Integrative Analysis of Protein Interaction Data

M. Fellenberg[1,2], K. Albermann[1], A. Zollner[1], H.W. Mewes[2], and J. Hani[1]

[1]Biomax Informatics GmbH
[2]MIPS/GSF, Max-Planck-Institute f. Biochemistry
82152 Martinsried, Germany
jean.hani@biomax.de
Phone +49 89 895574-0, Fax +49 89 895574-25

Abstract

We have developed a method for the integrative analysis of protein interaction data. It comprises clustering, visualization and data integration components. The method is generally applicable for all sequenced organisms. Here, we describe in detail the combination of protein interaction data in the yeast *Saccharomyces cerevisiae* with the functional classification of all yeast proteins. We evaluate the utility of the method by comparison with experimental data and deduce hypotheses about the functional role of so far uncharacterized proteins. Further applications of the integrative analysis method are discussed. The method presented here is powerful and flexible. We show that it is capable of mining large-scale data sets.

Introduction

The knowledge of the whole genomic DNA sequence of organisms has started a new era in biological research. It is now for the first time possible to identify and analyze all genetic elements of a single organism, which are the *Open Reading Frames* (ORFs), RNA and DNA elements. Beside the whole genome also the whole *proteome*, i.e. all proteins expressed by the genome of an organism is now accessible. The analysis of the proteome mainly depends on experimental biological data. The data are published, but in general not stored in electronic form.

Biological processes are mainly determined by *molecular interactions*, e.g. between DNA and proteins, proteins and proteins, or proteins and small molecules. Among these, *protein-protein interactions* play an especially important role since they are essential for almost every biological process. Protein-protein interactions are the fundamental prerequisites for such complex phenomena as control of the cell cycle, DNA replication, transcription, metabolism and signal transduction. The knowledge about the biological context of a single protein, especially of its interactions with other molecules, is mandatory for a precise understanding of its function in the cell.

Studying the functions of individual proteins in various organisms has shown that proteins do not function isolated in a cell but act either in *multi-protein complexes* or in *protein networks*. Often these multi-protein complexes act as highly efficient protein machines (Alberts & Miake-Lye 1992). These protein machines are assemblies of different protein subunits in which the allosteric movement of individual components are coordinated to carry out complicated tasks which need temporal and spatial coordination.

Besides their importance for the formation of multi-protein complexes, protein-protein interactions are involved in a number of other essential features. Proteins are directed to the correct compartments of cells by binding to other proteins; protein messengers bind to protein receptors on the outer surface of cell membranes to exchange signals between cells; proteins form structural connections between cells; some inhibitors of enzymes are proteins; proteins are modified and degraded by enzymes; protein-protein interactions are involved in large-scale movements in organisms, such as muscle contraction. A vast amount of protein-protein interaction data has been generated during the last decades. Recently developed *high-throughput approaches* for a systematic analysis of *genome-wide protein-protein interactions* are widely used, producing large-scale data sets (Fromont-Racine, Rain, & Legrain 1997; Ito *et al.* 2000; Uetz *et al.* 2000). The final goal of studying protein-protein interactions in a given organism is to produce complete *protein interaction maps*.

The MIPS Yeast Interaction Tables

One of the main challenges for the analysis and annotation of the genome of the bakers yeast *Saccharomyces cerevisiae* after completion of the sequencing project (Goffeau *et al.* 1997) was to integrate all available gene-related information of the public domain into a comprehensive yeast database, MYGD at www.mips.biochem.mpg.de/proj/yeast/ (Mewes *et al.* 2000).

Information is gathered from various sources, mainly the systematic functional analysis projects of yeast (Oliver *et al.* 1998) and the yeast literature. Efficient integration of information from the literature requires the application of a standardized terminology as much as possible.

For the annotation of protein-protein interactions we developed the following format. Each interaction consists of 6 different annotation fields: first interactor, second interactor, type of interaction, method the interaction was detected with, references, and free text for additional information.

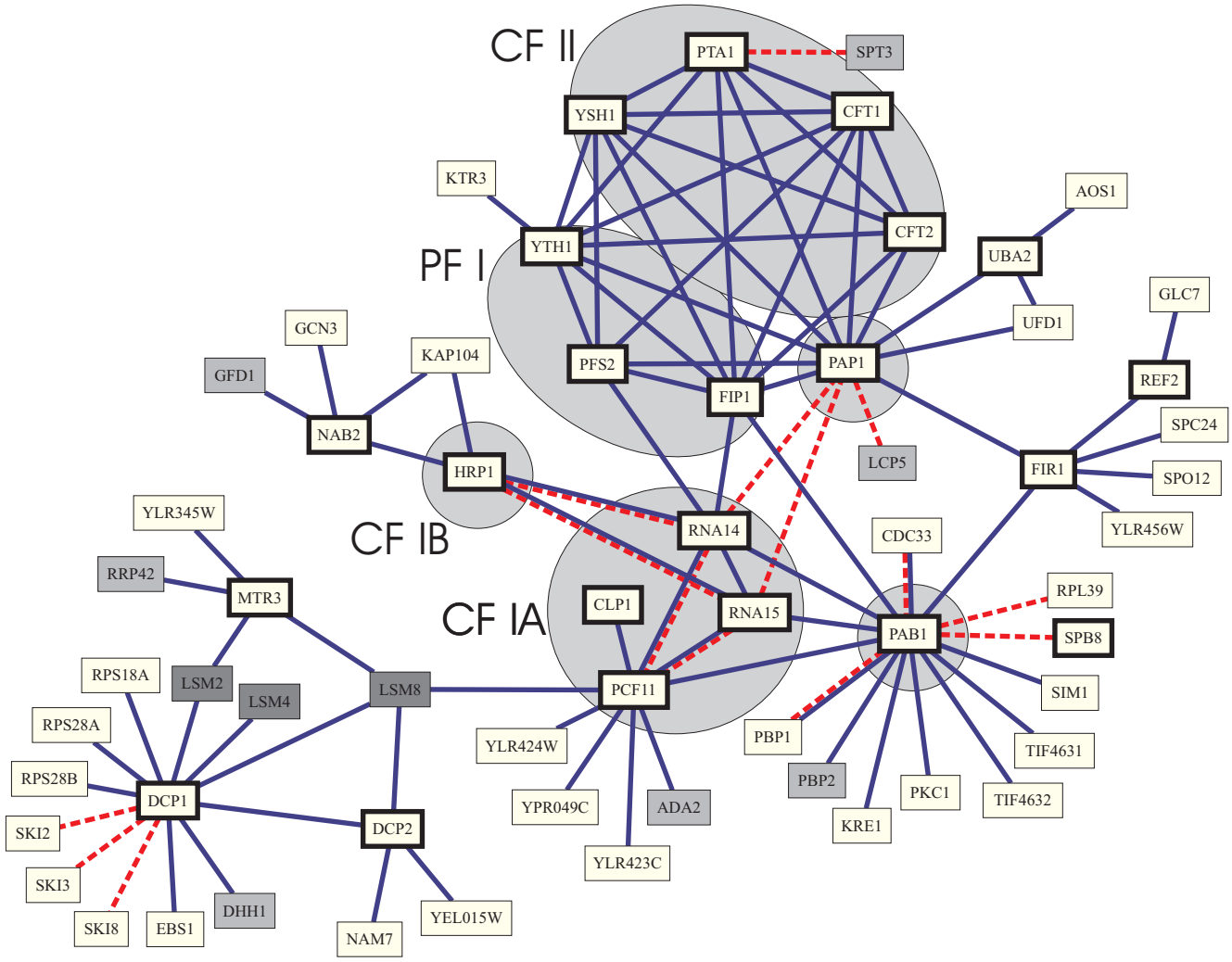

Figure 1: Display of the largest interaction cluster based on the functional category *mRNA processing (5'-, 3'-end processing, mRNA degradation)*. Proteins of the clustered category appear with thick borders, those of other functional categories have thin borders. Solid lines indicate physical interactions, dashed lines indicate genetic interactions. The genes with dark grey boxes belong to functional category *mRNA processing (splicing)*. Those with light grey boxes are genes of category *transcription*, the super-category of mRNA processing. Factors of the processing complex as described in (Zhao, Hyman, & Moore 1999) are enclosed within ellipses (cf. Table 1).

Two types of interactions are distinguished: physical and genetic interactions. The type of interaction is annotated according to the experimental method applied. Physical interactions are detected by e.g. coimmunoprecipitation, two hybrid assay, and affinity purification; genetic interactions are revealed by methods like extragenic suppression, multicopy suppression, synthetic lethality, and transdominant inhibition. Although genetic methods are quite powerful, they are often just a starting point for further biochemical or cell biological experiments since they only give indirect clues for the interaction of two proteins. This standardized annotation format allows the compilation of the gathered data into tables giving easy electronic access to the data (www.mips.biochem.mpg.de/proj/yeast/tables/interaction/).

So far data about interacting domains of individual proteins have not been systematically introduced into the data set. The MIPS Yeast Interaction Tables have been used for other types of presentation such as in the INTERACT database (Eilbeck *et al.* 1999).

The MIPS *yeast interaction tables* consist of 2016 genetic interactions and 3945 physical interactions as of February 2000. For those cases where the interaction type could not be identified from the literature we generated a supplementary table that now contains 197 unclassified interactions.

The Yeast Functional Catalogue

It was recognized that the usage of free text for the systematic functional description of proteins is not adequate for

ORF	Gene	Description	Functional categories	Factor
YDR448w	ADA2	general transcriptional adaptor	04.05.01.04	
YPR180w	AOS1	Smt3p activating enzyme subunit	06.07;06.13.01	
YOL139c	CDC33	translation initiation factor eIF4E	03.10;05.04	
YDR301w	CFT1	pre-mRNA 3'-end processing factor CF II subunit	**04.05.05**	CF II
YLR115w	CFT2	pre-mRNA 3'-end processing factor CF II subunit	**04.05.05**	CF II
YOR250c	CLP1	cleavage/polyadenylation factor IA subunit	**04.05.05**	CF IA
YOL149w	DCP1	mRNA decapping enzyme	**04.05.05**	
YNL118c	DCP2	suppressor protein of a yeast pet mutant	02.13;**04.05.05**	
YDL160c	DHH1	strong similarity to RNA helicases of the DEAD box family	04.99	
YDR206w	EBS1	similarity to EST1 protein	03.16	
YJR093c	FIP1	component of pre-mRNA polyadenylation factor PF I	**04.05.05**	PF I
YER032w	FIR1	interacts with the poly(A) polymerase	**04.05.05**	
YKR026c	GCN3	translation initiation factor eIF2B, 34 KD, alpha subunit	05.04	
YMR255w	GFD1	nuclear pore complex protein	04.07;08.01	
YER133w	GLC7	ser/thr phosphoprotein phosphatase 1, catalytic chain	01.05.04;02.19;03.04; 03.13;03.22;05.07	
YOL123w	HRP1	cleavage/polyadenylation factor IB	**04.05.05**;04.05.99;04.07	CF IB
YBR017c	KAP104	beta-karyopherin	08.01	
YNL322c	KRE1	cell wall protein	01.05.01	
YBR205w	KTR3	alpha-1,2-mannosyltransferase	01.05.01;06.07	
YER127w	LCP5	Ngg1p interacting protein	04.01.04	
YBL026w	LSM2	snRNP-related protein	04.05.99	
YER112w	LSM4	U6 snRNA associated protein	04.05.03	
YJR022w	LSM8	splicing factor	04.05.03	
YGR158c	MTR3	involved in mRNA transport	04.01.04;**04.05.05**;04.07	
YGL122c	NAB2	nuclear poly(A)-binding protein	**04.05.05**;04.07	
YMR080c	NAM7	nonsense-mediated mRNA decay protein	01.03.16;05.07	
YER165w	PAB1	mRNA polyadenylate-binding protein	**04.05.05**;05.04	Pab 1
YKR002w	PAP1	poly(A) polymerase	**04.05.05**	Pap 1
YGR178c	PBP1	Pab1p interacting protein	**04.05.05**	
YBR233w	PBP2	Pab1p interacting protein	04.99	
YDR228c	PCF11	component of factor CF I	**04.05.05**	CF IA
YNL317w	PFS2	component of pre-mRNA polyadenylation factor PF I	**04.05.05**	PF I
YBL105c	PKC1	ser/thr protein kinase	03.04;03.22; 10.01.05.11;11.01	
YAL043c	PTA1	pre-mRNA 3'-end processing factor CF II subunit	04.03.03;**04.05.05**	CF II
YDR195w	REF2	RNA 3'-end formation protein	**04.05.05**	
YMR061w	RNA14	component of factor CF I	04.05.01.04;**04.05.05**	CF IA
YGL044c	RNA15	component of factor CF I	**04.05.05**	CF IA
YJL189w	RPL39	ribosomal protein L39.e	05.01	
YDR450w	RPS18A	ribosomal protein S18.e.c4	05.01	
YOR167c	RPS28A	ribosomal protein S28.e.c15	05.01	
YLR264w	RPS28B	ribosomal protein S28.e.c12	05.01	
YDL111c	RRP42	rRNA processing protein	04.01.04	
YIL123w	SIM1	involved in cell cycle regulation and aging	03.22;11.11	
YLR398c	SKI2	antiviral protein and putative helicase	11.07	
YPR189w	SKI3	antiviral protein	11.99	
YGL213c	SKI8	antiviral protein	03.13;03.19;11.13	
YJL124c	SPB8	suppressor of PAB1	**04.05.05**	
YMR117c	SPC24	spindle pole body protein	03.22	
YHR152w	SPO12	sporulation protein	03.10;03.13	
YDR392w	SPT3	general transcriptional adaptor	03.07;04.05.01.04	
YGR162w	TIF4631	mRNA cap-binding protein (eIF4F), 150K subunit	05.04	
YGL049c	TIF4632	mRNA cap-binding protein (eIF4F), 130K subunit	05.04	
YDR390c	UBA2	E1-like (ubiquitin-activating) enzyme	**04.05.05**;06.07;06.13.01	
YGR048w	UFD1	ubiquitin fusion degradation protein	06.13.01	
YLR345w		similarity to 6-phosphofructo-2-kinases	01.05.04;02.01	
YLR277c	YSH1	pre-mRNA 3'-end processing factor CF II subunit	**04.05.05**	CF II
YPR107c	YTH1	component of pre-mRNA polyadenylation factor PF I	**04.05.05**	PF I

Table 1: The functionally classified proteins of the main cluster of category *mRNA processing (5'-, 3'-end processing, mRNA degradation)* (04.05.05), displayed in Figure 1. The description of the factors can be found in (Zhao, Hyman, & Moore 1999). Functional categories are explained in Table 5.

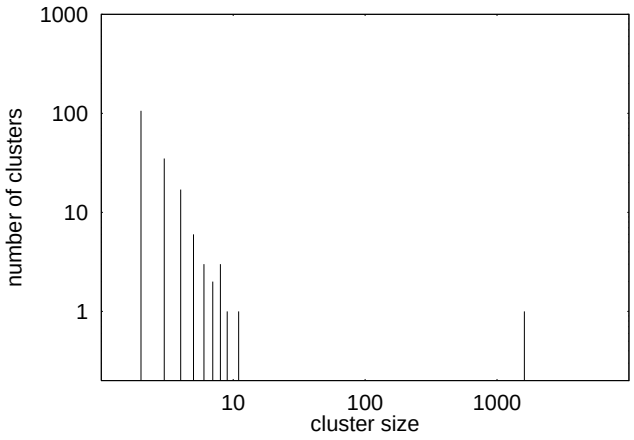

Figure 2: Histogram of cluster sizes. Cluster sizes are shown on the abscissa, the number of clusters of the respective cluster sizes are shown on the ordinate. Both scales are logarithmic. For the clustering, the ORFs of all functional categories have been considered. Though there are 174 small clusters with less than 12 ORFs, these contain a total of only 562 ORFs, while one cluster comprises 1721 ORFs.

computational tasks (Riley 1993). In analogy to the established EC catalogue (NC-IUBMB 1992) a hierarchical ordering of the gene products of a cell in terms of their function is an adequate solution for a systematic approach. Different levels of categories group together biochemical functionality according to their role in the organism in rough analogy to biochemical textbooks grouping the biochemical information in paragraphs, chapters and sections.

As the sequence of the yeast *Saccharomyces cerevisiae* was available in 1996 we have generated a special *functional catalogue* for yeast (Mewes *et al.* 1997) (www.mips.biochem.mpg.de/proj/yeast/catalogues/funcat/). Significant homologies of proteins to functionally characterized proteins as well as data from the literature derived from biochemical, genetic or phenotypic experiments are used to assign functions. Proteins can be assigned to more than one functional category. This allows a multidimensional annotation. The yeast functional catalogue is hierarchically organized. It contains 15 main categories, each containing 3 to 4 levels of subcategories. In total the catalogue consists of more than 200 functional categories. For 3793 out of 6359 yeast genes at least one of the functional categories is assigned, the remaining proteins are assigned to the category *unclassified proteins*.

The combination of different data sets leads to a more detailed level of information. Here we show how data of physical and genetic interactions can be combined with other types of biological data. We show in detail the combination with the functional categories.

Methods

The genetic interactions as well as the physical interactions are binary relations. They are ideally suited for visualization as graphs. Genes and proteins are modeled as nodes, the interactions are represented as edges between the respective nodes. A graph editor tool-kit can be employed for displaying the interaction graphs. We customized the LEDA graph editor (Mehlhorn & Näher 1999) for the graphical visualization of interaction graphs. The nodes are labeled with the systematic names or the gene names if available. Edges correspond to interactions and are drawn according to the interaction mode. Physical interactions are represented by solid edges, genetic interactions by dashed edges. A color code can be applied for a deeper characterization of the different methods by which the interactions have been detected. The algorithms of the editor create a suitable layout for the complex graphs, resulting in a clear, easy to grasp picture of the displayed interactions. The user can alter the graph by moving nodes and by deleting nodes and edges.

functional category	genes in this category	genes with interaction	genes in biggest cluster	interaction in all clusters
all	6359	2283	1721	6158
04	749	460	732	2416
04.05	542	336	558	1862
04.05.05	40	32	62	259

Table 2: Number of genes found in different categories and properties of the resulting clusters. Categories: *Transcription* (04), *mRNA processing* (04.05) and *5'-, 3'-end processing, mRNA degradation* (04.05.05).

In an iterative procedure we build *clusters of genes* due to the interactions annotated. Every single gene initially represents its own cluster. For every annotated interaction, where the interacting proteins are not in the same cluster, we join the two clusters involved. After the clustering, every cluster contains all the genes that interact either directly or indirectly. In a graph theoretic sense, modeling genes as nodes and interactions as edges, we build clusters of genes that belong to the same connected component of the whole interaction graph. Homomultimeric interactions are not considered.

The MIPS functional catalogue is well suited for selecting subsets of genes with related functions. We restrict the gene set to those genes sharing functional categories. The result is an interaction graph that consists of the genes belonging to the selected functional categories (nodes drawn with thick borders) and those directly interacting with them (nodes drawn with thin borders).

Results

Most of the categories of the MIPS functional catalogue contain a large number of genes due to the fact that biological processes in general involve a substantial number of proteins. The catalogue does not describe the exact role of single genes in the cell. The combination with other data adds the information necessary for an exact assignment of the genes to cellular processes.

Protein-protein interaction data in turn is not sufficient for

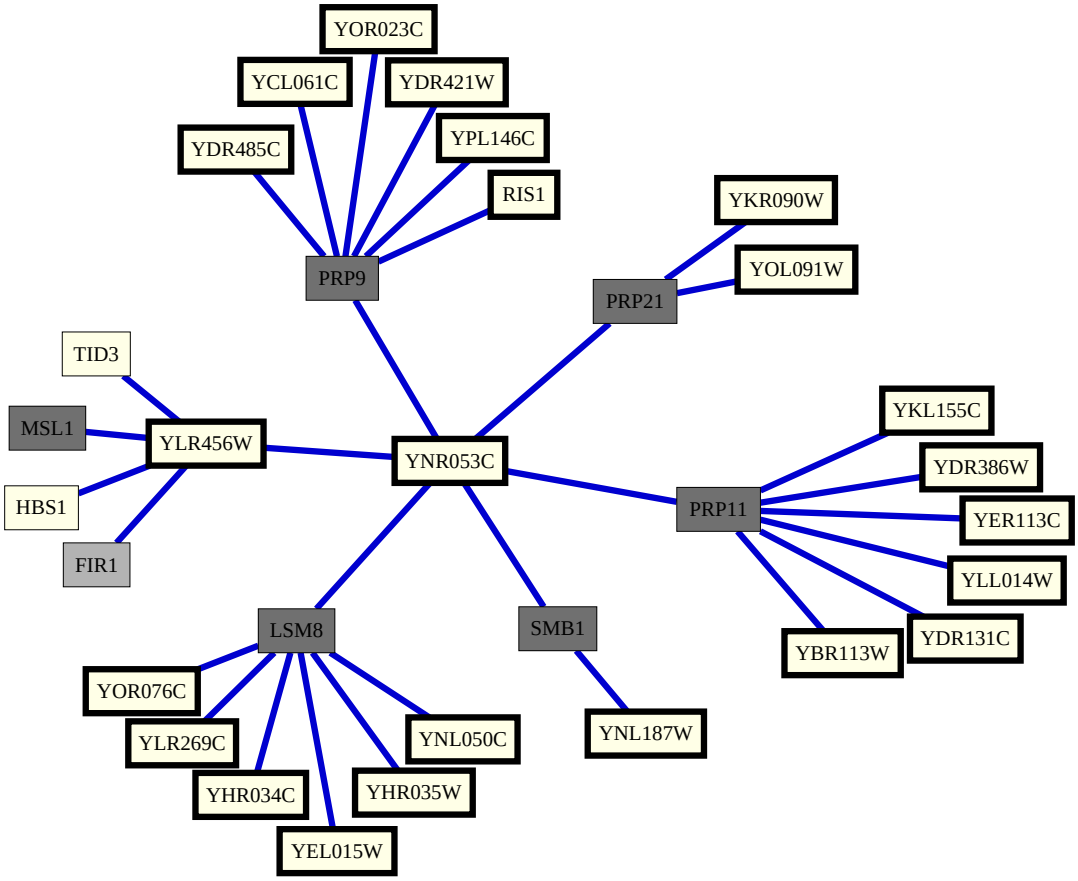

Figure 3: Part of an interaction cluster resulting from the gene set restricted to the non-characterized proteins. YNR053c directly interacts with five genes belonging to the functional category *mRNA processing (splicing)*, shown in dark grey. Another gene of this category is found in the second interaction level: *MSL1*. *FIR1* belongs to the same biological context, *mRNA processing (5'-, 3'-end processing, mRNA degradation)*. It is indicated in light grey. The graph is cut at the second interaction level with respect to YNR053c. Non-characterized proteins appear with thick borders, those with annotated functional categories have thin borders. The functional categories of the individual proteins are listed in Table 3. The solid lines indicate physical interactions.

ORF	Gene	Description	Functional categories	Interaction level
YDL030w	PRP9	pre-mRNA splicing factor	**04.05.03**	1
YDL043c	PRP11	pre-mRNA splicing factor	**04.05.03**	1
YER029c	SMB1	associated with U1 snRNP	**04.05.03**	1
YER032w	FIR1	interacts with the poly(A) polymerase	04.05.05	2
YIL144w	TID3	Dmc1p interacting protein	03.22	2
YIR009w	MSL1	strong similarity to snRNPs	**04.05.03**;06.10	2
YJL203w	PRP21	pre-mRNA splicing factor	**04.05.03**;06.10	1
YJR022w	LSM8	splicing factor	**04.05.03**	1
YKR084c	HBS1	eEF-1 alpha chain homologue	05.04	2
YLR456W		strong similarity to YPR172w	99	1

Table 3: The interacting partners of the non-characterized protein YNR053c. All interacting proteins of level 1 and those of level 2 with a described function are listed. The functional categories are described in Table 5.

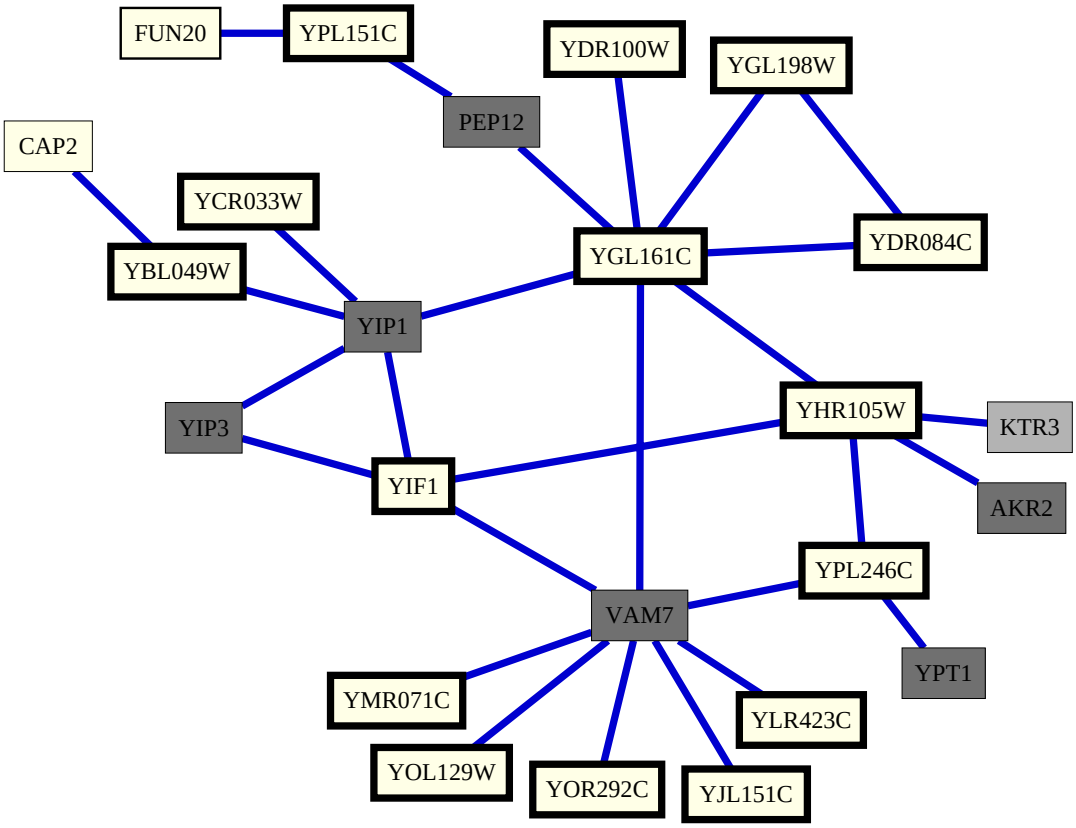

Figure 4: Part of an interaction cluster resulting from the gene set restricted to the non-characterized proteins. The graph is cut at the third interaction level with respect to YGL161c. Non-characterized proteins appear with thick borders, those with annotated functional categories have thin borders. The functional categories of the individual proteins are listed in Table 4. The solid lines indicate physical interactions.

ORF	Gene	Description	Functional categories	Interaction level
YAL032c	FUN20	required for RNA splicing	04.05.03	3
YBR205w	KTR3	alpha-1,2-mannosyltransferase	01.05.01;06.07	2
YFL038c	YPT1	GTP-binding protein of the rab family	**08.07**	3
YGL212w	VAM7	vacuolar morphogenesis protein	**08.13**;09.25	1
YGR172c	YIP1	golgi membrane protein	**08.07**	1
YIL034c	CAP2	F-actin capping protein, beta subunit	03.04;03.07	3
YNL044w	YIP3	involved in ER to golgi transport	06.04;**08.07**	2
YOR034c	AKR2	involved in constitutive endocytosis of Ste3p	**08.19**	2
YOR036w	PEP12	syntaxin (T-SNARE), vacuolar	06.04;**08.13**	1

Table 4: The interacting partners of the non-characterized protein YGL161c. All interacting proteins of level 1 and those of levels 2 and 3 with a described function are listed. The functional categories are described in Table 5.

a comprehensive description of the functional context. If no additional information, e.g. about the time, localization or function of the individual interactions is taken into account, i.e. using the pure clustering method described above, the proteome of an organism tends to be in a single cluster.

Verification of the Method

The MIPS yeast interaction tables contain 6158 individual interactions annotated for 2283 ORFs as of February 2000. Clustering these genes as described above leads to 175 clusters. While 106 clusters consist of just two genes and 35 clusters comprise three genes, there is one cluster containing 1721 genes (main cluster). The remaining 33 clusters consist of 4 to 11 genes (Figure 2). A reasonable visualization and interpretation can be computed for the small clusters while the interaction graph of the large one is too complex to produce a clear, easy to grasp representation. Thus the interaction graph of the vast majority of the genes cannot be shown. One possibility to reduce the complexity of the data is to combine the set of genes to be analyzed to functionally categorized genes.

For the functional category *transcription* and two of its subcategories, we show the properties of the gene set restricted to the respective category and the resulting cluster sizes (Table 2).

Restricting the interaction to the lowest hierarchical level, i.e. category *mRNA processing (5'-, 3'-end processing, mRNA degradation)* leads to a gene set that consists of 40 genes, 32 of these with annotated interactions. We use this well described category to verify the utility of the method. For the clustering all interactions annotated for the genes of this category are used, including those interactions with a gene not belonging to the category. 8 clusters result from this process, the largest cluster containing 62 proteins, 23 from the specified functional category, and 99 interactions between the proteins.

As expected, the interaction graph of this cluster (Figure 1) corresponds to the description of the *pre-mRNA 3'-end processing* in the literature (Zhao, Hyman, & Moore 1999). According to this review, the *yeast cleavage/polyadenylation complex* consists of 15 identified proteins. For 14 of them the corresponding genes are known. The complex is subdivided into five functionally distinct activities. CF IA, IB and II are described as sufficient for the cleavage reaction, and specific poly(A)-addition are described as to require CF IA and IB, Pap1, Pab1, and PF I. These functional factors can be identified in the interaction graph. The graph contains additional interactions to proteins of other functional categories, linking the cleavage/polyadenylation complex with neighbouring cellular processes like splicing and transcription (Table 1).

The fifteenth, missing protein has been identified by purification of the PF I complex. It is a protein of 58kd called Pfs1p, which has not been published yet. PFS1 is an essential gene containing a zinc knuckle (Zhao, Hyman, & Moore 1999). We did not find interaction data pointing to an ORF fulfilling these requirements. Thus it is not possible to speculate about this protein on the basis of the protein interaction data available.

Revealing the Biological Context of Non-Characterized Proteins

So far, roughly a third of the yeast ORFs are not functionally described (Mewes *et al.* 1997). Several systematic approaches have been developed in order to functionally classify these ORFs (Oliver *et al.* 1998). Since 1997 about 7% of the ORFs then called *proteins of unknown function* have been functionally described (cf. MYGD – www.mips.biochem.mpg.de/proj/yeast/ and YPD – www.proteome.com/databases/index.html).

We show here how the usage of protein interaction data can provide substantial clues as to the biological context of unknown proteins.

Restricting the gene set to the 2563 non-characterized proteins gives rise to 177 clusters. 39 of these clusters contain more than two unclassified genes. The biggest cluster, comprising 93 non-characterized and 66 characterized proteins, contains the ORFs YNR053c and YGL161c. We show the capabilities of our integrative method for the prediction of functions for so far non-characterized proteins by means of these ORFs. Evaluating a protein interaction graph, functional relationships between proteins can be deduced depending on the distance of the respective proteins. Two proteins that directly interact are most likely to be involved in the same biological process or pathway (Walhout *et al.* 2000). We thus concentrate on the direct surrounding of the considered ORFs. Therefore the interaction graphs of YNR053c (Figure 3) and YGL161c (Figure 4) have been cut at the second and third interaction level, respectively.

The functional context of YNR053c. YNR053c is known to interact directly with six other proteins (Fromont-Racine, Rain, & Legrain 1997), five of which have a known functional classification (Figure 3). These five all belong to the category *mRNA processing (splicing)* (cf. Table 3). All but four of the indirectly connected proteins of the second level are uncharacterized. The four classified proteins comprise Msl1p that is also involved in the mRNA splicing and Fir1p that is involved in 3'-end mRNA processing. The remaining proteins, Tid3p and Hbs1p belong to other functional categories. The central position of YNR053c in the described protein interaction network is a strong clue as to its functional role in mRNA splicing. For YLR456w, a non-characterized protein that directly interacts with YNR053c, a functional prediction is more difficult to make. The five interactors of this ORF belong to more diverse categories. Two of the interacting proteins are known to be involved in mRNA transcription (Msl1p, Fir1p). For the third interactor, YNR053c, there are strong indications for the participation in mRNA splicing as described above. The two remaining ORFs are involved in cell cycle control and translation, respectively. Considering the whole interaction context of YLR456w, allows to hypothesize on a potential cellular function in mRNA splicing.

The functional context of YGL161c. For YGL161c seven direct interactions with other proteins are known (Ito *et al.* 2000; Uetz *et al.* 2000), three of them have a known function (Figure 4). These three proteins, Vam7p, Yip1p,

Systematic number	Description of category
01	METABOLISM
01.03.16	polynucleotide degradation
01.05.01	C-compound and carbohydrate utilization
01.05.04	regulation of C-compound and carbohydrate utilization
02	ENERGY
02.01	glycolysis and gluconeogenesis
02.13	respiration
02.19	metabolism of energy reserves (glycogen, trehalose)
03	CELL GROWTH, CELL DIVISION AND DNA SYNTHESIS
03.04	budding, cell polarity and filament formation
03.07	pheromone response, mating-type determination, sex-specific proteins
03.10	sporulation and germination
03.13	meiosis
03.16	DNA synthesis and replication
03.19	recombination and DNA repair
03.22	cell cycle control and mitosis
04	TRANSCRIPTION
04.01.04	rRNA processing
04.03.03	tRNA processing
04.05.01.04	transcriptional control
04.05.03	mRNA processing (splicing)
04.05.05	mRNA processing (5'-, 3'-end processing, mRNA degradation)
04.05.99	other mRNA-transcription activities
04.07	RNA transport
04.99	other transcription activities
05	PROTEIN SYNTHESIS
05.01	ribosomal proteins
05.04	translation (initiation, elongation and termination)
05.07	translational control
06	PROTEIN DESTINATION
06.04	protein targeting, sorting and translocation
06.07	protein modification (glycolsylation, acylation, myristylation, palmitylation, farnesylation and processing)
06.10	assembly of protein complexes
06.13.01	cytoplasmic degradation
08	INTRACELLULAR TRANSPORT
08.01	nuclear transport
08.07	vesicular transport (Golgi network, etc.)
08.13	vacuolar transport
08.19	cellular import
09	CELLULAR BIOGENESIS (proteins are not localized to the corresponding organelle)
09.25	vacuolar and lysosomal biogenesis
10	CELLULAR COMMUNICATION/SIGNAL TRANSDUCTION
10.01.05.11	key kinases
11	CELL RESCUE, DEFENSE, CELL DEATH AND AGEING
11.01	stress response
11.07	detoxificaton
11.11	ageing
11.13	degradation of exogenous polynucleotides
11.99	other cell rescue activities
99	UNCLASSIFIED PROTEINS

Table 5: A part of the MIPS functional catalogue. The systematic numbers appearing in Tables 1, 3, and 4 are described. The complete functional catalogue is available at MIPS (www.mips.biochem.mpg.de/proj/yeast/catalogues/funcat/).

and Pep12p, are involved in intracellular transport. Additionally, two of three functionally classified proteins of the second interaction level, Yip3p and Akr2p, are also intracellular transport proteins. The third, Ktr3p, is an alpha-1,2-mannosyltransferase involved in glycosylation of proteins to be secreted. Ktr3p is likely to be localized in the golgi apparatus (Sipos, Puoti, & Conzelmann 1995). Thus the cellular role of Ktr3p is closely linked to intracellular transport processes. Even on the third interaction level of YGL161c one more protein, Ypt1p is described to be involved in intracellular transport processes. Considering all these data one can assume that YGL161c is involved in intracellular transport. In addition, there is some evidence for the non-characterized proteins YIF1, YHR105w, and YPL246c to be involved in the context of intracellular transport. All three have a central position in the described interaction network of intracellular transport processes.

Discussion

We have developed a method for the integrative analysis of protein interactions. We combined protein interaction data on *S. cerevisiae* proteins, systematically collected from the literature, with the functional classification data of all yeast proteins. A clustering is being performed to find maximal groups of proteins that are directly or indirectly connected to interaction networks. These networks are graphically represented using a graph editor toolkit. The visualization of protein interaction networks supports the comprehensive analysis of these large data sets.

The characteristic of the resulting cluster sizes (Figure 2) suggests that the more interactions are known for an organism the larger the clusters of genes become. We suspect that if all interactions are known, a single cluster of genes results from the described clustering method. It is thus necessary to focus on a subset of the whole set of genes. A more efficient and exploratory interpretation of the protein interaction data is enabled by the concentration on a certain biological context that is achieved by the use of the systematic functional categorisation of all yeast ORFs.

This combination of different data sets results in the necessary reduction of complexity. Our results show that the integrative analysis with a hierarchically organized data set allows to scale the complexity of the interaction graphs (Table 2).

The utility of the presented method has been shown by applying it to the well described functional context of mRNA processing. The analysis of the biological context of the uncharacterized proteins YNR053c and YGL161c shows the relevance of the method for mining the protein interaction data and formulating hypothesis about a functional classification of so far uncharacterized proteins.

It has been shown that interactions among proteins are conserved between the homologous proteins of various organisms (Uetz *et al.* 2000). In cross-genome analysis the presented method can be used for the prediction of protein interactions in other species. This is of particular interest with respect to sequences resulting from whole genome sequencing projects, e.g. the human genome project.

The integrative analysis method is easy to use and allows a comprehensive overview of the protein interaction data. It is very flexible, in that it can easily be applied to other types of large-scale data sets to be combined, e.g. of protein complexes, EC numbers, subcellular localization, and phenotypes (www.mips.biochem.mpg.de/proj/yeast/catalogues/). It is also possible and very promising to combine the protein interaction data with other data produced by high-throughput methods, the most prominent being expression data produced by DNA microarray technology (DeRisi, Iyer, & Brown 1997) and localization via GFP tagging (Brachat *et al.* 2000).

References

Alberts, B., and Miake-Lye, R. 1992. Unscrambling the puzzle of biological machines: the importance of the details. *Cell* 68(3):415–420.

Brachat, A.; Liebundguth, N.; Rebischung, C.; et al. 2000. Analysis of deletion phenotypes and GFP fusions of 21 novel *Saccharomyces cerevisiae* open reading frames. *Yeast* 16(3):241–253.

DeRisi, J. L.; Iyer, V. R.; and Brown, P. O. 1997. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278(5338):680–686.

Eilbeck, K.; Brass, A.; Paton, N.; and Hodgman, C. 1999. INTERACT: an object oriented protein-protein interaction database. In *Proc of the Seventh Int Conf on Intelligent Systems for Mol Biol*, 87–94. AAAI Press.

Fromont-Racine, M.; Rain, J.; and Legrain, P. 1997. Toward a functional analysis of the yeast genome through exhaustive two-hybrid screens. *Nat Genet* 16(3):277–282.

Goffeau, A.; Aert, R.; Agostini-Carbone, M.; et al. 1997. The yeast genome directory. *Nature* 387(6632 Suppl):1–105.

Ito, T.; Tashiro, K.; Muta, S.; et al. 2000. 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* 97(3):1143–1147.

Mehlhorn, K., and Näher, S. 1999. *The LEDA Platform of Combinatorial and Geometric Computing*. Cambridge University Press. http://www.mpi-sb.mpg.de/LEDA/.

Mewes, H.; Albermann, K.; Bähr, M.; et al. 1997. Overview of the yeast genome. *Nature* 387(6632 Suppl):7–65.

Mewes, H.; Frishman, D.; Gruber, C.; et al. 2000. MIPS: a database for genomes and protein sequences. *Nucleic Acids Res* 28(1):37–40.

NC-IUBMB. 1992. *Enzyme Nomenclature – Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (IUBMB)*. New York: Academic Press.

Oliver, S.; Winson, M.; Kell, D.; and Baganz, F. 1998. Systematic functional analysis of the yeast genome. *Trends Biotechnol* 16(9):373–378.

Riley, M. 1993. Functions of the gene products of escherichia coli. *Microbiol Rev* 57(4):862–952.

Sipos, G.; Puoti, A.; and Conzelmann, A. 1995. Biosynthesis of the side chain of yeast glycosylphosphatidylinositol anchors is operated by novel mannosyltransferases located in the endoplasmic reticulum and the golgi apparatus. *J Biol Chem* 270(34):19709–19715.

Uetz, P.; Giot, L.; Cagney, G.; et al. 2000. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* 403:623–627.

Walhout, A.; Sordella, R.; Lu, X.; et al. 2000. Protein interaction mapping in c. elegans using proteins involved in vulval development. *Science* 287:116–122.

Zhao, J.; Hyman, L.; and Moore, C. 1999. Formation of mRNA 3' ends in eukaryotes: mechanism, regulation, and interrelationships with other steps in mRNA synthesis. *Mol Biol Rev* 63(2):405–445.