Analysis of Gene Expression Data **Spring Semester, 2005**

Lecture 10: May 19, 2005

Lecturer: Roded Sharan *Scribe: Daniela Raijman and Igor Ulitsky*

10.1 Protein Interaction Networks

In the past we have discussed different types of biological networks. For example, in the context of gene expression we discussed networks in which each node represents a gene and an edge represents a similarity of gene expression profiles. In this lecture, protein interaction networks will be discussed. Protein interaction networks are graphs in which each node represents a protein, and an edge between two nodes represents an evidence for the presence of a physical interaction between the two proteins. A small subset of such a network can be seen in Figure 10.1. For example, such a physical interaction can be a kinase protein which *phosphorylates* another protein. The network does not contain information about the nature of the interaction (activation, de-activation, two proteins which participate in the same complex, etc.). The use of protein interaction networks has recently expanded due to the development of high-throughput technologies measuring these interactions, and the availability of large data sources containing interaction evidence for different species.

10.2 High-Throughput technologies for measuring protein-protein interactions

10.2.1 Yeast Two-Hybrid

The Yeast Two-Hybrid method [3] for detection of protein-protein interactions utilizes the fact that *Transcription Factors* commonly require two domains - a *DNA binding* domain and an *activation* domain promoting transcription. In order to find out which proteins interact with a certain protein of interest (termed *bait*), the bait is fused to a DNA binding domain which binds to the promoter of a reporter gene, while the other proteins (termed *prey*) are fused with an activation domain. When a physical interaction between the bait and some prey occurs, an expression of the reporter gene can be detected. Even transient interactions can be detected using this method. The method is summarized in Figure 10.2.

10.2.2 Protein coImmunoPrecipitation (coIP)

In this method, the bait protein is marked by a tag. At a certain point in time an antibody which recognizes the tag is used to trap the bait protein and precipitate it. In the precipitation process any protein which is in a physical contact or in the same complex with the bait is precipitated as well. Following this, mass spectrometry is used to determine the identity of the prey proteins. The advantage of this method is in its ability to discover interactions

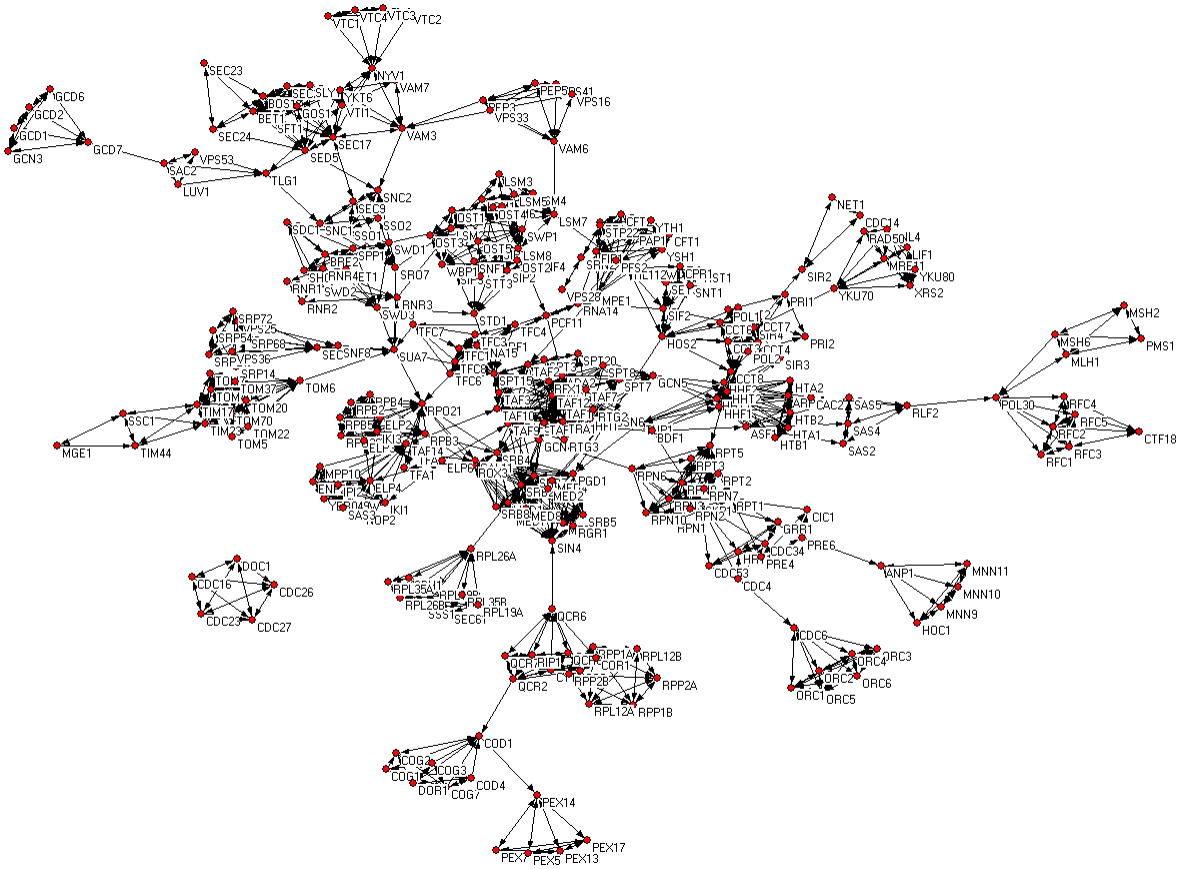

Figure 10.1: A small part of the budding yeast (*Saccharomyces cerevisiae*) protein interaction network [].

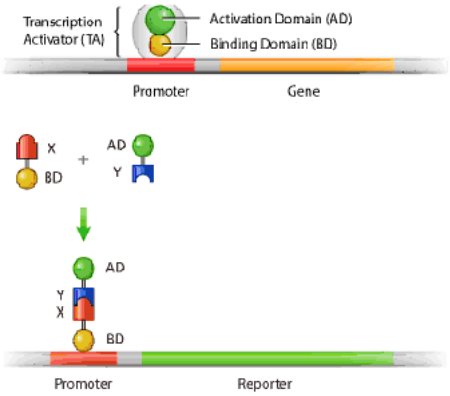

Figure 10.2: The Yeast Two-Hybrid technology for detecting protein-protein interactions. An interaction between a protein containing a DNA binding domain (bait) and an activation domain (prey) can be detected by measuring the expression level of the reporter gene.

between a single bait and multiple prey proteins which require no manipulation. The drawback here is that it is unknown which protein has been in a direct physical contact with which protein. Direct interactions can not be distinguished from interactions mediated by other proteins in a complex. A summary of this method can be seen in figure 10.3.

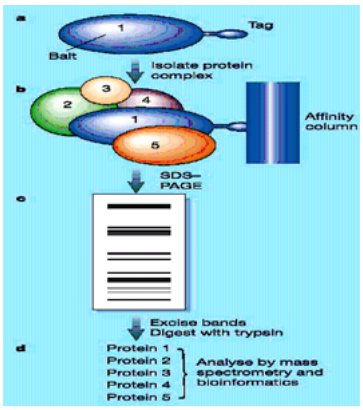

Figure 10.3: The coImmunoPrecipitation technology for detecting protein-protein interactions. A binding site for an antibody (TAP epitope) is inserted into the bait protein and it is used for precipitating proteins found in complex with the bait in the cell.

10.3 Scale-Free Networks

There are multiple ways by which networks can be characterized. One way is to analyze the network by the distribution of the node degrees. Recently it was noticed that most naturally occuring networks are scale-free [9]. Scale-free networks are characterized by a small number of high-degree nodes, which are termed *hubs*, and a large number of small-degree nodes. The distribution is of the form $P(k) = k^{-c}$ where c is a constant. This in contrast to a random graph, where the degree distribution is binomial. The scale-free nature of the protein interaction networks ($c \approx 2.5$) can be seen in Figure 10.4.

10.4 Data Processing

10.4.1 Protein Interaction Data Quality

Using protein interaction data, we would like to be able to infer biological complexes and pathways, uncovering the cellular machinery. However, the interaction data is noisy and incomplete, an issue that must be addressed for all purposes.

Figure 10.5 shows the overlap between interactions discovered in different experiments in yeast. The overlap is very small, even considering the fact that there is an estimated number of 20,000 interactions in yeast. It is currently believed that only about 50% of interactions are known, and that of the known interactions about 50% are incorrect.

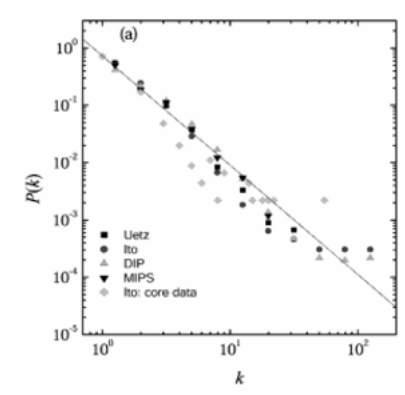

Figure 10.4: Distribution of node degrees in networks constructed from different large-scale experiments in yeast (log-scale). In all the experiments, regardless of the technique, linearity is maintained

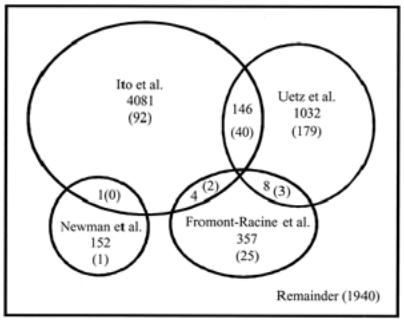

Figure 10.5: The overlap between the various studies containing protein interaction data *Ito*:[10],*Uetz*:[6],*Fromont*:[5],*Newman*:[4],.

10.4.2 Noise and Reliability Estimation

Study Reliability Estimation

One method of estimating the credibility of the data is to use an independent dataset, and estimate the percentage of true interactions, denoted as r (referring to the *reliability* of the data) . The assumption, based on several experiments, is that there is a correlation between expression similarity and physical interaction, meaning that a pair of proteins which interact are likely to have similar expression patterns for their corresponding genes. Under this assumption, for each interaction the correlation between the expression of the genes which code for the interacting proteins is measured. The correlations are divided into k bins. In each bin we expect to see true interactions (with probability r) and non-true interactions (with probability $1-r$). Let us denote by p_k the probability of a true interaction being in the k-th bin, and by q_k the probability of a non-interaction being in the k-th bin (of size n_k). $\sum_k p_k = 1, \sum_k q_k = 1$. ps and qs are assumed to be independent. The value of r is picked

such that the the likelihood of the data is maximized:

$$L(r) = \prod_k (rp_k + (1-r)q_k)^{n_k}$$

The reliability estimation results for several major studies can be seen in Table 10.1.

Table 10.1: The reliability (r) of protein interaction studies.

Data	*Interactions*	r
Uetz	1436	0.53
Ito2	1469	0.56
Ito8	276	0.88
TAP	17962	0.59

Edge Reliability Estimation

Another approach is to estimate for each *interaction* the probability that it is a true interaction. This is done using *SVM* as a classifier which can classify interactions as being true or false. For each interaction a vector of information is constructed : the number of times the interaction was observed, the gene expression correlation, the number of common neighbors of the interacting proteins have in the interaction network, etc. The information vectors are used as the input data for the classification, and the classifier is trained using a training set - interactions for which the true or false nature is known.

Logistic Regression

Instead of an SVM classifier, logistic regression can also be used for the classification. The logistic function $P(x) = logit(x) = \frac{1}{1+e^{-x}}$ has a sigmoidal form (Figure 10.6). The motivation behind the use of the logistic function for classification is that the more evidence we have, the less each individual evidence contributes to our decision. The logistic function is a discriminative model - it models the probability of the label (true or false). The posterior belief is modelled using the *logit* function. Parameters are learned using a training set, and the given a new observation x it can be classified as being true or false. The learning of the parameters is done by maximizing the likelihood of the data. The advantage is that the likelihood function in this case is *concave*, and therefore the best parameters can be found using a greedy algorithm.

10.5 Algorithms on Single Networks

10.5.1 Pathway Extraction

Once we obtain a processed network with a probability assigned for each edge, we want to find pathways (signalling pathway, metabolic pathways, etc.) in the network. A pathway is a heavy path in the network (a path with high-probability edges). In addition, there can be

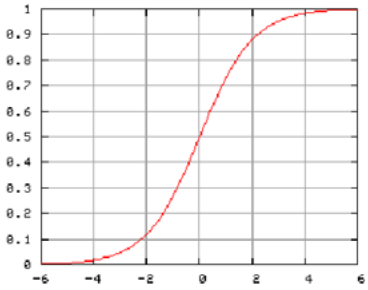

Figure 10.6: The logistic function.

different constraints on the path we are looking for. For example. For instance, we would like to be able to find a path which begins with a trans-membranal protein and ends with a transcription factor.

Color Coding

The problem of finding a simple path of a maximal weight is NP-Hard (by reduction from *Hamiltonian Path*). However, a probabilistic algorithm can be used to solve the problem correctly with high probability.

The idea of *Color Coding* [1] is as follows: We paint each node using a random color out of k possible colors (where k is the length of the path we are looking for). If we find a path in the colored graph which contains k different colors, then it is necessarily a simple path (no node is visited twice). *Dynamic Programming* can be used to find such a path.

Since the coloring is performed at random, it is possible that there exists a simple and heavy path of length k, but we colored two or more nodes in that path using the same color. The probability of that is $(\frac{k!}{k^k})$ which is approximately e^{-k}. Therefore, we would have to try e^k different random colorings in order to obtain the path of a maximal weight with high probability.

The color coding algorithm can be easily expanded to include different constraints with biological motivation:

• One can force the algorithm to find a path starting with a trans-membranal protein and ending with a transcription factor, by changing the start and end conditions.

• One can force the path to contain a specific protein by coloring it with a unique color.

• One can force the path to contain exactly one protein which belongs to a specific group (for example transport proteins) by coloring the group in a unique color.

• Other contraints can be added to the Dynamic Programming logic.

Results and Validation

There are two ways to evaluate path extraction results. One approach is to show that the weights of the paths are significantly high. For each path found, a path weight-based p-value is calculated by comparing with best paths in random graphs with the same degree distribution. Another is to show functional enrichment found in the proteins of the obtained

path. The results of such analysis can be seen in 10.7. Biological examples of comparisons between the known paths and paths extracted using the Color Coding algorithm can be seen in Figures 10.8 and 10.9.

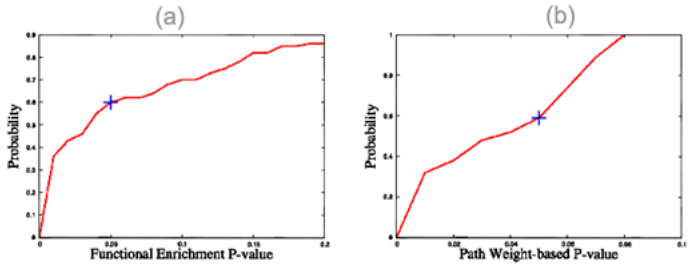

Figure 10.7: The assessment of the quality of the discovered pathways. Over 60% of the discovered pathways have a p-value smaller than 0.05 under both criterions.

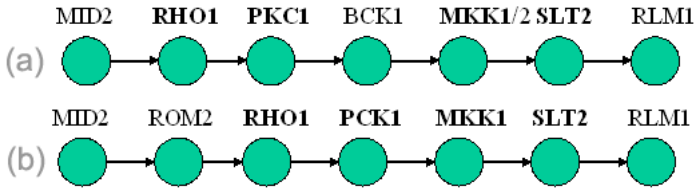

Figure 10.8: A comparison of the path obtained using the pathway discovery algorithm with the known pathway of MAP Kinase. **a** The known pathway. **b** The discovered pathway when forcing the same start and end points and the same length. The genes in **bold** designate genes common to the two paths.

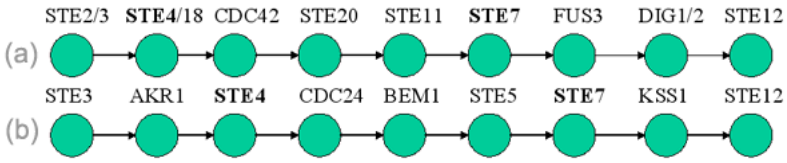

Figure 10.9: A comparison of the path obtained using the pathway discovery algorithm with the known pathway of Pheromone Response **a** The known pathway. **b** The discovered pathway when forcing the same start and end points and the same length. The discovered path contains proteins which are involved in pheromone response, but have a secondary role.

Segmented Pathways

When dealing with signal transduction pathways, we want the first protein in the path to be a membranal protein, and as we go down the path we want the obtained proteins to

become localized closer to the nucleus. Data about protein localization assayed *in vivo* can be incorporated to meet this demand. In addition to the color coding, each protein is given a number based on its proximity to the nucleus. When doing the Dynamic Programming, we will demand that the path is non-decreasing in terms of protein numbers. This approach was tested on known pathways and found to improve the results, as shown in Figure 10.10. In addition, this allows us to use less colors in the color coding.

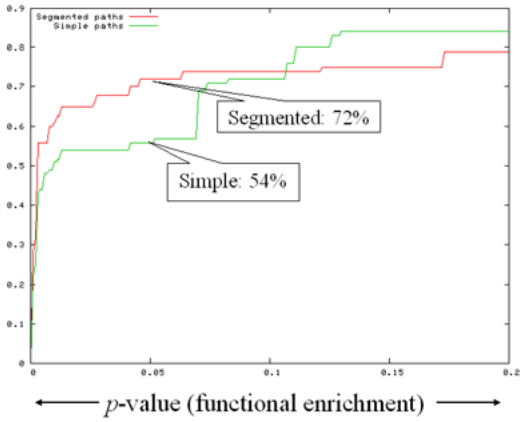

Figure 10.10: Performance comparison of the algorithm using segmented pathways (Three levels - membranal, cytoplasmatic, nuclear) as opposed an algorithm using simple pathways. Using segmentation, even as simple as this, improves performance.

Interval Constraints

Instead of assigning a number to each protein, it is possible to assign an interval. A path is said to be *consistent* if the lower bounds of the intervals are non-decreasing. The Dynamic Programming algorithm can be altered to handle intervals.

10.5.2 Extraction of Complexes

A complex is a set of proteins joined together to form a cellular machine. As a path in the network represents a pathway, a heavy subgraph in the network may represent a biological complex.
For further information on this topic refer to [8].

10.6 Algorithms on Multiple Networks

So far we have discussed construction and interpretation of single protein interaction networks. Now we shall discuss what can be done when we have more than one network. If a complex or a pathway are conserved in more than one network, then it is more likely to be true. When defining similarity between complexes or pathways in different networks, we

want a similarity on the protein-sequence level, as well as on the network topology level. An example of such similarity can be seen in Figure 10.11.

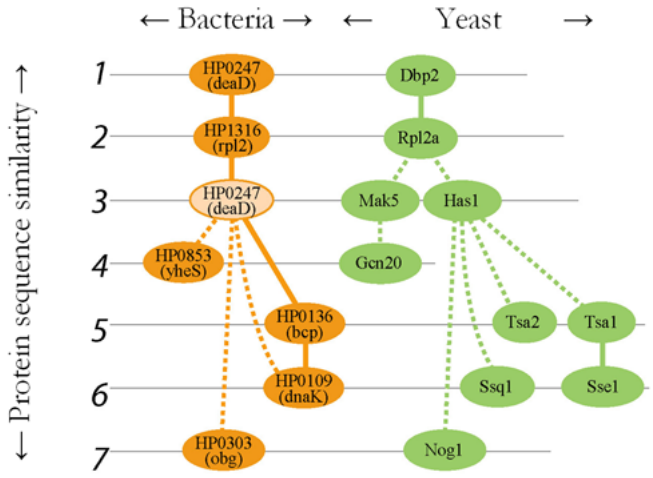

Figure 10.11: The homology between parallel pathways in bacteria and in yeast.

10.6.1 Network Alignment

We define a product graph as follows: Each node contains a set of similar proteins (by sequence similarity), one from each species. An edge between two nodes represents a conserved interaction between the proteins in one node and the proteins in the other node. A toy product graph example can be seen in Figure 10.13 [7].

10.6.2 Motivation

The above method can be used to produce different kinds of biological insights:
• Predict protein function
• Discover connections between known functions
• Decide on the best functional ortholog out of several sequence-match orthologs.
• Predict interactions in one species based on interactions between similar proteins in other species.

10.6.3 Validation of complexes

Validation of the results can be done in several ways.

Comparison to manual annotation

A predicted complex is said to be *pure* if at least half of its members belong to the same manually annotated complex. Out of the predicted complexes [7] 94% were pure, compared to 83% when applying the algorithm to yeast only.

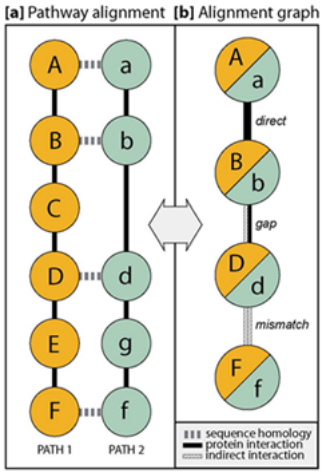

Figure 10.12: Example pathway alignment and merged representation. (a) Vertical solid lines indicate direct proteinprotein interactions within a single pathway, and horizontal dotted lines link proteins with significant sequence similarity (BLAST E value Ecutoff). An interaction in one pathway may skip over a protein in the other (protein C), introducing a "gap." Proteins at a particular position that are dissimilar in sequence (E value ¿ Ecutoff, proteins E and g) introduce a "mismatch." The same protein pair may not occur more than once per pathway, and neither gaps nor mismatches may occur consecutively. (b) Pathways are combined as a global alignment graph in which each node represents a homologous protein pair and links represent protein interaction relationships of three types: direct interaction, gap (one interaction is indirect), and mismatch (both interactions are indirect). Source: [2]

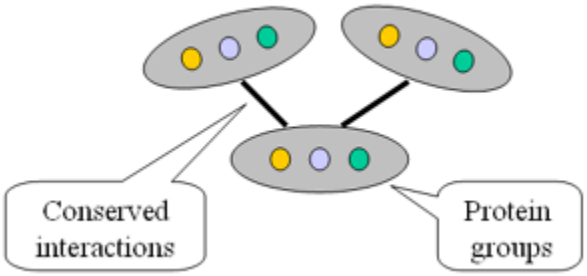

Figure 10.13: A product graph with 3 genes from 3 species. A node represents a set of sequence-similar proteins from the 3 species. The edges designate conserved interactions. Source: [7]

Function Prediction Cross-Validation

The proteins of a subnetwork are predicted to have a certain GO term if at least half the proteins in the subnetwork have this term. Cross-Validation of these predictions can be done by hiding the function of a portion of the known proteins, and checking if it is predicted

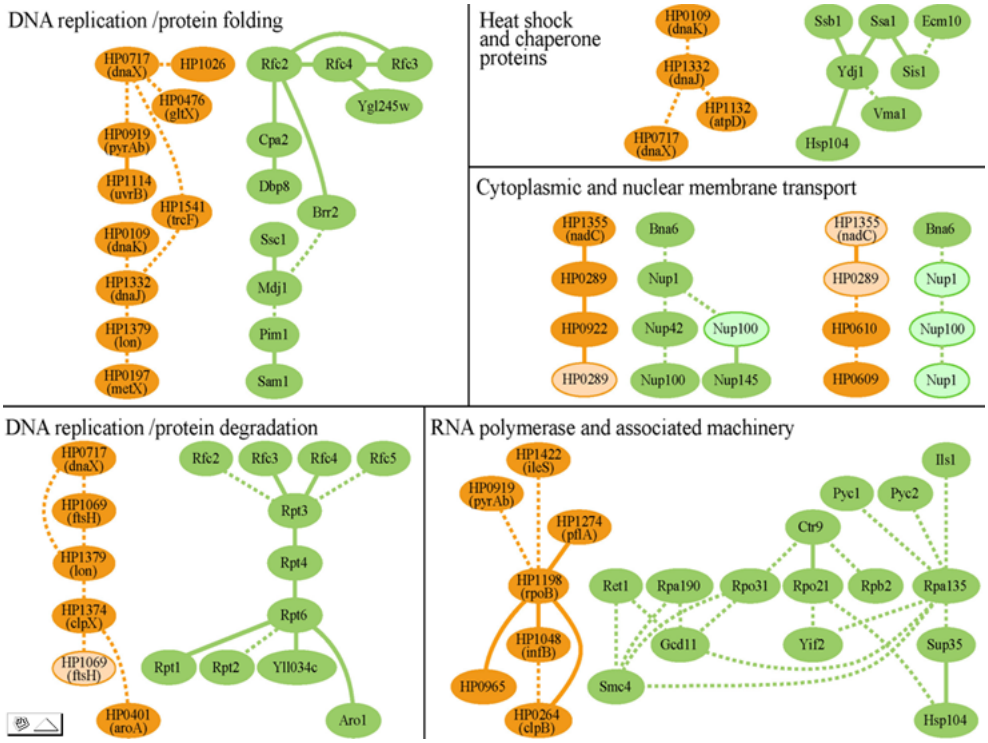

Figure 10.14: Top-scoring pathway alignments between bacteria and yeast. Source: [2]

accurately. Results are shown in Figure 10.16.

Interaction Prediction Cross-Validation

The same approach can be used to evaluate the accuracy of interaction prediction. A pair of proteins is predicted to interact if they co-occur in the same conserved complex, and there is evidence that a sequence-similar pair of proteins interact in the other two species. Results are shown in Figure 10.17.

Experimental Validation

Predicted interactions were experimentally tested. Results are shown in Figure 10.18

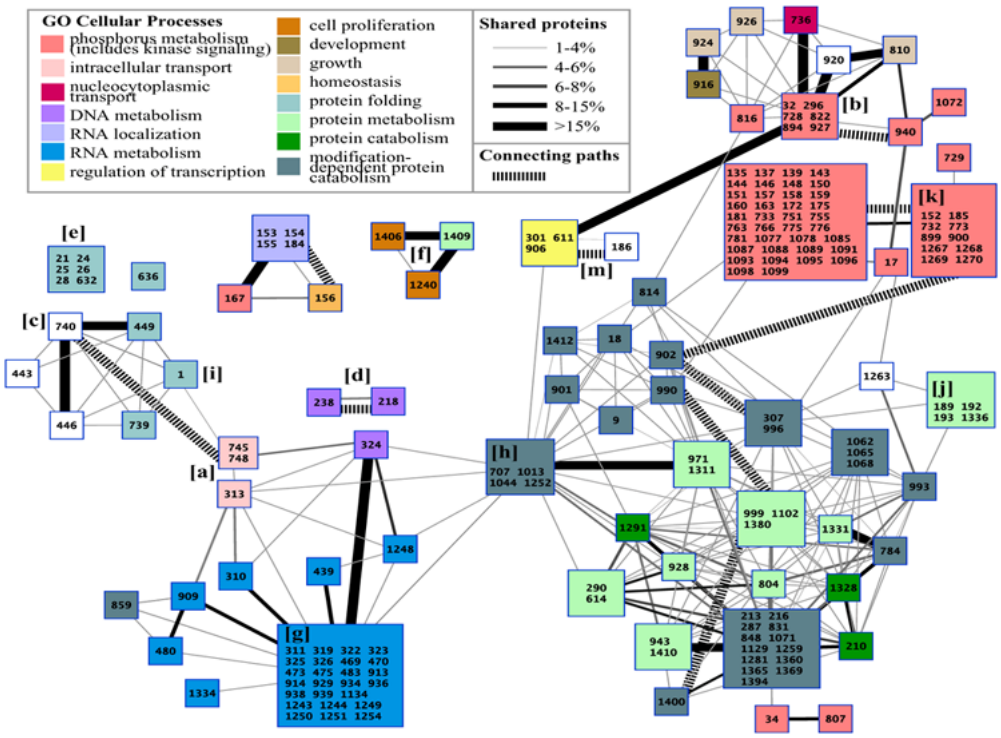

Figure 10.15: Modular structure of conserved clusters among yeast, worm, and fly. Multiple network alignment revealed 183 conserved clusters, organized into 71 network regions represented by colored squares. Regions group together clusters that share ¿15% overlap with at least one other cluster in the group and are all enriched for the same GO cellular process (P ¡ 0.05 with the enriched processes indicated by color). Cluster ID numbers are given within each square; numbers are not sequential because of filtering. Solid links indicate overlaps between different regions, where thickness is proportional to the percentage of shared proteins (intersection/union). Hashed links indicate conserved paths that connect clusters together. Source: [7]

Species	#Correct	#Predictions	Success rate (%)
Yeast	**114**	**198**	**58**
Worm	**57**	**95**	**60**
Fly	**115**	**184**	**63**

Figure 10.16: Cross validation results: Function Prediction

Species	Sensitivity (%)	Specificity (%)	*P-value*	Strategy
Yeast	50	77	1e-25	[1]
Worm	43	82	1e-13	[1]
Fly	23	84	5e-5	[1]
Yeast	9	99	1e-6	[2]+[1]
Worm	10	100	6e-4	[2]+[1]
Fly	0.4	100	0.5	[2]+[1]

Figure 10.17: Cross validation results: Interaction Prediction

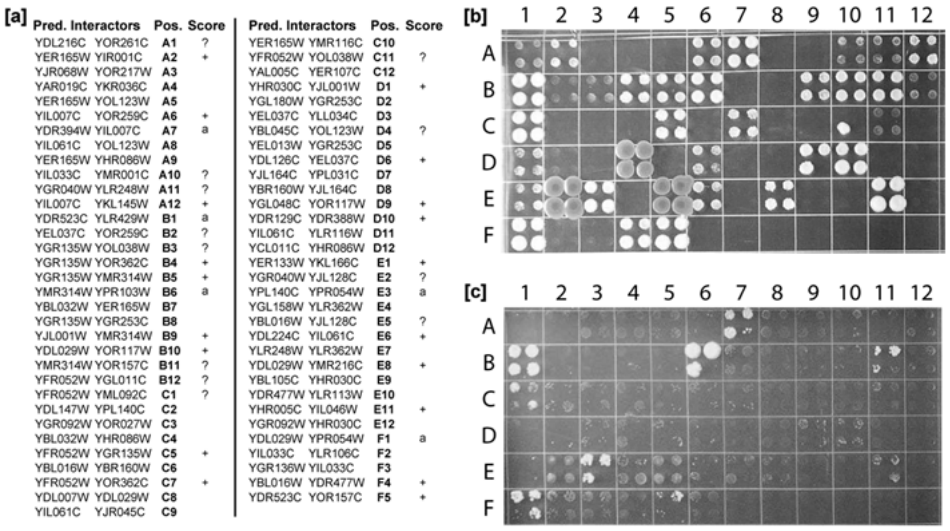

Figure 10.18: Verification of predicted interactions by two-hybrid testing. (a) Sixty-five pairs of yeast proteins were tested for physical interaction based on their co-occurrence within the same conserved cluster and the presence of orthologous interactions in worm and fly. Each protein pair is listed along with its position on the agar plates shown in b and c and the outcome of the two-hybrid test. (b) Raw test results are shown, with each protein pair tested in quadruplicate to ensure reproducibility. Protein 1 vs. 2 of each pair was used as prey vs. bait, respectively. (c) This negative control reveals activating baits, which can lead to positive tests without interaction. Protein 2 of each pair was used as bait, and an empty pOAD vector was used as prey. Activating baits are denoted by "a" in the list of predictions shown in a. Positive tests with weak signal (e.g., A1) and control colonies with marginal activation are denoted by "?" in a; colonies D4, E2, and E5 show evidence of possible contamination and are also marked by a "?". Discarding the activating baits, 31 of 60 predictions tested positive overall. A more conservative tally, disregarding all results marked by a "?", yields 19 of 48 positive predictions. Source: [7]

Bibliography

[1] N. Alon, Yuster R., and U. Zwick. Color coding. *J. ACM*, (42):844–856, 1995.

[2] Kelley BP, Sharan R, Karp RM, Sittler T, Root DE, Stockwell BR, and Ideker T. Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *Proc Natl Acad Sci U S A*, 100(20):11394–11399, September 2003.

[3] Wim Van Criekinge and Rudi Beyaert. Yeast two-hybrid: State of the art.

[4] Newman JR and Keating AE. Comprehensive identification of human bzip interactions with coiled-coil arrays. *Science*, 5628(3):2097–101, June 2003.

[5] Fromont-Racine M, Rain JC, and Legrain P. Toward a functional analysis of the yeast genome through exhaustive two-hybrid screens.

[6] Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M, Fields S, and Rothberg JM. A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae.

[7] Sharan R, Suthram S, Kelley RM, Kuhn T, McCuine S, Uetz P, and Sittler T. Conserved patterns of protein interaction in multiple species. *Proc Natl Acad Sci U S A*, 102(6):1974–1979, February 2005.

[8] Sharan R, Ideker T, Kelley B, Shamir R, and Karp RM. Identification of protein complexes by comparative analysis of yeast and bacterial protein interaction data. *RECOMB*, pages 282–289, 2004.

[9] Yook SH, Oltvai ZN, and Barabasi AL. Functional and topological characterization of protein interaction networks.

[10] Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, and Sakaki Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome.