Ontologies for proteomics -- Towards a systematic definition of structure & function that scales to the genome level:

Lan Ning[1], Gaetano T. Montelione[3] & Mark Gerstein[1, 2, †]

Departments of Molecular Biophysics & Biochemistry[1]
and Computer Science[2]
266 Whitney Avenue, Yale University
PO Box 208114, New Haven, CT 06520
(203) 432-6105, FAX (360) 838-7861

[3]Center for Advanced Biotechnology and Medicine
Department of Molecular Biology and Biochemistry
Rutgers University
Piscataway, NJ 08854-5638

lan@bioinfo.mbb.yale.edu
guy@cabm.rutgers.edu
Mark.Gerstein@Yale.edu

† Corresponding author

Abstract

A principle aim of structural and functional genomics is to elucidate the structures and functions of all the gene products in the genome. However, to adequately comprehend and analyze such a large amount of information we need new descriptions of proteins that scale to the genomic level. In short, we need a unified ontology for proteomics. Here we review progress towards this end, surveying the diverse approaches to systematic structural and functional classification and their progress towards developing standardized, unified descriptions for proteins. We focus particularly on systems to organize protein properties (both biophysical and biochemical) - as opposed to the classification of 3D protein folds, a subject has been reviewed extensively elsewhere. These systems are essential parts of the world-wide structural genomics effort. In relation to function, we survey the current classification approaches involving hierarchies, networks, and other graph structures (i.e. DAGs) and then describe a new approach to classification based on defining a protein's function through systematic enumeration of molecular interactions.

***Key words*: proteome; structure, function; ontology**

Introduction

After recent successes in genome-sequencing projects, the focus of large-scale biology has shifted from DNA to RNA and proteins, and the main challenge for bioinformatics is to integrate the ever growing amount of data to fully ascribe the biological role of proteins, cells, and ultimately, organisms [1]. Such task calls for the development of systematic systems describing how we should conceptualize and represent key information on proteins that can scale up to genomic level and be sufficiently standardized to support datamining (Fig. 1). These systematic descriptions go by the formal term of ontologies [2, 3]. Descriptions of protein structure and function, as well as the language used to describe experimental protocols in protein production, were originally crafted for individual proteins. These notions have progressed rapidly in recent years towards systematic representation, but are still isolated from each other, and are under intensive study and debate. In this paper we review some of the currently established representation systems of structural and functional genomics, and then describe a grid-like structure that defines protein function through molecular interactions.

Toward an Ontology for Structural Genomics

Structural genomics has emerged as one of the core areas of post-genomic studies [4-5]. It has major goals of helping in the determination of biochemical function for uncharacterized proteins and also in comprehensively surveying the range of folds adopted by proteins [6-9]. One of the main areas of ontological interest in structure genomics is defining a classification scheme for 3D protein folds. Classifying protein

folds has a number of important aspects, such as the possibility of doing this either manually or automatically via computer program. There has been considerable progress on this problem and there are currently a number of popular schemes organizing the protein structural universe including FSSP [10], CATH [11], and SCOP [12]. There have been quite a few recent reviews on this subject and we direct the reader to these for more details [1,13-16]. Here we wish to focus on other ontological issues raised by structural genomics, namely the systematic description of protein properties.

Structure determination requires a large number of experimental steps that go from cloning, expression, purification, biophysical characterization, to structure determination via NMR spectroscopy or X-ray crystallography. Traditionally the labor-intensive experimental structural biology had mainly been hypothesis driven and conducted on single-protein level. The success of the Human Genome Project has encouraged the construction of high-throughput pipelines aiming at characterizing proteins on a large-scale and eventually obtaining 3D protein structural information about them. Moreover, the highly variable characteristics of proteins make structural genomics projects fundamentally orders of magnitudes more complex than genome sequencing projects [17]. It is therefore essential that specifications or ontologies be developed to standardize the information about protein properties and make them amenable to retrospective analysis.

Across the world, several large-scale structural genomics projects have been initiated [18-20]. In the United States, nine pilot studies have been started under the NIH Protein Structure Initiative to develop and implement high-throughput technologies required for structural genomics [21]. Each of these centers is supported by database systems and underlining ontology structures. Here we use a database we created for one of the centers as an example to illustrate some key issues in developing specifications and ontologies for protein properties.

SPINE: An Integrated Tracking Database for the NESG

Northeast Structural Genomics Consortium (NESG) [22] is a multi-institutional collaboration emphasizing proteins from model eukaryotes. The consortium is geographically widespread, requiring a centralized repository to integrate and manage the data generated that are accessible to all the participating members. SPINE (Structural Proteomics in the Northeast) is the centralized tracking database for the consortium. As structural genomics is a new and rapidly evolving field, it was important to allow for database evolution to follow the development of the high-throughput process, rather than to take a top-down approach in which the database could restrict the development of the experimental technologies. A critical issue in designing a system of this kind is determining the fundamental “unit” to be tracked by the database. Initially the expression "construct" was chosen, and the best experimental results for the expression, purification, and characterization of each construct was recorded as attributes for this single entity. As the database expanded with more and more targets entered from various labs, it became obvious then than the primary objects being tracked were actually protein "targets", or the

proteins (or protein domains) themselves, each of which was being produced in multiple "construct" forms. Moreover, the need emerged to record experimental information on different levels, including not only the best conditions of cloning, expression, purification, etc., but also the sub-optimal ones such that future datamining can be conducted on multiple fronts. The improved database schema (Fig.2) better captures the work flow of the structural genomics pipeline at the NESG. In the current conceptualization, each "target" can be cloned into multiple "constructs", which are subsequently "expressed" under various fermentation conditions and then purified using multiple methods. The resulting protein "batches" are used for various biophysical characterizations (e.g., oligomerization state, monodispersity, crystallization, circular dichroism analysis) and structure determination by X-ray crystallography or NMR spectroscopy. The protein or nucleic acid (e.g. plasmid or cDNA) material generated at each step of the process is assigned a unique "Protein / DNA Sample ID", which is associated through the database with the complete history of the sample, as well as its specific storage location in the laboratory, reflecting the fact that the properties of each protein are contingent its particular preparation history. Each such sample is derived from a specific parent sample by a specific process, with one-to-many relationships from start to end. Relationships between samples (e.g., a set of plasmids within a 96-well plate), as well as the history of sample locations and transfers from one laboratory to another within the consortium, are also stored in these SPINE database records.

Information with disparate formats and types creates difficulty for data mining, therefore another key issue of the system is the standardization of experimental data sets.

Towards this end we introduced numerical values in place of the text descriptors sometimes used by experimentalists, as highlighted in Fig. 2. For a multi-institutional collaborative effort, it is important to accommodate the needs of various consortium projects where different experimental methodologies are used. Fields from existing data sets were used to develop a consensus of experimental parameters, which was in turn adapted to the current database framework. Using standardized solubility data from SPINE we were able to conduct decision tree analysis for optimization of target selection [23].

SPINS: Standardized ProteIn NMR Storage.

Another critical component of the structural genomics pipeline involve organizing the raw data, intermediate results, and final structure depositions into the public domain for each of hundreds of experimental structures determined by X-ray crystallography and NMR spectroscopy. Ontologies and databases for these processes, which will be invaluable to structural genomics and traditional structural biology projects, alike, are currently under active development [24-26]. An example of one such ontology and database is SPINS (Standardized ProteIn NMR Storage), a data dictionary and object-oriented relational database for archiving protein NMR spectra [27]. Modern protein NMR spectroscopy laboratories have a rapidly growing need for an easily queried local archival system of raw experimental NMR datasets. SPINS is an object-oriented relational database that provides facilities for high-volume NMR data archival, organization of analyses, and dissemination of results to the public domain by automatic preparation of the header files required for submission of data to the BioMagResBank

(BMRB). SPINS coordinates the process from data collection to BMRB deposition of raw NMR data by standardizing and integrating the storage and retrieval of these data in a local laboratory file system. SPINS also includes a user-friendly internet-based graphical user interface, which is integrated with certain NMR data collection software. To ensure smooth integration of SPINS data into the NMRStar format used by the BMRB, efforts are made to keep the SPINS data model as consistent as possible with the related and partially overlapping NMRStar data model [28], as well as with the evolving CCPN data dictionary and model of experimental NMR data [26, 29].

SPINS v1.0 and its associated data dictionary represent the first phase of a multi-phase process integration project, providing organization, archiving, and simple submission to the BMRB of time domain FID files and all the information needed to describe and reproduce these data. Relatively few raw NMR data sets (FIDS) are currently available in the public domain and routine archiving of such data using tools like SPINS will have significant scientific value. Through the activities of the NESG, SPINS is evolving into a central agent which integrates the entire process of NMR-based protein structure determination. As the protein spectroscopist progresses through the resonance assignment and structure determination process, the evolving SPINS database serves as the central archive, logging important information critical for documenting and reproducing each step of the NMR data analysis process, and generating intermediate files in appropriate formats for the supported specific software applications, forming the core of an automated data analysis process. The SPINS data dictionary is also designed to be consistent with evolving structural genomics project databases, such as SPINE.

Finally, SPINS will also capable of auto-submission of the associated intermediate and final data files generated in the process of NMR resonance assignments and structure analysis to the public domain BioMagResBank [25] and Protein Data Bank [30] in a fully validated format. Similar efforts are in progress for NMR data organization by the CCPN Network [26, 29], and for X-ray crystallographic data organization by the PHENIX project [24].

Other Efforts to Standardized Structural Genomics Information

There are a number of prominent efforts to standardize structural genomics information. Currently the information from various centers and labs is highly scattered. Before solved structures are deposited into the Protein Data Bank (PDB), there need to be coordination of the selection and production progress of protein targets in order to minimize the waste of resources and efforts on the overlapping target pools that have been identified by the various centers. For this purpose, TargetDB [31] was created as a target registration database, originally for registration and tracking information for NIH P50 structural genomics centers, and later expanded to include target data from worldwide structural genomics and proteomics projects. Participating centers provide status and tracking information on the progress of their targets in XML format based on the Document Type Definition (DTD) defined by TargetDB, which in turn provides display and query interface to the target information. The minimal contents of TargetDB was defined by a subcommittee of the International Structural Genomics Organization so as to allow for its rapid implementation, though in the future the range of structural

genomics information consolidated in TargetDB and shared across the world-wide structural genomics community is envisioned to greatly expand.

Another structural genomics database aimed at improving communication in the field by providing a repository of project progress information is PRESAGE (Protein Resource Entailing Structural Annotation of Genomic Entities) [32]. The fundamental unit in PRESAGE is an annotation, either experimental or prediction, with subsidiary and additional varieties, as submitted by researchers worldwide.

The PDB as the sole repository of macromolecular structural data is aiming to validate all data in its archive and release a uniform data format as well as a guideline for future deposition, so as to facilitate systematic analyses and integration with other biological and structural databases. The depository is also expanding their data dictionary to include a comprehensive experimental data collection and refinement that previously were embedded in unstructured REMARK records in PDB files [33].

Toward an Ontology for Functional Genomics

It is important to recognize that functional characterization must be done on several levels. A particular gene product can be characterized with respect to its genetic or physiological function (e.g., expression of the gene product codes for a particular cellular fate or lineage), a cellular function (e.g., the protein regulates microtubule assembly), the biochemical function (e.g., the protein is a kinase), and the biophysical function (e.g., the

details of individual residue pKa's and local electrostatic potential in determining the specificity and mechanism of phosphorylation). Solving the structure of a protein can often provide valuable clues towards elucidating its biochemical and biophysical function, and there numerous examples in structural genomics of how having a structure suggested a function for an uncharacterized protein [34-37].

However, structure rarely provides deep insights into genetic, physiological, or cellular function. These other levels of protein function can also be characterized by various high-throughput functional genomics methods, using oligonucleotide and cDNA microarrays, gene disruption through transposon insertion [38] or deletion [39,-40], yeast two hybrid assays [41-44], proteome microarrays [45-47] and the TAP-tagging method [48, 49], or through homology-based annotation transfer based on the idea that proteins of similar sequence and structure are presumably descended from a common ancestral protein, and have related functions [50-52]. 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. For protein pairs that share the same fold, usually 30-40 % sequence identity is required for function to be conserved [53-54]. Examples also exist where proteins of high sequence and structural similarity perform disparate functions, such as lysozyme and α-lactalbumin; or proteins with different structural folds have identical function, such as subtilisin and chymotrypsin [55].

Systematic representation of protein function

Early functional annotation tended to be recorded as simple phrases, which are nonstandard, highly unstable, and have no organized structure among functions. Many humorous examples can be taken from the fly (e.g. Redtape, roadblock, starry night) [56]. Moreover, function has been described from different angles dependent on the experimental perspective. Biochemists often characterize protein function in terms of molecular reactions. 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.

One approach is the hierarchical representation adopted by most functional ontologies such as the Gene Ontology (GO)[57], the MIPS Functional Classification Catalogue [58] and the Enzyme Commission (EC) classification [59]. Fig. 3a shows a simplified hierarchy in [26] to represent enzyme and non-enzyme function. Sharing of classification numbers indicates 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 Gene Ontology Consortium has been highly successful in creating a structured and precisely defined controlled vocabulary for describing gene function

across several organisms [57]. GO classifies genes into three parallel categories, i.e. three directed acyclic graphs: biological process, molecular function and cellular component. This is for defining the function of a gene at various levels, including its biochemical activities, biological roles as well as cellular structure. Nodes can often be reached from multiple paths, which facilitates the representation and comparison of genes with multiple functions or involved in more than one process. GO Consortium has recently launched an umbrella web site called GOBO (Global Open Biological Ontologies) for structured shared controlled vocabularies for use within the genomics and proteomics domain [60].

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.

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 [61]. 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.

Protein-protein interaction maps represent a population of interacting proteins displayed as networks or circuits. An example is shown in Fig. 3b. 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 [42]. The more recently developed proteome microarray technology allows for direct analysis of a variety of interactions, including interactions between proteins [45-47]. Protein interactions have also been predicted by computational methods based on genomic sequence [62] or mRNA expression [63]. We found that gene expression data is sometimes more meaningful when they are grouped under a protein complex scheme rather then a functional classification scheme [64].

Protein-protein interaction maps have not only confirmed the existence of previously known complexes and pathways but also shed light on the discovery of new complexes and crosstalk between previously unlinked pathways [43, 65]. 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 such that their respective orthologs in other organisms also interact [66].

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. 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 the answer question of functional equality more precisely, one needs to integrate functional information from a variety of resources including pathway and interaction maps.

For two non-identical but related functions, the degree of similarity is much harder if not impossible to answer using natural language-based ontologies. When comparing two GO terms, their names and positions in the hierarchy often do not provide full information on the level of similarity between them. 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 function may only be meaningful 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 [67].

Consider the following more complex questions such as: 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 potentially be easy to solve if function were represented in numerical form. Here we describe a grid-like structure that represents protein function in term of interaction probabilities (recently proposed in [64]) and discuss its potential application in function prediction.

Construction and potential application of the function grid

In the 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 (Fig. 4a). 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.

There are several issues that 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. 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 (Fig. 4b). Third, when information on

molecular interaction from multiple organisms is collected, how are we going to integrate them, i.e. should homologs be treated as different fields of the same protein or different proteins? One reasonable decision is to construct individual matrices for each organism, and keep evolutionary relations between homologs in another table. This way the similarity and difference between interaction partners among homologs 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. 4c-d shows how two steps in the MAP kinase pathway involved in the maintenance of cellular integrity [42] are represented in the interaction grid.

The interaction grid can be combined with sequence and structural features, 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.

Conclusion

The availability of fully sequenced genomes challenges bioinformatics to elucidate the structure, interactions and functions of proteins on genomic scale. Ontology systems are needed that can facilitate calculation of function together with other biological data. Such ontologies should aim at capturing all dimensions of protein structure and function and should keep up with the phenomenal rate at which biological data are being produced. Although there is a well-studied link between protein structure and function [55, 68], the

ontologies adopted in the each field, such as CATH that describes protein folds, SPINE that delineates biophysical characteristics, and GO that represents function, have been developed separately and remain largely isolated. One potential connection point would be the description of active sites in protein structure that illustrates its function. Current structural and 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. Future progress in this field is likely towards increased uniformity, more refined data structure and higher level of standardization to support datamining.

References

[1] Thornton JM: **From Genome to Function.** *Science*, 2001. **292**: 2095-2097.

[2] Genesereth M, Nilsson N. **Logical Foundations of Artificial Intelligence.** Morgan Kaufmann; 1987.

[3] Gruber TR: **A translation approach to portable ontologies.** *Knowledge Acquisition*, 1993. **5**:199-220.

[4] Zarembinski TI, et al.: **Structure-based assignment of the biochemical function of a hypothetical protein: a test case of structural genomics.** *Proc Natl Acad Sci USA* 1998, **95**: 15189-15193.

[5] Cort JR, Koonin EV, Bash PA, Kennedy MA: **A phylogenetic approach to target selection for structural genomics: solution structure of YciH.** *Nucl Acids Res* 1999, **27**:4018-4027.

[6] Montelione GT, Anderson S: **Structural genomics: keystone for a Human Proteome Project.** *Nat Struct Biol,* 1999. **6**:11-12.

[7] Burley SK, et al.: **Structural genomics: beyond the human genome project.** *Nat Genet,* 1999. **23**:151-157.

[8] Montelione GT: **Structural genomics: an approach to the protein folding problem.** *Proc Natl Acad Sci USA*, 2001. **98**: 13488-13489.
* The author reviews recent progress of genomic-scale three-dimensional (3D) protein structure determination and its role in elucidating the protein folding problem.

[9] Vitkup D, Melamud E, Moult J, Sander C: **Completeness in structural genomics.** *Nat Struct Biol* 2001, **8**:559-566.

[10] Holm L, Sander C: **The FSSP database: fold classification based on structure-structure alignment of proteins.** *Nucleic Acids Res,* 1996. **1**:206-209
** The authors describe the FSSP database which is a c classification of 3D protein folds based on an all-against-all comparison of structures currently in the PDB.

[11] Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM: **CATH--a hierarchic classification of protein domain structures.** *Structure* 1997 5:1093-108
** The authors describe recent developments to the CATH domain database of protein structural families along with other protein family resources, and also reveal important caveats in transferring functional data between homologous proteins.

[12] Hubbard TJ, Ailey B, Brenner SE, Murzin AG, Chothia C: **SCOP: a Structural Classification of Proteins database.** *Nucleic Acids Res,* 1999. **27**:254-256.

** The authors describe in detail the hierarchical levels of SCOP and its application as a source of data to calibrate sequence search algorithms and for the generation of population statistics on protein structures.

[13] Hadley C, Jones DT: **A systematic comparison of protein structure classifications: SCOP, CATH and FSSP.** *Structcture Fold Des,* 1999. **7**:1099-1112.
* The authors compare the three most widely used and comprehensive databases systematically to determine their overall agreement in classifying protein structures. They describe the advantages each method offers depending on biological requirements.

[14] Thornton JM, Todd AE, Milburn D, Borkakoti N, Orengo CA: **From structure to function: approaches and limitations.** *Nat Struct Biol,* 2000. **Suppl:** 991-994

[15] Teichmann SA, Murzin AG, Chothia C: **Determination of protein function, evolution and interactions by structural genomics.** *Curr Opin Struct Biol,* 2001. **11**: 354-363.
* The authors review recently efforts on the determination of the function and evolutionary relationships of proteins by experimental structural genomics and the discovery of protein-protein interactions by computational structural genomics.

[16] Brenner SE, Levitt M: **Expectations from structural genomics.** *Protein Sci,* 2000. **9:**197-200.

[17] Stevens RC, Yokoyama S, Wilson IA: **Global Efforts in Structural Genomics.** *Science* 2001, **294**:89-92.
* The authors describe the worldwide initiative in structural genomics and highlight the Second International Structural Genomics Meeting of 2001.

[18] Heinemann U: **Structural genomics in Europe: slow start, strong finish?** *Nat Struct Biol,* 2000. **Suppl:** 940-942

[19] Terwilliger TC: **Structural genomics in North America.** *Nat Struct Biol,* 2000. **Suppl:** 935-939.

[20] Yokoyama S, et al.: **Structural genomics projects in Japan.** *Nat Struct Biol,* 2000. **Suppl:** 943-945.

[21] Burley FC, Bonanno JB: **Structuring the universe of proteins**. *Annu Rev Genomics Hum Genet* 2002, **3**:243-262.

[22] http://www.nesg.org

[23] Bertone P et al.: **SPINE: an integrated tracking database and data mining approach for identifying feasible targets in high-throughput structural proteomics**. *Nucleic Acids Res*.2001, **29**: 2884-2898.

** The authors describe a structural genomics project tracking database specifically designed to enable distributed scientific collaboration via the Internet as well as an active vehicle to standardize proteomics data in a form that would enable systematic data mining.

[24] Adams PD, et al.: **PHENIX: building new software for automated crystallographic structure determination.** *Acta Crystallogr D Biol Crystallogr,* 2002. **58**:1948-1954.

[25] Seavey BR, Farr EA, Westler WM, Markley JL: **A relational database for sequence-specific protein NMR data.** *J. Biomol. NMR*, 1991. **1**:217-236

[26] Fogh R: **The CCPN project: an interim report on a data model for the NMR community.** *Nat Struct Biol,* 2002. 9:416-418.

[27] Baran M, Moseley HNB, Sahota G, Montelione GT: **SPINS: A data dictionary and object-oriented relational database for archiving protein NMR spectra.** *J. Biomol. NMR*, 2002. **24**: 113-121.
** The authors describe SPINS, an object-oriented relational database that provides facilities for high-volume NMR data archival, organization of analyses, and automatic submission of results to the public domain.

[28] http://www.bmrb.wisc.edu/

[29] http://www.bio.cam.ac.uk/nmr/ccp/

[30] Berman HM, et al.: **The Protein Data Bank.** *Nucleic Acids Res* 2000. **28**:235-242.
* The authors describe the short-term and long-term goals of the PDB, the data deposition systems, and how to obtain further information.

[31] http://targetdb.rutgers.edu/index.html
** A centralized registration database created by the PDB for target sequences from the nine P50 NIH structural genomics centers and from other genomics centers worldwide, updated weekly.

[32] Brenner SE, Barken D and Levitt M: **The PRESAGE database for structural genomics.** *Nucleic Acids Res.* 1999, **27**: 251-253
* The authors describe the PRESAGE database as a collaborative resource to which researchers add annotations indicating current experimental status, structural predictions and suggestions, aimed at enhancing communication among structural genomics researchers.

[33] Westbrook J, et al.: **The Protein Data Bank: unifying the archive.** *Nucl Acids Res* 2002, **30**: 245-248.
** The authors describe validation process of all data in the PDB archive and the release of a uniform archive for the structural genomics community.

[34] Newkirk K, et al.: **Solution NMR structure of the major cold shock protein (CspA) from Escherichia coli: Identification of a binding epitope for DNA.** *Proc Natl Acad Sci USA* 1994, **91**: 5114-5118.

[35] Feng W, Tejero R, Zimmerman DE, Inouye M, Montelione GT: **Solution NMR structure and backbone dynamics of the major cold shock protein (CspA) from Escherichia coli: Evidence for conformational dynamics in the proposed ssRNA-binding site.** *Biochemistry,* 1998, **37:** 10881 - 10896.

[36] Cort JR, Chiang Y, Zheng D, Montelione GT, Kennedy MA: **NMR structure of conserved eukaryotic protein ZK652.3 from *C. elegans*: A ubiquitin-like fold.** *Proteins: Struct. Funct. Genetics,* 2002. **48**: 733-736.

[37] Yang H, et al.: **BRCA2 function in DNA binding and recombination from a BRCA2-DSS1-ssDNA structure.** *Science,* 2002. **297**:1837-48

[38] Ross-Macdonald P: **Large-scale analysis of the yeast genome by transposon tagging and gene disruption.** *Nature,*1999. **402**:413-418.

[39] Ni L and Snyder M: **A Genomic Study of the Bipolar Bud Site Selection Pattern in *Saccharomyces cerevisiae*.** *Mol. Biol. Cell*, 2001. **12**:2147-2170.

[40] Winzeler EA et al.: **Functional characterization of the *Saccharomyces cerevisiae* genome by comprehensive and precise gene deletion and massively parallel analysis.** *Science,* 1999. **285**:901-906.

[41] Ito T, 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.** *Proc Natl Acad Sci USA*, 2000. **97**: 1143-1147

[42] Schwikowski B, Uetz P and Fields S: **A network of protein-protein interactions in yeast.** *Nat. Biotechnol*., 2000. **18**: 1257-1261.

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

[44] Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M and Sakaki Y: **A comprehensive two-hybrid analysis to explore the yeast protein interactome.** *Proc Natl Acad Sci USA,* 2001. **98**:4569-4574.

[45] MacBeath G and Schreiber SL: **Printing Proteins as Microarrays for High Throughput Function Determination.** *Science,* 2000. **289**: 1760-1762.

[46] Zhu H et al.: **Analysis of yeast protein kinases using protein chips**. *Nat Genet,* 2000. **26**:283-289.

[47] Zhu H et al.: **Global analysis of protein activities using proteome chips.** *Science,* 2001. **293**: 2101-2105.

[48] Gavin AC et al.: **Functional organization of the yeast proteome by systematic analysis of protein complexes.** *Nature*, 2002. **415**:141-147.

[49] Ho Y et al.: **Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry.** *Nature*, 2002. **415**:180-183.

[50] Bork P, Koonin EV: **Protein sequence motifs.** *Curr Opin Struct Biol*, 1996. **6**:366-376.

[51] Zhang Z, et al. **Protein sequence similarity searches using patterns as seeds.** *Nucl Acids Res,* 1998. **26**: 3986-3990.

[52] Attwood TK, et al., **PRINTS prepares for the new millennium.** *Nucl Acids Res,* 1999. **27**:220-225.

[53] Hegyi H and Gerstein M: **Annotation Transfer for Genomics: Measuring Functional Divergence in Multi-domain Proteins.** *Genome Res.,* 2001. **11**:1632-1640.

[54] Wilson CA, Kreychman J and Gerstein M: **Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores.** *J Mol Biol,* 2000. **297**:233-249.

[55] Todd AE, Orengo CA and Thornton JM: **Evolution of function in protein superfamilies, from a structural perspective.** *J Mol Biol,* 2001. **307**: 1113-1143
* The authors explore the functional variation of homologous enzyme superfamilies containing two or more enzymes and find that almost all superfamilies exhibit functional diversity generated by local sequence variation and domain shuffling.

[56] Vacek M: **A Gene by Any Other Name.** *American Scientist,* 2001. **89.**

[57] Ashburner M, et al.: **Gene ontology: tool for the unification of biology.** *Nat Genet,* 2000. **25**:25-29.
** The authors describe GO, a structured and precisely defined controlled vocabulary for describing gene function across several organisms. GO consists of three DAGs which define gene function at various levels, including its biochemical activities, biological roles as well as cellular structure.

[58] Mewes HW, et al.: **MIPS: a database for genomes and protein sequences.** *Nucleic Acids Res,* 2000. **30**:31-34.

[59] Webb EC: **Enzyme Nomenclature 1992. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology.** *Academic Press New York,* 1992.

[60] http://www.geneontology.org/doc/gobo.html

[61] Karp, PD: **An ontology for biological function based on molecular interactions.** *Bioinformatics,* 2000. **16:** 269-285.
* The author describes the functional ontology developed for the EcoCyc database which encodes a diverse array of biochemical processes. The ontology is validated through its use to support complex functional queries for the EcoCyc DB.

[62] Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D: **Detecting Protein Function and Protein-Protein Interactions from Genome Sequences.** *Science,* 1999. **285**:751-753.

[63] Jansen R, Greenbaum D, Gerstein M, **Relating whole-genome expression data with protein-protein interactions.** *Genome Res*, 2002. **12**:37-46.
* The authors investigate the relationship of protein-protein interactions with mRNA expression levels, by integrating a variety of data sources for yeast and find that subunits of the same protein complex show significant coexpression, both in terms of similarities of absolute mRNA levels and expression profiles.

[64] Lan N, Jansen R, Gerstein M: **Towards a systematic definition of protein function that scales to the genome level: Defining function in terms of interactions.** *Proc IEEE,* in press.

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

[66] Matthews LR et al. **Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or interologs.** *Genome Res*, 2001. **11**:2120-2126.

[67] Antonini E and Brunoni M: **Hemoglobin and myoglobin in their reactions with ligands.** Amsterdam, Holland: Borth-Holland; 1971.

[68] Hegyi H and Gerstein M: **The relationship between protein structure and function: a comprehensive survey with application to the yeast genome.** *J Mol Biol,* 1999. **288**: 147-164.
* The authors systematically investigate the relationship between protein function and structure and find that the major SCOP fold classes have different propensities to carry out certain broad categories of functions.

[69] Kumar A, et al.: **Subcellular localization of the yeast proteome.** *Genes Dev,* 2002. **16**:707-719.

[70] Cho R, et al.: **A genome-wide transcriptional analysis of the mitotic cell cycle**. *Mol Cell,* 1998. **2**: 65-73.

[71] Hughes T, et al.: **Functional discovery via a compendium of expression profiles.** *Cell,* 2000. **102**:109-126.

Fig.1. History, current status and future perspective of protein ontology

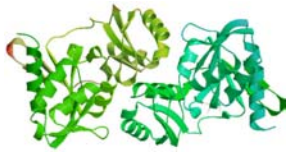
Early descriptions of protein structure, function and biophysical properties consist of natural phrases as well as experimental data in diverse format. Genome-scale representation systems in the form of hierarchical structure, directed acyclic graph or grid-like structure is being developed for structural and functional proteomics, respectively. Future progress in this field is likely towards a unified system with higher level of standardization to support datamining.

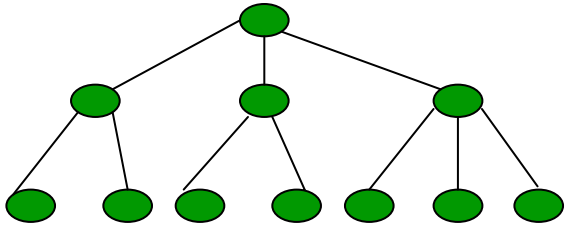
CHAOS

CURRENT

FUTURE

Structure





Function

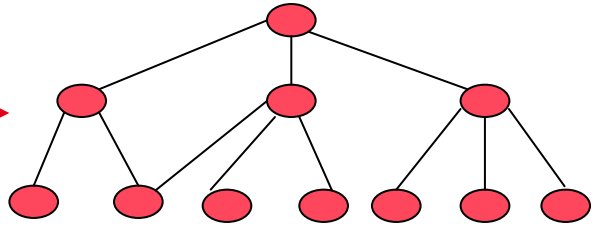
actin

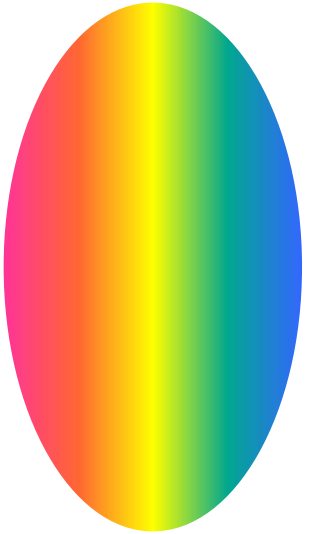
protein kinase

DNA binding

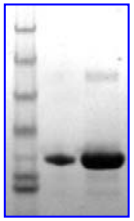
Involved in apoptosis

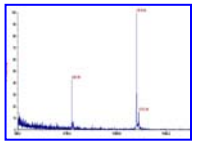
histone synthetic lethal

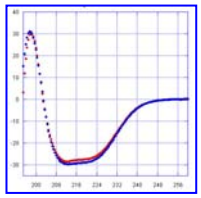


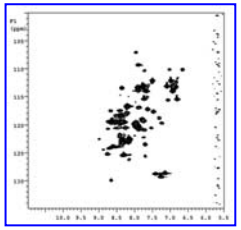


Properties









	expression	solubility	oligomer	HSQC	.....
protein 1	8	5	dimer	promising	.....
protein 2	6	3	monomer	good	.....
protein 3	9	3	dimer	aggregated	.....
protein 4	5	0	monomer	-	.....
protein 5	0	-	-	-	.....
protein 6	8	4	tetramer	good	.....
.....	.....	.....	.....	.....	.....

Fig. 2. Simplified schema of SPINE database

The major tables in SPINE and their relationship are shown. Some of the key attributes are listed. Standardized fields using numerical values are in red, while fields using controlled vocabulary are in green. Shaded fields are those we used in datamining of SPINE [23]

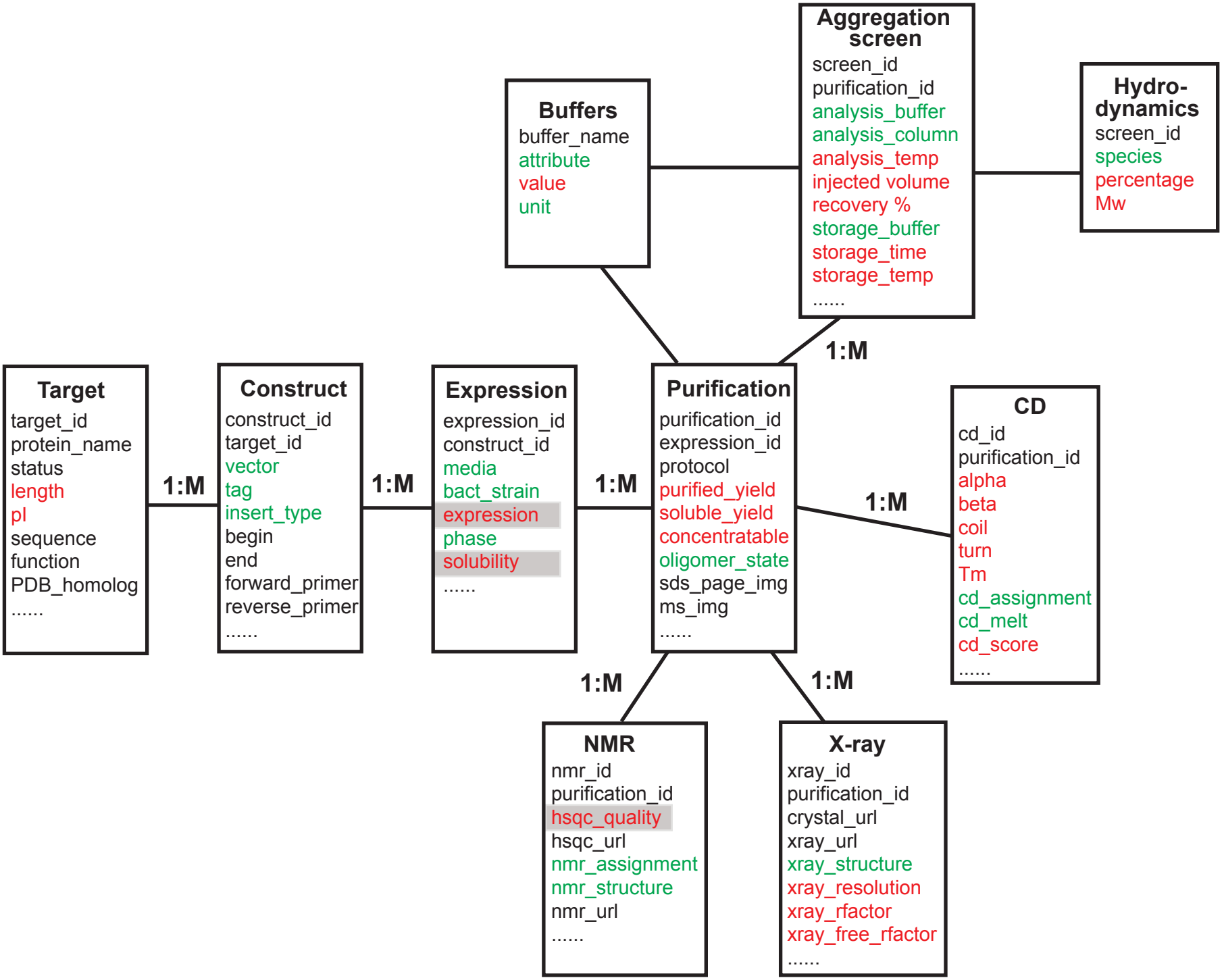
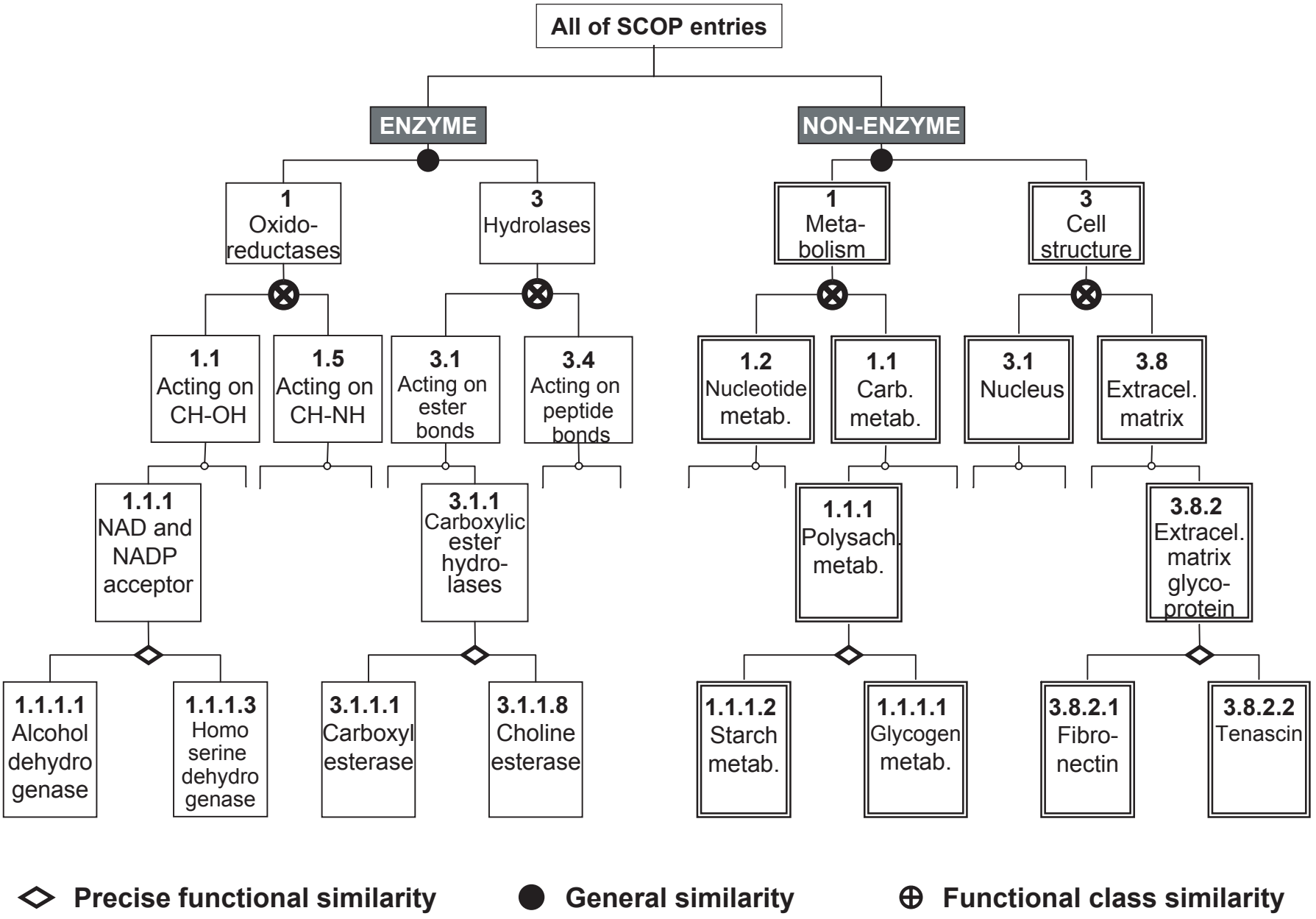

Target
target_id
protein_name
status
length
pI
sequence
function
PDB_homolog
.....
1:M
Construct
construct_id
target_id
vector
tag
insert_type
begin
end
forward_primer
reverse_primer
.....
1:M
Expression
expression_id
construct_id
media
bact_strain
expression
phase
solubility
.....
1:M
Buffers
buffer_name
attribute
value
unit
Aggregation screen
screen_id
purification_id
analysis_buffer
analysis_column
analysis_temp
injected volume
recovery %
storage_buffer
storage_time
storage_temp
.....
Hydro-dynamics
screen_id
species
percentage
Mw
1:M
Purification
purification_id
expression_id
protocol
purified_yield
soluble_yield
concentratable
oligomer_state
sds_page_img
ms_img
.....
1:M
CD
cd_id
purification_id
alpha
beta
coil
turn
Tm
cd_assignment
cd_melt
cd_score
.....
1:M
NMR
nmr_id
purification_id
hsqc_quality
hsqc_url
nmr_assignment
nmr_structure
nmr_url
.....
1:M
X-ray
xray_id
purification_id
crystal_url
xray_url
xray_structure
xray_resolution
xray_rfactor
xray_free_rfactor
.....

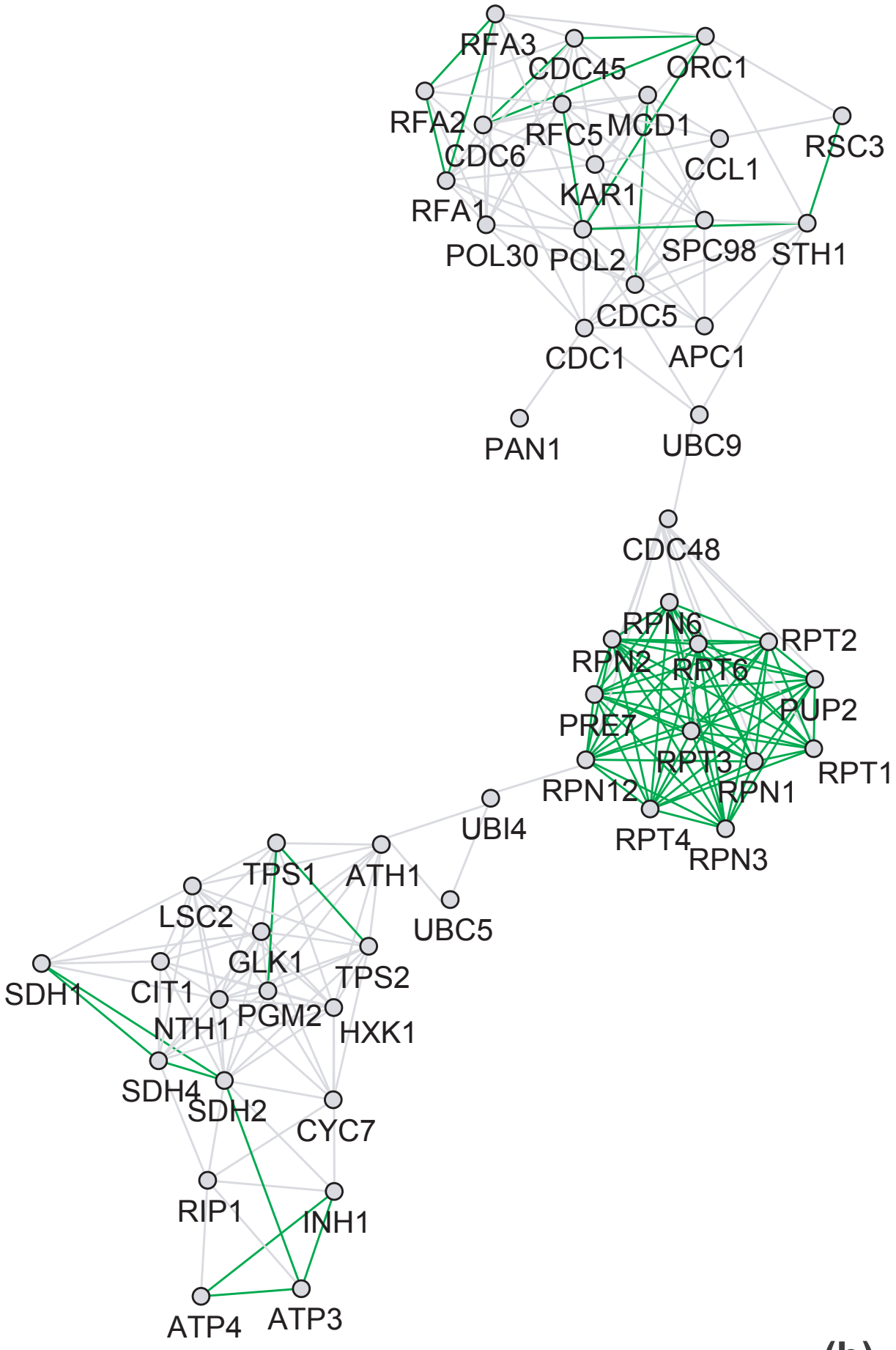
Fig. 3. Systematic representation of protein function

(a). Hierarchical scheme for functional classification, adapted from ref. 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 non-enzymes, 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 non-enzyme numbers, depending on category) then they have the same precise function.

(b) Example of a yeast protein network. The green edges represent protein-protein interactions from the MIPS complexes catalog [58], two yeast-two hybrid datasets [43-44], and two in-vivo pull-down datasets [48-49]. 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 [69-71].


All of SCOP entries
ENZYME
NON-ENZYME
1 Oxido-reductases
3 Hydrolases
1 Meta-bolism
3 Cell structure
1.1 Acting on CH-OH
1.5 Acting on CH-NH
3.1 Acting on ester bonds
3.4 Acting on peptide bonds
1.2 Nucleotide metab.
1.1 Carb. metab.
3.1 Nucleus
3.8 Extracel. matrix
1.1.1 NAD and NADP acceptor
3.1.1 Carboxylic ester hydro-lases
1.1.1 Polysach. metab.
3.8.2 Extracel. matrix glyco-protein
1.1.1.1 Alcohol dehydro genase
1.1.1.3 Homo serine dehydro genase
3.1.1.1 Carboxyl esterase
3.1.1.8 Choline esterase
1.1.1.2 Starch metab.
1.1.1.1 Glycogen metab.
3.8.2.1 Fibro-nectin
3.8.2.2 Tenascin
Precise functional similarity
General similarity
Functional class similarity

(a)


RFA3
CDC45
ORC1
RFA2
RFC5
MCD1
RSC3
CDC6
CCL1
KAR1
RFA1
POL30
POL2
SPC98
STH1
CDC5
CDC1
APC1
PAN1
UBC9
CDC48
RPN6
RPT2
RPN2
RPT6
PUP2
PRE7
RPT3
RPT1
RPN12
RPN1
RPT4
RPN3
UBI4
TPS1
ATH1
LSC2
UBC5
GLK1
TPS2
SDH1
CIT1
PGM2
NTH1
HXK1
SDH4
SDH2
CYC7
RIP1
INH1
ATP4
ATP3

(b)

Fig. 4. Functional grid and its application in functional prediction

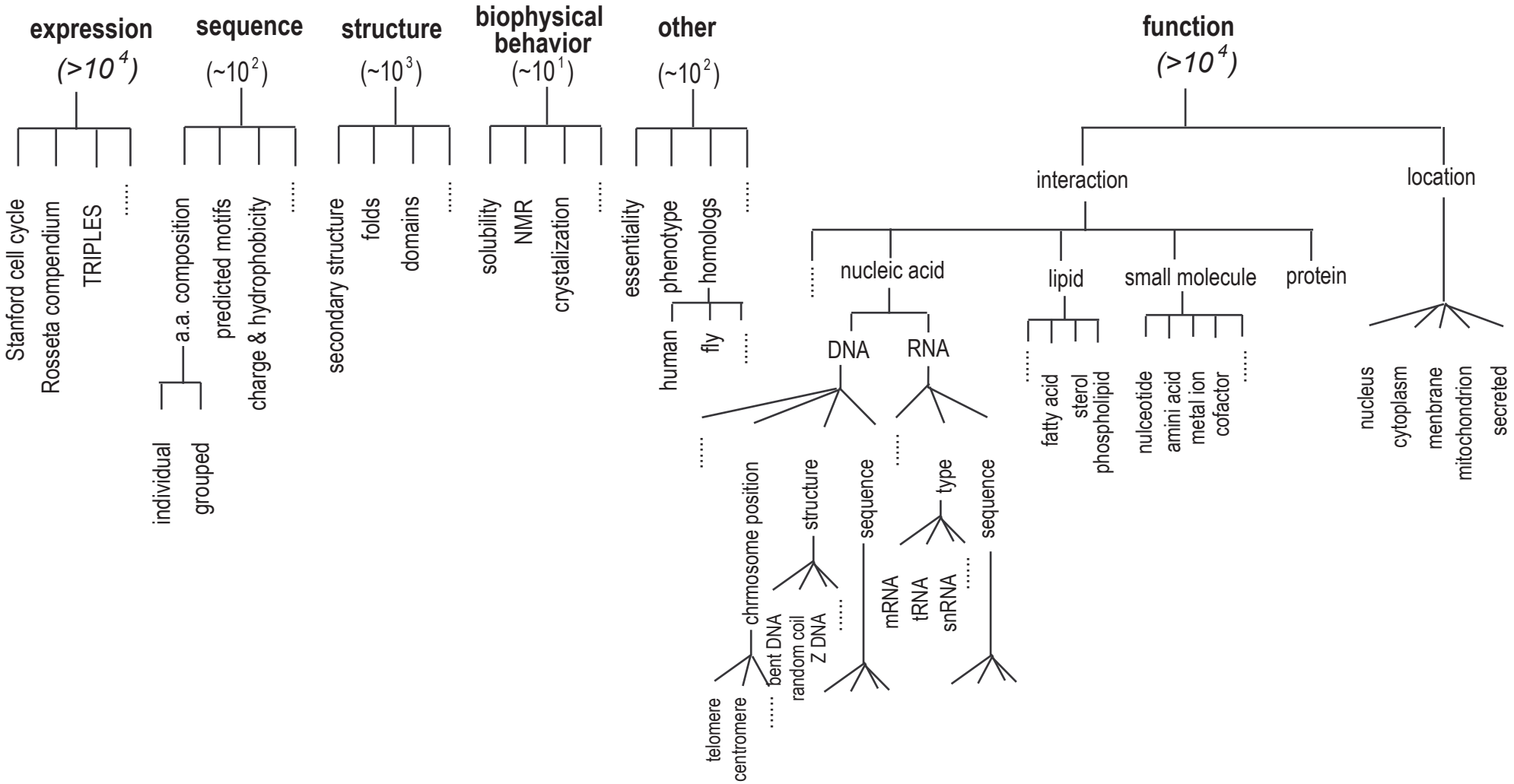
(a). A simplified example of Interaction Grid. The function of each protein is defined as a row vector that consists of the probability of binding to various ligands. The grid is filled with data collected from GO, EC, yeast two-hybrid system interactions and proteome chip experiments. For information 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 (TAS & IDA). Using the data from proteome chip experiments, we define the binding probability of each protein by normalizing its binding signal against the lowest value of all proteins that are known to bind the ligand. 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.

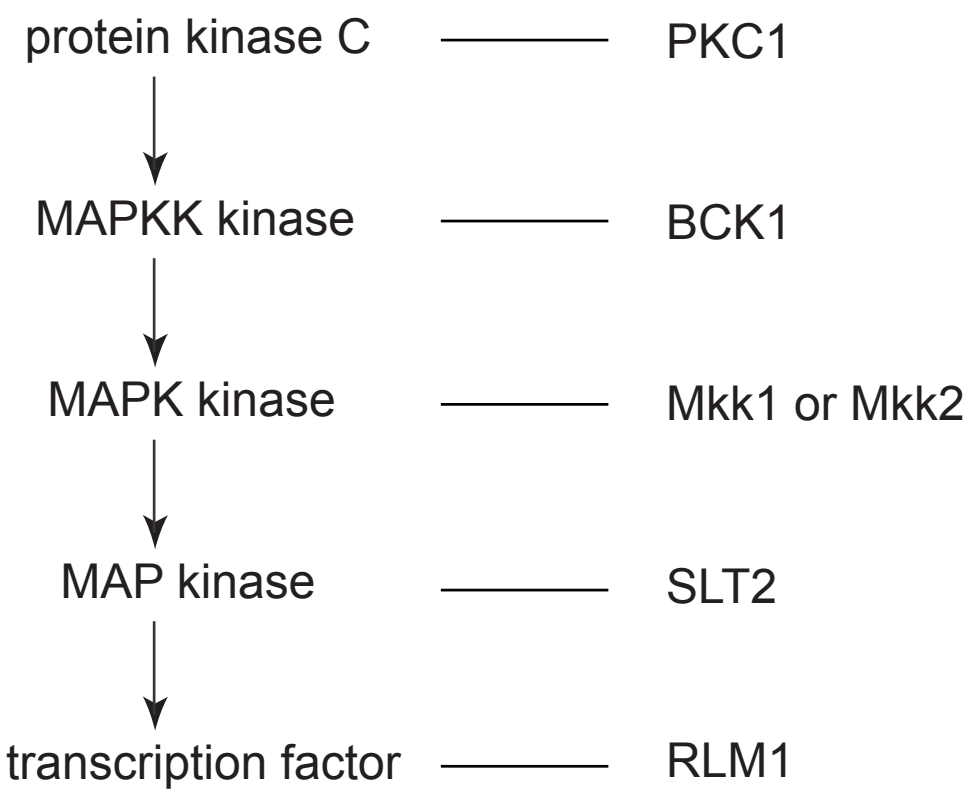
(b). Hierarchical organization of the functional grid. The fields in the functional grid can be grouped into a hierarchical structure, such that the data mining can be performed at various levels. The range of potential number of fields (columns) for each group is indicated in parentheses. Areas where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

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

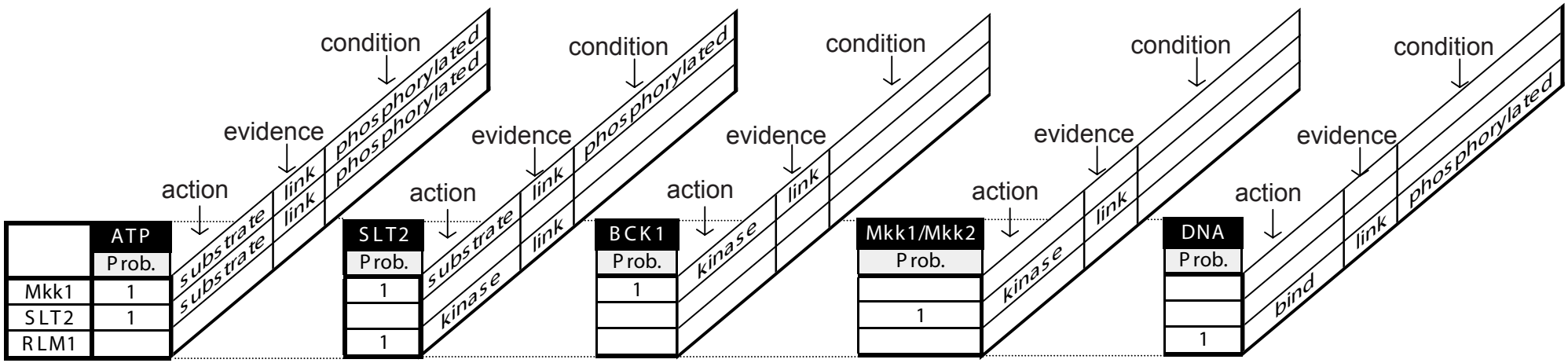
	nucleic acids		small molecules					proteins			
	DNA	**RNA**	**ATP**	**Metal**	**CoA**	**NAD**	**……**	**G protein**	**CDC28**	**Calmodulin**	**……**
protein 1	1.0	0	0	0	0	0	……	0	0	0	……
protein 2	0	0.9	0	0	0	0	……	0	0	0	……
protein 3	1.0	0	1.0	0	0	0	……	0	0	0	……
protein 4	0	0	0	0	0.8	0	……	0	0	1.0	……
protein 5	1.0	0	0	0	0	0	……	0	0.9	0	……
protein 6	0.9	0					……				……
protein 7	0	0.8					……				……
……	……	……	……	……	……	……	……	……	……	……	……

(a)


expression
(>10^4)
Stanford cell cycle
Rosseta compendium
TRIPLES
.....
sequence
(~10^2)
a.a. composition
individual
grouped
predicted motifs
charge & hydrophobicity
.....
structure
(~10^3)
secondary structure
folds
domains
.....
biophysical behavior
(~10^1)
solubility
NMR
crystalization
.....
other
(~10^2)
essentiality
phenotype
homologs
.....
human
fly
.....
function
(>10^4)
interaction
location
.....
nucleic acid
lipid
small molecule
protein
DNA
RNA
.....
fatty acid
sterol
phospholipid
nulceotide
amini acid
metal ion
cofactor
.....
nucleus
cytoplasm
menbrane
mitochondrion
secreted
.....
chrmosome position
structure
sequence
.....
type
sequence
telomere
centromere
.....
bent DNA
random coil
Z DNA
.....
mRNA
tRNA
snRNA
.....


protein kinase C
PKC1
MAPKK kinase
BCK1
MAPK kinase
Mkk1 or Mkk2
MAP kinase
SLT2
transcription factor
RLM1

(a)


condition
evidence
action
phosphorylated
phosphorylated
link
link
substrate
substrate
ATP
Prob.
Mkk1
1
SLT2
1
RLM1
SLT2
Prob.
1
1
BCK1
Prob.
1
Mkk1/Mkk2
Prob.
1
DNA
Prob.
1
kinase
bind

(b)