do with its function in a very unclear way.

When the function of a gene cannot be inferred from its name, one has to resort to database records or the literature. Early functional annotation tended to be recorded as simple phrases, which are nonstandard, highly unstable, and have no organized structure among functions. Moreover, function has been described from different angles, depending on the experimental perspective. Biochemists characterize protein function in terms of molecular interaction. Cell biologists describe protein function as its role in a cellular process. Geneticists characterize genes by the phenotype of their mutations. Standard ontology systems that integrate these various conceptualizations in genomics and define exact specifications of function need to be established in order to facilitate cross-query and annotation transfer as well as a variety of projects that entail interoperation of the ever-increasing biological databases.

V. Hierarchical Representation of Protein Function

One approach is the hierarchical representation adopted by most functional ontologies such as the Gene Ontology (GO) Consortium [29], the Munich Information Center for Protein Sequences (MIPS) Functional Classification Catalog, [30] and the Enzyme Commission (EC) classification [31]. Fig. 1(a) shows a simplified hierarchical structure that Wilson *et al.* [26] adopted to represent enzyme and nonenzyme function. Sharing of classification numbers indicate functional similarity. One can trace up and down the hierarchy to find whether one function is part of another function, and whether or not (but not quantitatively) there is any commonality between two functions, i.e., whether they descend from the same broad function.

The ENZYME system developed by the EC classifies enzymes by reaction type, and can be applied to proteins in many different organisms [31]. However, it is applicable only to enzymes, and has two major drawbacks. First, it does not consider catalytic reaction mechanisms, and therefore often ignores obvious similarities [32]. Second, it presumes a $1:1:1$ relationship between gene, protein, and reaction,



(a)

Fig. 1. Systematic representation of protein function. (a) Hierarchical scheme for functional classification, adapted from [26]. In a tree-structured schema, functional similarity is measured by the height of common ancestor. In practice, the path of each node from the root is encoded into a classification number, and comparison is done by scanning the classification numbers from left to right. If two proteins are both enzymes or both nonenzymes, then they possess general functional similarity. If they share the first component of their classification numbers, then they are in the same functional class. If they share the first three components of their enzyme numbers (or the equivalent for nonenzyme numbers, depending on category) then they have the same precise function.

while an enzyme can be multifunctional, or an enzyme can be formed by polypeptides from two different genes [26].

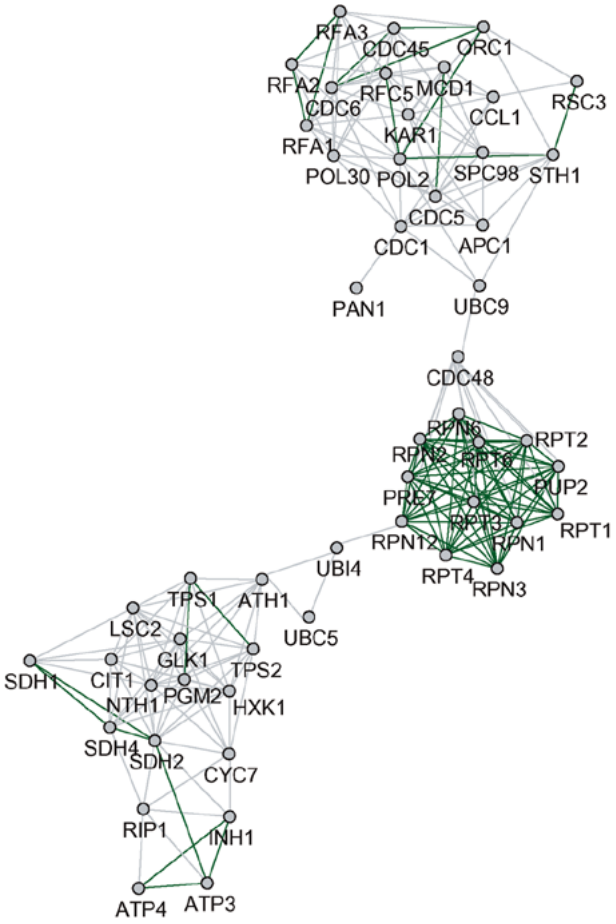
The functional role categories developed at GenProtEC for *E. coli* (http://genprotec.mbl.edu/) and the MIPS Functional Classification Catalog for yeast (http://mips.gsf.de/proj/yeast/catalogues/) are organized according to a hierarchical decomposition of cellular processes. These classifications integrate enzyme and nonenzyme functions from the start and are widely used. However, they are each applicable to only a single organism, and therefore cannot be readily applied in annotation transfer.

The Gene Ontology Consortium has been highly successful in creating a structured and precisely defined controlled vocabulary for describing gene function across several organisms [29]. The GO project started as a joint project between FlyBase, the *Saccharomyces* Genome Database, and Mouse Genome Informatics, attempting to merge the fly, yeast, and mouse functional classification schemes. As of December 2001 it has integrated genes from seven organisms; participating groups including the *Arabidopsis* Information Resource, Pombase, WormBase, Compugen, and the Institute for Genomic Research, and is closely integrated with InterPro, which facilitates maintenance of the association of protein motifs with functional descriptions [33].

GO classifies functions as a directed acyclic graph (DAG). Nodes can often be reached from multiple paths, which facilitates the representation and comparison of genes that have multiple functions or that are involved in more than one process. GO classifies genes into three parallel categories, i.e., three DAGs: biological process, molecular function, and cellular component. This allows for defining the function of a gene at various levels, including its biochemical activities and biological roles as well as cellular structure.

VI. Network Graph Representing Protein Function

Another approach to global representation of gene function is through network graphs, including pathway maps and protein–protein interaction maps. These graphs differ from the hierarchical representation in that each node is not a function, but a protein or a substrate/product of a reaction. The link between two nodes indicates an interaction. They can provide a framework from which complex regulatory information can be extracted.



(b)

Fig. 1. (*Continued.*) Systematic representation of protein function. (b) Example of a yeast protein network. The green edges represent protein-protein interactions from the MIPS complexes catalog [30], two yeast–two hybrid datasets [7], [8], and two *in vivo* pull-down datasets [12], [13]. The gray edges stem from a computational analysis of different data indicating protein-protein interactions; these data include information on whether two proteins are localized in the same subcellular compartment, whether they are coexpressed under the same physiological conditions, and whether they are involved in the same biological processes [56]–[58].

One example of a pathway graph is EcoCyc, an ontology that describes metabolic pathways and other cell functions of the *E. coli* genome by encoding information about the molecular interaction of *E. coli* genes [34]. It uses distinct frames to represent the molecule and its chemically modified forms, and then models its interactions by labeling it substrate, catalyst, modulator, or cofactor in a reaction. One can choose to view the global structure of the entire network or to search for the position of a specific gene in its local network.

Protein–protein interaction maps represent a population of interacting proteins displayed as networks or circuits. An example is shown in Fig. 1(b). The yeast two-hybrid system is one of the major methodologies for large-scale analysis of protein–protein interactions. Interaction maps combining yeast two-hybrid studies with previous annotations have been generated [6]. The more recently developed proteome microarray technology allows for direct analysis of a variety of interactions, including interactions between proteins [9]–[11]. There are several public databases containing protein interaction maps, including Myriad's Pronet Database (http://www.myriad-pronet.com/), Curagen's Pathcalling Yeast Interaction Database (http://portal.curagen.com), the Biomolecular Interaction Network Database (BIND, http://www.bind.ca/), and the Database of Interacting Proteins (DIP, http://dip.doe-mbi.ucla.edu/).

Protein interactions have also been predicted by computational methods based on genomic sequence [36] or mRNA expression [37]. We found that gene expression data are sometimes more meaningful when they are grouped under a protein complex scheme rather then a functional classification scheme (see Fig. 2). A comparison of the MIPS functional catalog with the MIPS complexes catalog shows that there are many proteins assigned to more than one function, but there are only very few proteins that are members of more than one protein complex. A related observation is that protein complexes often form a functional unit as a whole, whereas the individual proteins exhibit no or only a reduced number of functions themselves. Functional versatility is thus often created on the level of protein complexes. These properties of the complexes catalog often allow a less ambiguous and more straightforward analysis of the observations of genomic experiments, such as cDNA microarray expression data. However, a disadvantage of the complexes classification is that it characterizes a much smaller number of proteins than the functional classification (in MIPS, 3687 proteins are functionally classified, whereas only 1137 proteins are assigned to complexes).

Protein–protein interaction maps have not only confirmed the existence of previously known complexes and pathways but have also shed light on the discovery of new complexes and crosstalk between previously unlinked pathways [7], [38]. An interaction map generated in one species can potentially be used to predict interactions in another species, presuming that large numbers of physically interacting proteins in one organism have evolved in a correlated fashion so that their respective orthologs in other organisms also interact [39].

VII. Limitations of the Current Ontology Systems

Up to now, ontologies that define gene function as hierarchical structure are all based on natural language. Although a protein's function can be defined with relative accuracy through a controlled vocabulary and cross-linked hierarchical structures, the use of natural language limits the precision of function definition and potential applications of computational automation.

The most basic question in functional computation is whether two proteins have the same function. This appears easy to answer by directly comparing GO terms or MIPS functional categories. However, this functional equality is relative and approximate, since natural language-based ontologies may not be fine-tuned enough to reflect the complex cellular function and regulation of each gene. To answer the question of functional equality more precisely, one needs to integrate functional information from a variety

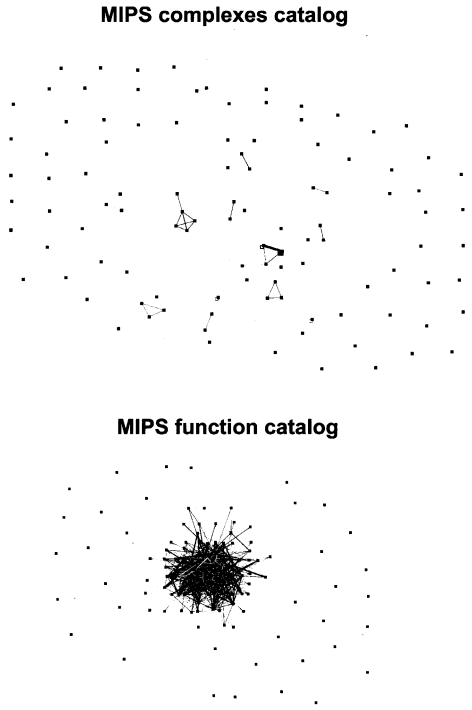


Fig. 2. Protein complex versus functional classification. (top) Undirected graph representation of the second hierarchy level of the MIPS complexes catalog, where each node represents one class (i.e., protein complex). Nodes are connected to one another if two protein complexes share at least one protein, in the sense that two copies of the same protein occur in each of both complexes. It is obvious that most protein complexes represent protein classes disjoint from other protein complexes. Furthermore, for those protein complexes that share common proteins, the amount of overlap is relatively small. (bottom) Graph representation of the second hierarchy level of the MIPS functional catalog, with each node representing a functional protein class and each edge indicating at least one shared protein. Although some functional classes are not connected to others, there is a large cluster of functional classes that are strongly interrelated.

of resources, including pathway and interaction maps, which is no easy task.

For two nonidentical but related functions, the degree of similarity is much harder if not impossible to answer using natural language-based ontologies. When comparing two functional GO terms, their names and positions in the GO hierarchy often do not provide full information on the level of similarity between them. For example, both hexokinase (GO:0004396) and carbamoyl–phosphate synthase (GO:0004088) are adenosine triphosphate (ATP)–binding proteins, but such information is not available from the GO hierarchy. In this case it would be helpful to resort to EC, which lists the known reactions catalyzed for each enzyme class and contains information on molecular interactions. Moreover, there are multifunctional proteins or proteins involved in multiple cellular processes that can be associated with more than one GO term in each of three level categories. On the other hand, certain functions may be meaningful only in terms of protein complexes. In such cases the interaction network graph may provide a more accurate picture of the protein. Another situation is that two genes may have the same cellular function but are under different regulation. For example, myoglobin and hemoglobin are both heme proteins that bind molecular oxygen. However, oxygen binding of hemoglobin is under regulation of allosteric effectors including oxygen, iron, CO_2 and 2,3-bisphosphoglycerate, whereas there is no allosteric regulation with the oxygen-binding property of myoglobin [40].

Consider the following more complex questions: Is the function of protein X more similar to protein Y than to protein Z? Among a group of proteins with known function, are there subgroups that are more closely related? Can novel function be deduced based on known function and other features of a protein? These questions can be easy to solve if function is represented using numeric values to allow the use of standard data-mining algorithms. Here we propose a grid-like structure that represents protein function in term of interaction probabilities and discuss its potential application in function prediction.

VIII. Construction and Potential Application of the Function Grid

In our function grid, the proteome interaction map is represented as a matrix, as each protein is associated with a row vector that consists of the probability of binding to various ligands [see Fig. 3(a)]. Denote by P the set of proteins, and L the set of binding ligands. The function grid G can be defined as mapping

$$g\colon P \times L \rightarrow [0, 1]$$

as

$$g(p_i, l_j) = G_{i,j}.$$

As an initial effort, we constructed the grid with the complete proteome from the budding yeast *Saccharomyces cerevisiae*. The interaction data were collected from GO, EC, yeast two-hybrid system interactions [5], [8], and proteome chip experiments [11]. The dimension of each row vector can potentially be infinite, as it expands when experimental data for previously unknown ligands become available.

Functional similarity between two proteins can be defined by the cosine of the angle between the two corresponding vectors. Then proteins can be grouped according to function similarity using a number of clustering methods.

The interaction grid can be combined with sequence and structural features and cellular localization, as well as expression data, to make up a more comprehensive grid, which can be used for data mining as we deduce novel interactions based on known ones. Fig. 3(b) is a schematic representation of how DNA-binding proteins can be predicted combining the interaction grid with other genomic datasets.

Several issues need to be addressed in designing the interaction grid. First, there needs to be a systematic way to define a binding probability, which determines the accuracy of the calculations. Information gathered from previous annotations or experiments, represented in the form of either a phrase or signal intensity, should be converted to a value between zero and one, depending on the source or nature of experiments [see Fig. 3(a)]. By associating each binding

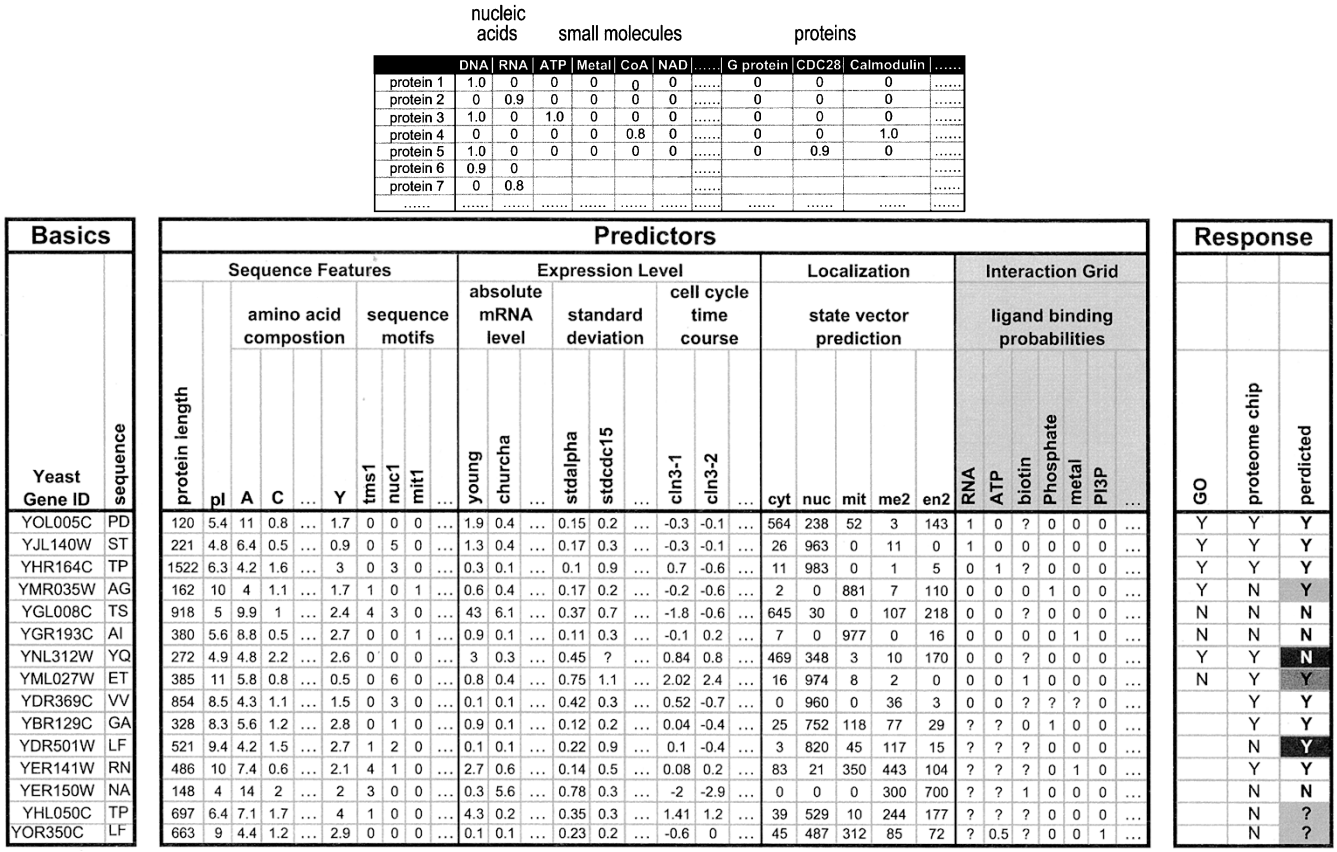


Fig. 3. Function grid and its application in functional prediction. (a) A simplified example of an interaction grid. The function of each protein is defined as a row vector that consists of the probability of binding to various ligands. The grid is filled with data collected from GO, EC, yeast two-hybrid system interactions, and proteome chip experiments. For information gathered from GO, based on the GO evidence code associated with each entry (defined at http://www.geneontology.org/GO.evidence.html), we assigned probabilities from 0.8 (NR) to 1.0 [traceable author statement (TAS) and inferred from direct assay (IDA)]. Using the data from proteome chip experiments, we define the binding probability of each protein (P_i) by normalizing its binding signal (S_i) against the lowest value of all proteins that are known to bind the ligand (S_m):$P_i = \min(S_i/S_m, 1)$. The value is left empty when binding probability is unknown. The dimension of each row vector can be expanded when experimental data for previously unknown ligands become available. (b) Schematic display combining the interaction grid with other genomic information to predict DNA binding. The interaction grid is combined with sequence features, expression data, and localization information to predict DNA binding. Prediction results are compared with the GO annotation and the protein chip data as a control. Each predication result can have several possible outcomes: consistent with the GO annotation and the protein chip results (white); consistent with respect to only one of the two controls (gray); inconsistent with both the GO annotation and the protein chip (black). In some cases, it may be impossible to predict whether the protein binds DNA or not.

probability with an evidence field, which records where the data were collected and indicates how the probability was assigned, we have created an evidence system similar to that of GO.

Second, we need to consider what and how many ligands to put into the grid, and the relationship between these ligands. On the one hand, we want to collect every possible piece of information on molecular interactions. In the meantime, these ligands need to be grouped into a hierarchical structure, allowing the function grid to be viewed and mined at multiple levels (see Fig. 4).

Third, the initial function grid was constructed of yeast proteins. When information on molecular interaction from other organisms is collected, how are we going to integrate them, i.e., should homologues be treated as different fields of the same protein or as different proteins? Our decision is to construct individual matrices for each organism and keep evolutionary relations between homologues in another table. This way the similarity and difference between interaction partners among homologues can be easily calculated by calculating the distance between the respective binding vectors.

The fourth point is concerned not so much with data mining but more with the power of this interaction grid system to represent gene function in the context of cellular regulation. Apart from probability and evidence, each reaction has two extra fields of action and condition, to indicate the reaction type and regulation of this interaction. Fig. 5 shows how two steps in the MAP kinase pathway involved in the maintenance of cellular integrity [41] are represented in the interaction grid.

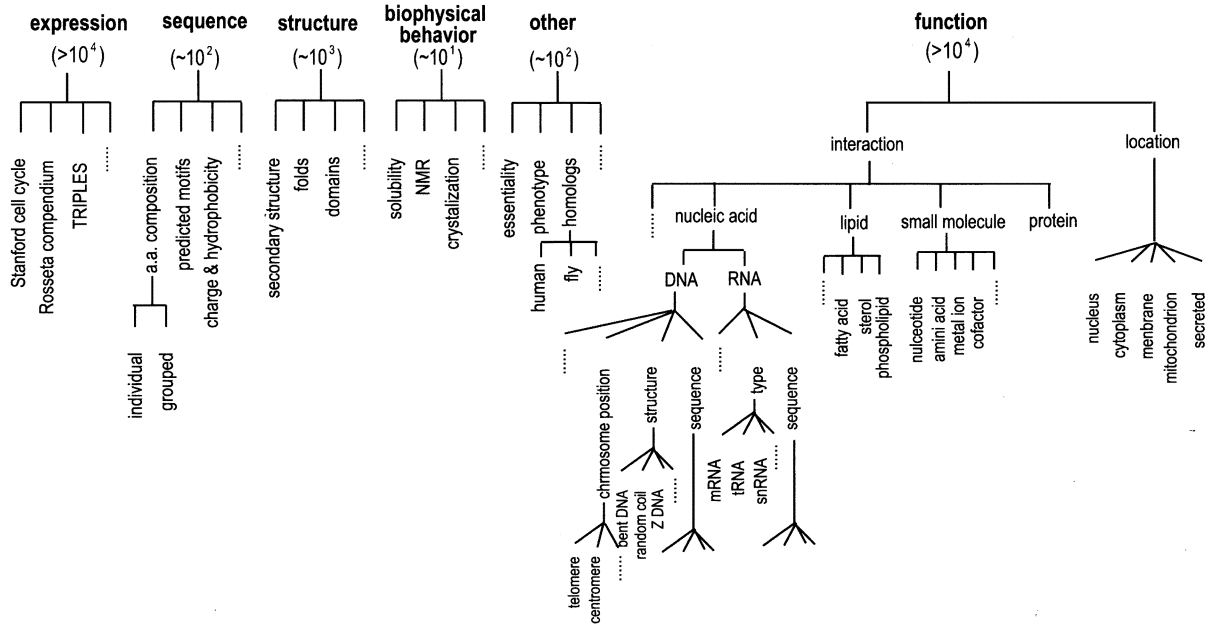


Fig. 4. Hierarchical organization of the function grid. The fields in the function grid can be grouped into a hierarchical structure, so that the data mining can be performed at various levels. The range of potential number of fields (columns) for each group is indicated in parentheses. Areas where rapid expansion is expected in the near future are in *italic*.

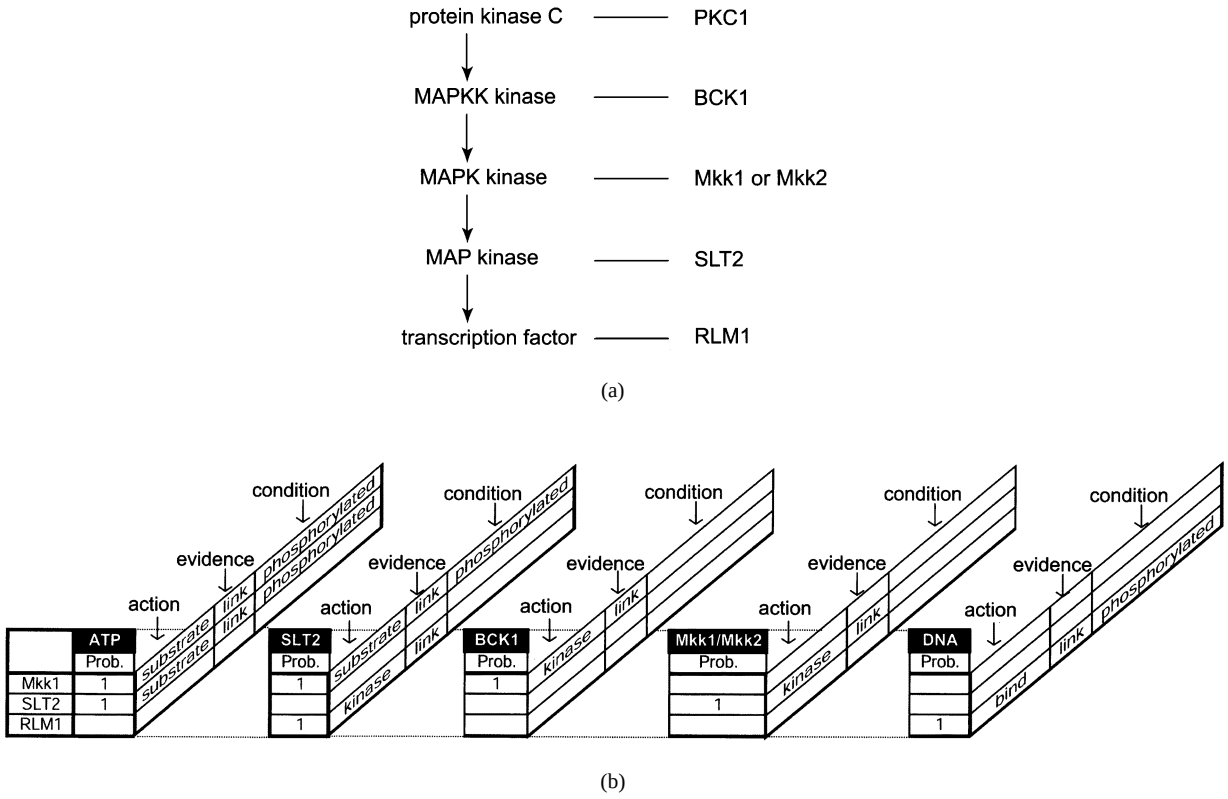


Fig. 5. Representation of part of a signal transduction pathway. Here, we show (a) schematic representation of some of the main components of yeast protein kinase C cascade and (b) how part of this cascade is represented in the interaction grid. Mkk1 phosphorylates SLT2 when phosphorylated by BCK1. SLT2 phosphorylates RLM1 SLT2 when phosphorylated by Mkk1 or Mkk2. RLM1 binds DNA when phosphorylated by SLT2. "Link" in the evidence field refers to the original publication.

Table 1
Protein Properties Selected for DNA-Binding Prediction

Feature	Description	Number
C(r)	Single-residue composition (occurence over sequence length); r = A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y	20
C(c)	Combined amino acid compositions; c = KR, NQ, DE, ST, LM, FWY, HKR, AVILM, DENQ, GAVL, SCTM	11
P(c)	Probability of localization in one of the five general cellular compartments; c=nuc, cyt, en2, me2, mit	5
B(l)	Probability of binding to various ligands l=RNA, ATP, protein, phosphate, metal, 6 phospholipids, NAD, CoA	13
pI	pI (Isoelectric Point) values (from-MIPS).	1
Hphobe	Maximum GES hydrophobicity score of a 20-residue window, iin kcal/mol	1
CPLX (x)	Normalized low-complexity sequences; x=s(short) or l (long)	2
young	Absolute mRNA expression in a GeneChip experiment (Holstege et al., 1998).	1
sagem_phase	Absolute mRNA expression of g/m phase proteins in the SAGE experiment (Velculescu et al., 1997).	1
sagel_phase	Absolute mRNA expression of l phase proteins in the SAGE experiment (Velculescu et al., 1997).	1
sages_phase	Absolute mRNA expression of s phase proteins in the SAGE experiment (Velculescu et al., 1997).	1
churcha	Church Absolute mRNA Expression (a)	1
churchalpha	Church Absolute mRNA Expression (alpha)	1
churchgal	Church Absolute mRNA Expression (gal)	1
churchheat	Church Absolute mRNA Expression (heat)	1
samson	Samson Absolute mRNA Expression	1
stdcdc15	the cdc15 arrest time series experiment in Yeast Cell Cycle Analysis Project (Spellman et al., 1998).	1
stdcdc28	Standard deviation in mRNA expression level over time (i.e. expression fluctuation) for a protein in the cdc28 time series experiment in Yeast Cell Cycle Analysis Project (Spellman et al., 1998).	1
stdalpha	the Alpha-factor arrest time series experiment in Yeast Cell Cycle Analysis Project (Spellman et al., 1998).	1
stdelu	Standard deviation in mRNA expression level over time (i.e. expression fluctuation) for a protein in the elutriation time series experiment in Yeast Cell Cycle Analysis Project (Spellman et al., 1998).	1
stddiauxic	the diauxic shift experiment (DeRisi et al., 1997).	1

Sequence features include amino acid compositions, biochemical properties, and entropy complexity measures based on the SEG program, secondary structure prediction, and hydrophobicity scores on the GES scale. Interaction probabilities are assigned based on GO and EC annotation, and protein chip experiments. Expression data are collected from multiple sources.

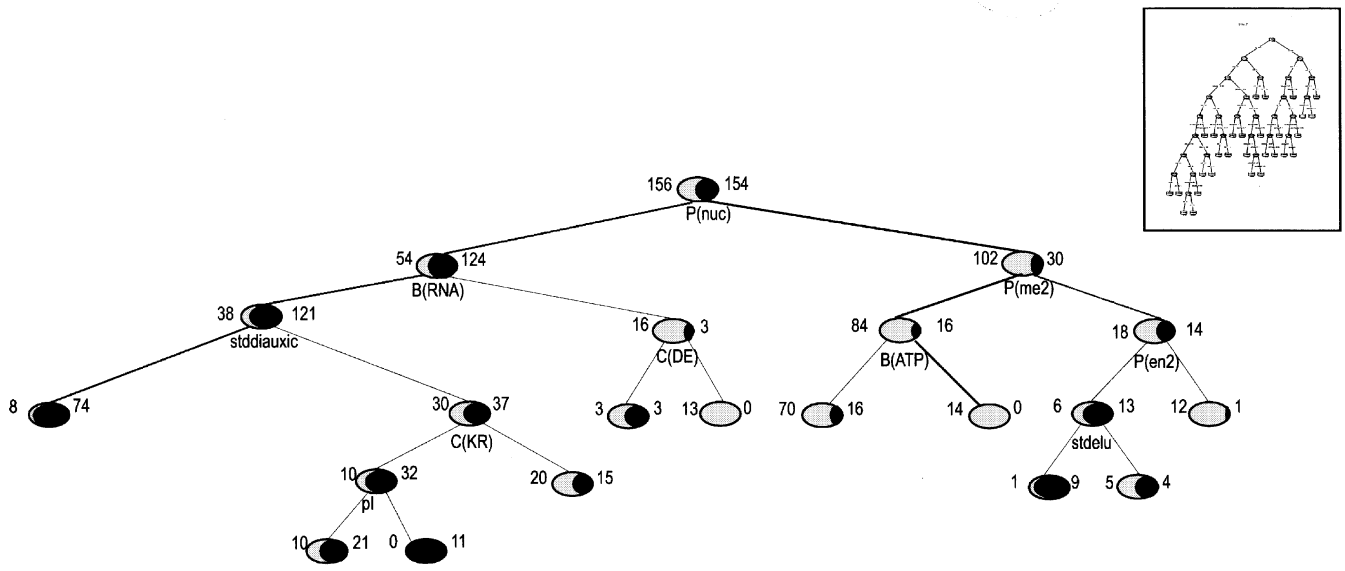
IX. Case Study—Prediction of DNA Binding

DNA-binding proteins play an essential role in the genetic activities of an organism. Structural analyzes reveal a high diversity of protein–DNA complex geometries found in nature, while the main mode of interaction in more than one-half of the protein families is the interaction between α helices and the DNA major groove [42]. The most significant DNA-binding motifs identified include helix-turn-helix, zinc-coordinating, and leucine zipper. A number of applications have been developed to predict protein–DNA binding through the search of particular motifs [43], [44] or a molecular docking simulation [45], [46]. These methods have gained only limited success, largely because of the complexity of determination factors of protein–DNA binding. Also, they require that the protein in question contain certain known DNA-binding motifs or have its structure solved in an uncomplex form [42].

In this paper, as an illustration of the data-mining application of the function grid, we applied supervised learning algorithms to probe the possibility of predicting DNA-binding activities of proteins that may not have a distinct DNA-binding motif or solved structure, using our combined function grid. As a proof of principle, we chose decision-tree learning for its ease of interpretation. Straightforward rules can be inferred by traversing the tree from root to leaf nodes. Moreover, classification can be based on an arbitrary mixture of symbolic and numeric values, and it is not necessary to scale the variables relative to one another. The model is generally robust in the presence of missing values, which is important because in many cases the probability of interaction with certain ligands is unknown [47].

We selected a training set of 156 DNA-binding proteins and 154 non-DNA-binding proteins based on GO annotation. Only proteins with the most reliable GO evidence (codes TAS or IDA), and thereby having a DNA-binding probability of 1.0, were selected as DNA-binding. Non-DNA-binding proteins were randomly selected from proteins with known molecular function that are not characterized as DNA-binding proteins in GO.

We selected the following features as predictor: a total of 53 sequence features, including amino acid composition, hydrophobicity, occurrence of low-complexity regions, etc.; probability of localization in one of the five generalized cellular compartments [48]; expression data from 14 different



Level	Feature splitting thresholds
1	nuc=0.646
2	RNA=0.5; me2=0.052
3	stddiauxic=0.255; DE=11.64; ATP=0.5; en2=0.052
4	KR=11.1; stdelu=0.18
5	pI=8.2

Fig. 6. Decision tree for DNA-binding prediction. DNA-binding and non-DNA-binding proteins are indicated by the numbers to the left and right of each node, respectively. The inset shows the complete tree, the upper section of which is pruned and used for rule extraction. The bold path leads to DNA-binding proteins, and the dashed one leads to non-DNA-binding ones. Nonmembrane nuclear proteins that bind to ATP are likely to bind DNA; proteins are predicted not to bind DNA if they are not likely to be localized in the nucleus, do not bind RNA, and have low standard deviation in diauxic shift.

experiments, including absolute value and standard deviation in each experiment [49]–[52]; probability of interaction with RNA, ATP, protein, phosphate, metal ion, nicotinamide adenine dinucleotide, coenzyme A, and six phospholipids, as summarized in Table 1.

A decision tree was constructed to partition the data, and stratified tenfold cross-validation was performed, where repeatedly a randomized 90% of the data was set for training and the remaining 10% for testing. The cross-validation approach resulted in an overall prediction success of 65%–70%. Only the top levels of the tree are significant in terms of yielding a generalized concept and deriving useful rules. Fig. 6 illustrates the upper five levels of the tree built on 310 proteins and subject to cross-validation. Two interesting paths are highlighted. Following the right path of the tree, DNA-binding proteins are selected, provided that they have a high probability of being localized in the nucleus, are not likely to be membrane proteins, and bind ATP. Non-DNA-binding proteins are selected by the left branch of the tree that have low probability of being nuclear proteins, do not bind RNA, and have low standard deviation in diauxic shift. Apparently most of these findings are consistent with the expected cellular location and molecular interaction of DNA-binding proteins, in that DNA-binding proteins should be located in the nucleus and many of them are involved in transcription, which involves RNA and ATP binding.

We also tried to predict DNA binding using the support vector machine (SVM) algorithm. This algorithm, widely used in pattern recognition, represents a method of finding binary classification rules from examples, for which they can guarantee the lowest error rate on new observations [53]. It had previously been employed to predict functional classes from gene expression data [54]. The data set applied to construct the decision tree was divided in half. One half was used to construct an SVM model using the SVM Light software [55]. The trained network was tested on the other half. This procedure was repeated ten times; the average percentage of correct identifications was 72%.

X. Conclusion

The availability of fully sequenced genomes challenges bioinformatics to elucidate the structure, interactions, and

functions of proteins on a genomic scale. Ontology systems are needed that can facilitate calculation of functions together with other biological data. Such ontology should aim at capturing all dimensions of protein function and should keep up with the phenomenal rate at which biological data are being produced. Current functional ontology systems are mainly based on natural language, which has limitations in the precision of function definition and therefore cannot readily support calculation of functional similarity.

We have proposed a grid-style representation of protein function through molecular interactions, and proved in principle that this function grid can be combined with various genomic data to predict new functions. One main concern about constructing this grid is how to collect experimental data on interactions and turn them into binding probability with relative accuracy, which requires careful consideration. With the rapid accumulation of proteome interaction data, we expect a significant increase in the power of function prediction. Therefore, we believe the development of this function grid system will prove highly valuable for global representation and calculation of protein function in the postgenomic era.

REFERENCES

[1] J. M. Thornton, "From genome to function," *Science*, vol. 292, pp. 2095–2097, 2001.

[2] P. Ross-Macdonald, "Large-scale analysis of the yeast genome by transposon tagging and gene disruption," *Nature*, vol. 402, pp. 413–418, 1999.

[3] L. Ni and M. Snyder, "A genomic study of the bipolar bud site selection pattern in *Saccharomyces cerevisiae*," *Mol. Biol. Cell.*, vol. 12, pp. 2147–2170, 2001.

[4] E. A. Winzeler *et al.*, "Functional characterization of the *Saccharomyces cerevisiae* genome by comprehensive and precise gene deletion and massively parallel analysis," *Science*, vol. 285, pp. 901–906, 1999.

[5] T. Ito *et al.*, "Toward a protein–protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins," in *Proc. Nat. Acad. Sci. USA*, vol. 97, 2000, pp. 1143–1147.

[6] B. Schwikowski, P. Uetz, and S. Fields, "A network of protein–protein interactions in yeast," *Nature Biotechnol.*, vol. 18, pp. 1257–1261, 2000.

[7] P. Uetz *et al.*, "A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*," *Nature*, vol. 403, pp. 623–627, 2000.

[8] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki, "A comprehensive two-hybrid analysis to explore the yeast protein interactome," in *Proc. Nat. Acad. Sci. USA*, vol. 98, 2001, pp. 4569–4574.

[9] G. MacBeath and S. L. Schreiber, "Printing proteins as microarrays for high throughput function determination," *Science*, vol. 289, pp. 1760–1762, 2000.

[10] H. Zhu *et al.*, "Analysis of yeast protein kinases using protein chips," *Nature Genetics*, vol. 26, pp. 283–289, 2000.

[11] H. Zhu *et al.*, "Global analysis of protein activities using proteome chips," *Science*, vol. 293, pp. 2101–2105, 2001.

[12] A. C. Gavin *et al.*, "Functional organization of the yeast proteome by systematic analysis of protein complexes," *Nature*, vol. 415, pp. 141–147, 2002.

[13] Y. Ho *et al.*, "Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry," *Nature*, vol. 415, pp. 180–183, 2002.

[14] C. M. Fraser *et al.*, "The minimal gene complement of *Mycoplasma genitalium*," *Science*, vol. 270, pp. 397–403, 1995.

[15] C. M. Fraser *et al.*, "Complete genome sequence of *Treponema pallidum*, the syphilis spirochete," *Science*, vol. 281, pp. 375–388, 1998.

[16] R. L. Tatusov *et al.*, "Metabolism and evolution of *Haemophilus influenzae* deduced from a whole-genome comparison with *Escherichia coli*," *Curr. Biol.*, vol. 6, pp. 279–291, 1996.

[17] R. L. Tatusov, E. V. Koonin, and D. J. Lipman, "A genomic perspective on protein families," *Science*, vol. 278, pp. 631–637, 1997.

[18] P. Bork and E. V. Koonin, "Protein sequence motifs," *Curr. Opin. Struct. Biol.*, vol. 6, pp. 366–376, 1996.

[19] Z. Zhang *et al.*, "Protein sequence similarity searches using patterns as seeds," *Nucleic Acids Res.*, vol. 26, pp. 3986–3990, 1998.

[20] T. K. Attwood *et al.*, "PRINTS prepares for the new millennium," *Nucleic Acids Res.*, vol. 27, pp. 220–225, 1999.

[21] D. Eisenberg, E. M. Marcotte, I. Xenarios, and T. O. Yeates, "Protein function in the post-genomic era," *Nature*, vol. 405, pp. 823–826, 2000.

[22] S. E. Brenner, "Errors in genome annotation," *Trends Genetics*, vol. 15, pp. 132–133, 1999.

[23] P. D. Karp, "A protocol for maintaining multidatabase referential integrity," in *Pac. Symp. Biocomput.*, 1996, pp. 438–445.

[24] P. Karp, "What we do not know about sequence analysis and sequence databases," *Bioinformatics*, vol. 14, pp. 753–754, 1998.

[25] H. Hegyi and M. Gerstein, "Annotation transfer for genomics: Measuring functional divergence in multi-domain proteins," *Genome Res.*, vol. 11, pp. 1632–1640, 2001.

[26] C. A. Wilson, J. Kreychman, and M. Gerstein, "Assessing annotation transfer for genomics: Quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores," *J. Mol. Biol.*, vol. 297, pp. 233–249, 2000.

[27] A. E. Todd, C. A. Orengo, and J. M. Thornton, "Evolution of function in protein superfamilies, from a structural perspective," *J. Mol. Biol.*, vol. 307, pp. 1113–1143, 2001.

[28] M. Vacek, "A gene by any other name," *Amer. Sci.*, vol. 89, no. 6, p. 500, Nov.–Dec. 2001.

[29] M. Ashburner *et al.*, "Gene ontology: Tool for the unification of biology," *Nature Genetics*, vol. 25, pp. 25–29, 2000.

[30] H. W. Mewes *et al.*, "MIPS: A database for genomes and protein sequences," *Nucleic Acids Res.*, vol. 30, pp. 31–34, 2000.

[31] E. C. Webb, *Enzyme Nomenclature 1992. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology*. New York: Academic, 1992.

[32] M. Riley, "Systems for categorizing functions of gene products," *Curr. Opin. Struct. Biol.*, vol. 8, pp. 388–392, 1998.

[33] S. Lewis, M. Ashburner, and M. G. Reese, "Annotating eukaryote genomes," *Curr. Opin. Struct. Biol.*, vol. 10, pp. 349–354, 2000.

[34] P. D. Karp, "An ontology for biological function based on molecular interactions," *Bioinformatics*, vol. 16, pp. 269–285, 2000.

[35] C. L. Tucker, J. F. Gera, and P. Uetz, "Toward an understanding of complex protein networks," *Trends Cell. Biol.*, vol. 11, pp. 102–106, 2001.

[36] E. M. Marcotte, M. Pellegrini, H.-L. Ng, D. W. Rice, T. O. Yeates, and D. Eisenberg, "Detecting protein function and protein–protein interactions from genome sequences," *Science*, vol. 285, pp. 751–753, 1999.

[37] R. Jansen, D. Greenbaum, and M. Gerstein, "Relating whole-genome expression data with protein–protein interactions," *Genome Res.*, vol. 12, pp. 37–46, 2002.

[38] M. Fromont-Racine *et al.*, "Genome-wide protein interaction screens reveal functional networks involving Sm-like proteins," *Yeast*, vol. 17, pp. 95–110, 2000.

[39] L. R. Matthews *et al.*, "Identification of potential interaction networks using sequence-based searches for conserved protein–protein interactions or interologs," *Genome Res.*, vol. 11, pp. 2120–2126, 2001.

[40] E. Antonini and M. Brunoni, *Hemoglobin and Myoglobin in Their Reactions with Ligands*. Amsterdam, The Netherlands: North-Holland, 1971.

[41] J. J. Heinisch, A. Lorberg, H. P. Schmitz, and J. J. Jacoby, "The protein kinase C-mediated MAP kinase pathway involved in the maintenance of cellular integrity in *Saccharomyces cerevisiae*," *Mol. Microbiol.*, vol. 32, pp. 671–680, 1999.

[42] N. M. Luscombe, D. Greenbaum, and M. Gerstein, "What is bioinformatics? A proposed definition and overview of the field," *Methods Inform. Med.*, vol. 40, pp. 346–358, 2001.

[43] M. Suzuki *et al.*, "DNA recognition code of transcription factors," *Protein Eng.*, vol. 8, pp. 1–4, 1995.

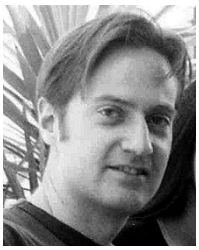
[44] Y. Mandel-Gutfreund and H. Margalit, "Quantitative parameters for amino acid–base interaction: Implications for prediction of protein–DNA binding sites," *Nucleic Acids Res.*, vol. 26, pp. 2306–2312, 1998.

[45] P. Aloy *et al.*, "Modeling repressor proteins docking to DNA," *Proteins*, vol. 33, pp. 535–549, 1998.

[46] M. J. Sternberg, H. A. Gabb, and R. M. Jackson, "Predictive docking of protein–protein and protein–DNA complexes," *Curr. Opin. Struct. Biol.*, vol. 8, pp. 250–256, 1998.
[47] P. Bertone *et al.*, "SPINE: An integrated tracking database and data mining approach for identifying feasible targets in high-throughput structural proteomics," *Nucleic Acids Res.*, vol. 29, pp. 2884–2898, 2001.
[48] A. Drawid and M. Gerstein, "A Bayesian system integrating expression data with sequence patterns for localizing proteins: Comprehensive application to the yeast genome," *J. Mol. Biol.*, vol. 301, pp. 1059–1075, 2000.
[49] F. C. Holstege *et al.*, "Dissecting the regulatory circuitry of a eukaryotic genome," *Cell*, vol. 95, pp. 717–728, 1998.
[50] V. E. Velculescu *et al.*, "Characterization of the yeast transcriptome," *Cell*, vol. 88, pp. 243–251, 1997.
[51] P. T. Spellman *et al.*, "Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization," *Mol. Biol. Cell.*, vol. 9, pp. 3273–3297, 1998.
[52] J. L. DeRisi, V. R. Iyer, and P. O. Brown, "Exploring the metabolic and genetic control of gene expression on a genomic scale," *Science*, vol. 278, pp. 680–686, 1997.
[53] C. Cortes and V. Vapnik, "Support vector networks," *Mach. Learning*, vol. 20, pp. 273–297, 1995.
[54] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler, "Support vector machine classification and validation of cancer tissue samples using microarray expression data," *Bioinformatics*, vol. 16, pp. 906–914, 2000.
[55] T. Joachims, *Making Large-Scale SVM Learning Practical. Advances in Kernel Methods—Support Vector Learning*, B. Schölkopf, C. Burges, and A. Smola, Eds. Cambridge, MA: MIT Press, 1999. [Online]. Available: http://www-ai.cs.uni-dortmund.de/DOKUMENTE/joachims_99a.pdf.
[56] A. Kumar, S. Agarwal, J. A. Heyman, S. Matson, M. Heidtman, S. Piccirillo, L. Umansky, A. Drawid, R. Jansen, Y. Liu, K. H. Cheung, P. Miller, M. Gerstein, G. S. Roeder, and M. Snyder, "Subcellular localization of the yeast proteome," *Genes Dev.*, vol. 16, pp. 707–719, 2002.
[57] R. J. Cho, M. J. Campbell, E. A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. G. Wolfsberg, A. E. Gabrielan, D. Landsman, D. J. Lockhart, and R. W. Davis, "A genome-wide transcriptional analysis of the mitotic cell cycle," *Mol. Cell.*, vol. 2, pp. 65–73, 1998.
[58] T. R. Hughes, M. J. Marton, A. R. Jones, C. J. Roberts, R. Stoughton, C. D. Armour, H. A. Bennett, E. Coffey, H. Dai, Y. D. He, M. J. Kidd, A. M. King, M. R. Meyer, D. Slade, P. Y. Lum, S. B. Stepaniants, D. D. Shoemaker, D. Gachotte, K. Chakraburtty, J. Simon, M. Bard, and S. H. Friend, "Functional discovery via a compendium of expression profiles," *Cell*, vol. 102, pp. 109–126, 2000.

Ning Lan received the B.S. degree in biochemistry from Peking University, Beijing, China, in 1995 and the Ph.D. degree in genetics from Duke University, Durham, NC, in 2000.

She is currently a Postdoctoral Fellow in the Department of Molecular Biophysics and Biochemistry at Yale University, New Haven, CT. Her postdoctoral research was supported by an NLM fellowship. Her research interests include genome annotation, biological database design, and integrative data analysis.



Ronald Jansen received the B.E. degree from Dartmouth College, Hanover, NH, in 1996, the M.S. degree in chemical engineering from the Technical University of Aachen, Aachen, Germany, in 1997, and the Ph.D. degree in computational genomics from Yale University, New Haven, CT in 2002.

He is currently a Postdoctoral Fellow in the Department of Molecular Biophysics and Biochemistry. His dissertation research was supported by an IBM Research Fellowship and has led to 16 publications in the bioscience literature. His main research interests are in the analysis of genome-wide expression data and protein–protein interactions.

Mark Gerstein received the A.B. degree in physics from Harvard College, Cambridge, MA, in 1989 and the Ph.D. degree in biophysics from Cambridge University, Cambridge, U.K., in 1993. He also did postdoctoral work at Stanford University, Stanford, CA.

He is an Associate Professor in the Department of Molecular Biophysics and Biochemistry at Yale University, New Haven, CT. He is also a joint appointee in the Department of Computer Science. He has published appreciably in biological science journals. His research interests include bioinformatics, and he is particularly interested in large-scale integrative surveys, biological database design, macromolecular genometry, molecular simulation, genome annotation, gene expression analysis, and Bayesian systems for data mining.

Dr. Gerstein has received a number of young investigator awards (e.g., from the Navy and Keck foundations).