Biological networks
Outline

1. Learning biological networks: experiments
2. Statistical properties of the networks
3. Understanding networks structure: motifs, modules, etc
4. From structure to function
5. Compare/align networks
6. Dynamics of networks
Outline

1. Learning biological networks: experiments
2. Statistical properties of the networks
3. Understanding networks structure: motifs, modules, etc
4. From structure to function
5. Compare/align networks
6. Dynamics of networks
Protein-protein Interactions

Yeast two-hybrid assay:
Does a protein A interact with B?

BAIT

Doesn’t active transcription

PREY

Doesn’t bind DNA

Activate transcription and grow on -his media
Protein-protein Interactions

**Large scale yeast two-hybrid assay:**
Find pairs of interacting proteins

A comprehensive two-hybrid analysis to explore the yeast protein interactome.
Ito T et al, *PNAS* 2001

A comprehensive analysis of protein-protein interactions in *S. cerevisiae*.
Protein-protein Interactions

2001

- proteins $m=4000$
- interactions $n=6500$
Protein-protein Interactions

Baker’s yeast:
9000 interactions
3000 proteins
(combined, 2006)
High-Throughput Yeast Two-Hybrid Assays for Large-Scale Protein Interaction Mapping

Albertha J. M. Walhout and Marc Vidal

Next-generation sequencing to generate interactome datasets

Figure 1 | Stitch-seq interactome mapping. (a) Outline of interactome mapping using different sequencing technologies. Each DNA fragment in each interacting pair is PCR-amplified individually and Sanger-sequenced; the association is tracked via position on the plate (top). Or each pair of DNA fragments is placed on the same PCR amplicon by PCR stitching; the amplicons are then collected and subjected to next-generation sequencing (bottom). (b) Outline of a PCR-stitching experiment.
High-Throughput Yeast Two-Hybrid Assays for Large-Scale Protein Interaction Mapping

Albertha J. M. Walhout and Marc Vidal

Next-generation sequencing to generate interactome datasets

Haiyuan Yu, Evan Weiner, Nenad Svrzakap, Edward Rietman, Kouros Salehi, Michael E Cusick, Pascal Braun.
High-Throughput Yeast Two-Hybrid Assays for Large-Scale Protein Interaction Mapping

Albertha J. M. Walhout and Marc Vidal

Next-generation sequencing to generate interactome datasets

(a) HI1 (2,754 Y2H interactions)  HI-NGS (1,166 Y2H interactions)

(b) Number of proteins vs. Number of Y2H interactors per protein
Evidence for Network Evolution in an *Arabidopsis* Interactome Map

*Arabidopsis Interactome Mapping Consortium*†

Plants have unique features that evolved in response to their environment, giving rise to complex cellular networks that underlie plant growth and development. To better understand these networks, we describe a proteome-wide binary protein-protein interaction map of the plant *Arabidopsis thaliana* containing about 6200 highly characterized proteins. A global organization of plant biological processes is observed in the resulting network, together with large numbers of links between proteins and pathways. We observe a dynamic interaction landscape, suggesting a model of evolution of protein networks. This and future plant interactome maps should facilitate understanding of plant biology and crop improvement.

Classical genetic and molecular approaches have provided fundamental insights into the organization of processes such as growth control or development and molecular descriptions of genotype-to-phenotype relationships for a variety of plant species. The *Arabidopsis* proteome is characterized by its high degree of functional conservation, with a complex architecture of interactions that is lacking from other plant species (e.g., rice, tomato, and S2), despite the high degree of conservation of genotype-to-phenotype relationships at the cellular level.

*All authors with their affiliations and contributions are listed at the end of the paper.
†To whom correspondence should be addressed. E-mail: marc.vidal@dfci.harvard.edu; edger@pslk.edu; pascal.braun@dfci.harvard.edu; david.hill@dfci.harvard.edu*

**5664 interactions between 2661 proteins**
Evidence for Network Evolution in an Arabidopsis Interactome Map

Arabidopsis Interactome Mapping Consortium†

Plants have unique features that evolved in response to their environments and ecosystems. A full account of the complex cellular networks that underlie plant-specific functions is still missing. We describe a proteome-wide binary protein-protein interaction map for the plant Arabidopsis thaliana containing about 6200 highly reliably detected 2700 proteins. A global organization of plant biological processes and analyses of the resulting network, together with large numbers of predicted links between proteins and pathways, we observe a dynamic rewiring of gene duplication events, providing evidence for a model of evolution for networks. This and future plant interactome maps should facilitate understanding plant biology and improve crops.

Classical genetic and molecular approaches have provided fundamental understanding of processes such as growth control or development and molecular descriptions of genotype-to-phenotype relationships for a variety of plant systems. Arabidopsis thaliana, the model plant, is functionally underrepresented by the biological organization and complex and dynamic processes it is lacking for Arabidopsis (A. thaliana) and A. suecica (A. suecica), depriving us of a genotype-to-phenotype relationship identified at the system level.

Fig. 3. Communities in Al-1_main (bottom) and in a typical randomized network (top left) (fig. S9). Only the largest connected component of each network is shown. Colored regions indicate enriched GO annotations summarized by the indicated terms (table S10). (Upper right) Distribution of randomized networks as a function of the total number and number of GO annotation enriched communities they contain. White arrow, position of the shown randomized network; red dot and arrow, position of Al-1_main; GA, gibberellic acid; JA, jasmonic acid; TCA, tricarboxylic acid.
Interaction landscape of membrane–protein complexes in *Saccharomyces cerevisiae*

Mohan Babu1,2*, James Vlasblom3,4*, Shuye Pu3, Xinghua Guo1, Chris Graham1, Björn D. M. Bean5, Helen E. Burston5, Franco J. Vizeacoumar1, Jamie Snider1, Sadhna Phanse1, Vincent Fong1, Yuen Yi C. Tam4, Michael Davey6, Olha Hnatshak1, Naigeet Bajaj1, Sharmanta Chandran1, Thanuja Punna1, Constantine Christopoulous1, Victoria Wong1, Analyn Yu1, Gouqing Zhong1, Joyce Li1, Igor Staglar1,2,6, Elizabeth Conibear1, Shoshana J. Wodak1,4,6, Andrew Emilii1,6 & Jack F. Greenblatt1,6

Figure 2 | Global organization of yeast MP complexes. Predicted MP clusters (subunits shown as similarly coloured nodes) inferred from the integrated network of high-confidence PPI (edges), demarcated according to primary compartment annotations. Representative complexes at the periphery highlight some of the findings of our study, including novel complexes and known complexes with new components. Our purifications were most successful for MPs localized to the Golgi and endoplasmic reticulum, a bias reflected in the highlighted examples. For each complex, previously reported components (red nodes), novel subunits (yellow nodes) and previously reported but not yet validated interactors (pink nodes) are displayed.
Agreement between databases

Literature curation of protein interactions: measuring agreement across major public databases

Andrei L. Turinsky1, Sabry Razick2,3, Brian Turner1, Ian M. Donaldson2,4 and Shoshana J. Wodak1,5,6,*

Figure 5. Pairwise agreement between databases for yeast-only and human-only co-citations. Shown is a pictorial summary of the agreement levels between pairs of databases for shared publications, where both databases annotated all the interactions.

The edge color indicates the fraction of the total number of shared (co-cited) publications contributed by the database pair. The edge color represents the agreement on less than half of the interactions or proteins, whereas shades of blue for agreement on more than half of interactions or proteins for human-only co-citations.
Regulatory network

Mining logic gates in prokaryotic transcriptional regulation networks

Rafael Silva-Rocha, Victor de Lorenzo*
Centro Nacional de Biotecnología, CSIC, Campus de Cantoblanco, Madrid 28049, Spain
Received 10 January 2008; accepted 28 January 2008
Measuring gene expression

Total RNA

- Oligo dT enrichment
- Fragmentation
- Random hexamer primed cDNA synthesis
  HiSeq™ 2000 sequencing
- Mapping to gene

Reference gene

Gene function analysis
Network of co-expression

**Human Gene Coexpression Network**

Gene A

Gene B

Pattern of expression changes

Coexpress (r=0.7)

Samples

**Conclusions/Significance:**

The identification of stable and reliable human gene to gene coexpression networks is essential to unravel the interactions and functional correlations between human genes at an omic scale. This work contributes to this research area.

**Methodology/Principal Findings:**

Exploration and analysis of gene expression data using genome-wide microarrays is a technique often used in genomic studies to find coexpression patterns and locate groups of co-transcribed genes. This kind of studies has been used in model organisms, like yeast, to discover gene functions, to define biological processes, and to find related transcription factors and their products. The wide microarrays is a technique often used in genomic studies to identify new non-described gene associations and it allows to place them in a functional context of some unknown non-assigned genes based on their interactions with known gene families. All these features leverage evolution and the high correlation of these groups with the high conservation of gene coexpression groups along bioinformatic studies are: the functional information associated to the placement in a functional context of some unknown non-assigned genes based on their interactions with known gene families. More than 60% are house-keeping and essential genes. The network displays new non-described gene associations and it allows to introduce strong bias and produce a lot of biological noise in the results. In fact, it is well known that cancer cells have altered disease datasets (mainly from cancer) in such meta-analyses may include heterogeneous datasets, mixing 'normal' samples with 'disease altered' samples, and, when they correspond to human, very often they include heterogeneous datasets, mixing normal with disease-altered samples. Moreover, the technical noise present in genome-wide expression global 'omic' scale are not focused in many cases on human research area.

Despite the described interest, coexpression studies done at the 'omic' scale are not focused on human samples and when they correspond to human very often include heterogeneous studies, mixing normal with disease-altered samples. The inclusion of many human gene expression large studies is a well reported problem that many times is not addressed with robust statistical methods, and the estimation of errors in the data is not provided.

The results provide a confident human gene coexpression network derived from tissue transcriptomic profiles. Further functional analysis of a subset core network, validated by two independent methods, are calculated by assignment to biological pathways. The results provide a confident human gene coexpression network avoiding both pathological and technical noise. To achieve this we describe a new method that combines several statistical and computational strategies: robust normalization and statistical validations; and estimation of the statistical accuracy and coverage of the data. All these methods provide a series of validations; and estimation of the statistical accuracy and coverage of the data. All these methods provide a series of validations.
Metabolic networks
Metabolic networks

(a) $S1 + S2 \xrightarrow{E} P1 + P2$

(b)

(c)

<table>
<thead>
<tr>
<th>Organism</th>
<th>Genes in Genome</th>
<th>Genes in Model</th>
<th>Reactions</th>
<th>Metabolites</th>
<th>Date of reconstruction</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Haemophilus influenzae</em></td>
<td>1,775</td>
<td>296</td>
<td>488</td>
<td>343</td>
<td>June 1999</td>
<td>[3]</td>
</tr>
<tr>
<td><em>Escherichia coli</em></td>
<td>4,405</td>
<td>660</td>
<td>627</td>
<td>438</td>
<td>May 2000</td>
<td>[5]</td>
</tr>
<tr>
<td><em>Saccharomyces cerevisiae</em></td>
<td>6,183</td>
<td>708</td>
<td>1,175</td>
<td>584</td>
<td>February 2003</td>
<td>[6]</td>
</tr>
<tr>
<td><em>Mycobacterium tuberculosis</em></td>
<td>4,402</td>
<td>661</td>
<td>939</td>
<td>828</td>
<td>June 2007</td>
<td>[10]</td>
</tr>
<tr>
<td><em>Bacillus subtilis</em></td>
<td>4,114</td>
<td>844</td>
<td>1,020</td>
<td>988</td>
<td>September 2007</td>
<td>[11]</td>
</tr>
<tr>
<td><em>Synechocystis sp. PCC6803</em></td>
<td>3,221</td>
<td>633</td>
<td>831</td>
<td>704</td>
<td>October 2008</td>
<td>[12]</td>
</tr>
<tr>
<td><em>Salmonella typhimurium</em></td>
<td>4,489</td>
<td>1,083</td>
<td>1,087</td>
<td>774</td>
<td>April 2009</td>
<td>[13]</td>
</tr>
<tr>
<td><em>Arabidopsis thaliana</em></td>
<td>27,379</td>
<td>1,419,</td>
<td>1,567</td>
<td>1,748</td>
<td>February 2010</td>
<td>[14]</td>
</tr>
</tbody>
</table>
Outline

1. Learning biological networks: experiments
2. **Statistical properties of the networks**
3. Understanding networks structure: motifs, modules, etc
4. From structure to function
5. Compare/align networks
6. Dynamics of networks
Graph theory

Nodes: proteins
Edges: interactions

• Node degree
• Components
• Shortest path
• Diameter=Longest Shortest Path
Graph theory

Nodes: proteins
Edges: interactions

• Node degree $d(k)$

• Components

• Shortest path

• Diameter = Longest Shortest Path
Graph theory

Nodes: proteins
Edges: interactions

- Node degree $d(k)$
- Components
- Shortest path
- Diameter = Longest Shortest Path
Graph theory

Nodes: proteins
Edges: interactions

• Node degree $d(k)$
• Components
• Shortest path
• Diameter = Longest Shortest Path
Graph theory

Nodes: proteins
Edges: interactions

- Node degree
- Components
- Shortest path
- Diameter = Longest Shortest Path
Graph theory

Nodes: proteins
Edges: interactions

- Node degree
- Components
- Shortest path
- Diameter = Longest Shortest Path
Graph theory

Nodes: proteins
Edges: interactions

• Node degree
• Components

• Shortest path

• Diameter = Longest Shortest Path
Graph theory

Nodes: proteins
Edges: interactions

• Node degree
• Components
• Shortest path
• Diameter = <Shortest Path>

Clustering coefficient

\[ c_i = \Pr\{\Delta_{jk} = 1 | \Delta_{ij} = \Delta_{ik} = 1\} \]
it’s not a random graph!

Andreas Wagner

Degree distribution

\[ P(d) \propto d^{2.5} \]
it’s not a random graph!

Table 1
Comparison of Statistical Features Between Random Graphs and the Yeast Protein Interaction Network

<table>
<thead>
<tr>
<th></th>
<th>Random Graphs</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yeast</td>
<td>ER</td>
<td>PL</td>
<td></td>
</tr>
<tr>
<td>Whole graph</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nodes</td>
<td>985</td>
<td>984.02 (10.39)</td>
<td>970.7 (81.57)</td>
<td></td>
</tr>
<tr>
<td>Degree</td>
<td>1.83</td>
<td>1.85 (0.98)</td>
<td>1.64 (1.76)</td>
<td></td>
</tr>
<tr>
<td>No. of components</td>
<td>163</td>
<td>108 (8)*</td>
<td>266.3 (30.6)*</td>
<td></td>
</tr>
<tr>
<td>Giant component</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nodes</td>
<td>466</td>
<td>624.0 (38.7)*</td>
<td>336.9 (86)</td>
<td></td>
</tr>
<tr>
<td>Degree</td>
<td>2.3</td>
<td>2.07 (1.05)</td>
<td>2.50 (2.6)</td>
<td></td>
</tr>
<tr>
<td>Clustering coefficient (×10⁻³)</td>
<td>22</td>
<td>0.59 (0.9)*</td>
<td>4.02 (2.3)*</td>
<td></td>
</tr>
<tr>
<td>Characteristic path length</td>
<td>7.14</td>
<td>15.88 (1.76)*</td>
<td>6.01 (1.14)</td>
<td></td>
</tr>
</tbody>
</table>

Wagner MBE 2000
it’s not a random graph!

The large-scale organization of metabolic networks

H. Jeong*, B. Tombor†, R. Albert*, Z. N. Oltvai† & A.-L. Barabási†

* Department of Physics, University of Notre Dame, Notre Dame, Indiana 46556, USA
† Department of Pathology, Northwestern University Medical School, Chicago, Illinois 60611, USA

In a cell or microorganism, the processes that generate mass, energy, information transfer and cell-fate specification are seamlessly integrated through a complex network of cellular constituents and reactions. However, despite the key role of these networks in sustaining cellular functions, their large-scale structure is essentially unknown. Here we present a systematic comparative mathematical analysis of the metabolic networks of

IT’S ALMOST **SCALE-FREE (＝POWER-LAW) GRAPH**

Figure 2 Connectivity distributions $P(k)$ for substrates. a, Archaeoglobus fulgidus (archae); b, E. coli (bacterium); c, Caenorhabditis elegans (eukaryote), shown on a log–log plot, counting separately the incoming (In) and outgoing links (Out) for each substrate. $k_{in}$ ($k_{out}$) corresponds to the number of reactions in which a substrate participates as a product (educt). The characteristics of the three organisms shown in a–c and the exponents $\gamma_{in}$ and $\gamma_{out}$ for all organisms are given in Table 1 of the Supplementary Information. d, The connectivity distribution averaged over all 43 organisms.
Random vs power-law

Other power-law networks


Other power-law networks:

• Metabolic network

• Network of social interactions:
  scientific collaborations, actors in films

• The Internet:
  links, physical connections
The web of human sexual contacts

Promiscuous individuals are the vulnerable nodes to target in safe-sex campaigns.

Unlike clearly defined ‘real-world’ networks, social networks tend to be subjective to some extent because the perception of what constitutes a social link may differ between individuals. One unambiguous type of contact is sexual contact, and hence an individual’s sexual behaviour can reveal the structure of a sexual-network that the cumulative distribution of number of different sex-partners in a year decays as a scale-free power-law with a similar exponent for both men and women. The scale-free networks of human sexual contacts is most efficient way to prevent sexually transmitted diseases.

Many real-world networks exhibit scale-free properties, where a few nodes have a very high degree, and the majority of nodes have a low degree. This is in contrast to random networks where degrees are distributed according to a Poisson distribution. The scale-free property is characterized by a power-law distribution of node degrees, which can be described by the equation $P(k) \sim k^{-\gamma}$, where $P(k)$ is the probability of a node having degree $k$, and $\gamma$ is the exponent.

**Figure 2** Scale-free distribution of the number of sexual partners for females and males. **a**, Distribution of number of partners, $k$, in the previous 12 months. Note the larger average number of partners for male respondents; this difference may be due to ‘measurement bias’ — social expectations may lead males to inflate their reported number of sexual partners. Note that the distributions are both linear, indicating scale-free power-law behaviour. Moreover, the two curves are roughly parallel, indicating similar scaling exponents. For females, $\alpha = 2.54 \pm 0.2$ in the range $k > 4$, and for males, $\alpha = 2.31 \pm 0.2$ in the range $k > 5$. **b**, Distribution of the total number of partners, $k_{tot}$ over respondents’ entire lifetimes. For females, $\alpha_{tot} = 2.1 \pm 0.3$ in the range $k_{tot} > 20$, and for males, $\alpha_{tot} = 1.6 \pm 0.3$ in the range $20 < k_{tot} < 400$. Estimates for females and males agree within statistical uncertainty.
Random vs power-law


The network of protein-protein interactions (and other molecular biological networks) are scale-free networks!

WHY?

• Scale-free networks are “better”...

OR/AND

• Biological networks became scale-free due to evolution.
Random

• Removal of a randomly picked node **significantly** increases the average path.
• All nodes are of equal “importance”.

Power-law

• Removal of a random node **slightly** increases the average path.
• Removal of a highly-connected node leads to drastic increase of the average path!

**Let’s check it**

**POWER LAW NETWORKS**

• Tolerant to random “attacks”,
• But more sensitive to targeted attacks!
Figure 2 Changes in the diameter $d$ of the network as a function of the fraction $f$ of the removed nodes. 

**a.** Comparison between the exponential (E) and scale-free (SF) network models, each containing $N = 10,000$ nodes and 20,000 links (that is, $k = 4$).
Figure 3 Network fragmentation under random failures and attacks. The relative size of the largest cluster $S$ (open symbols) and the average size of the isolated clusters $s$ (filled symbols) as a function of the fraction of removed nodes $f$ for the same systems as in Fig. 2. The size $S$ is defined as the fraction of nodes contained in the largest cluster (that is, $S = 1$ for $f = 0$).
Figure 3 Network fragmentation under random failures and attacks. The relative size of the largest cluster $S$ (open symbols) and the average size of the isolated clusters $s$ (filled symbols) as a function of the fraction of removed nodes $f$ for the same systems as in Fig. 2. The size $S$ is defined as the fraction of nodes contained in the largest cluster (that is, $S = 1$ for $f = 0$). a, Fragmentation of the exponential network under random failures (squares) and attacks (circles). b, Fragmentation of the scale-free network under random failures (blue squares) and attacks (red circles). The inset shows the error tolerance curves for the whole range of $f$, indicating that the main cluster falls apart only after it has been completely deflated.
Figure 3 Network fragmentation under random failures and attacks. The relative size of the largest cluster $S$ (open symbols) and the average size of the isolated clusters $s$ (filled symbols) as a function of the fraction of removed nodes $f$ for the same systems as in Fig. 2. The size $S$ is defined as the fraction of nodes contained in the largest cluster (that is, $S = 1$ for $f = 0$). a, Fragmentation of the exponential network under random failures (squares) and attacks (circles). b, Fragmentation of the scale-free network under random failures (blue squares) and attacks (red circles). The inset shows the error tolerance curves for the whole range of $f$, indicating that the main cluster falls apart only after it has been completely deflated
Random | Power-law

Equally stable to random failures

More sensitive to attacks

POWER-LAW NETWORKS

- Tolerant to random “attacks”
- But more sensitive to targeted attacks!
Outline

1. Learning biological networks: experiments
2. Statistical properties of the networks
3. Understanding networks structure: motifs, modules, etc
4. From structure to function
5. Compare/align networks
6. Dynamics of networks
Lethality and centrality in protein networks

The most highly connected proteins in the cell are the most important for its survival.
A simple physical model for scaling in protein–protein interaction networks

Eric J. Deeds*, Orr Ashenberg†, and Eugene I. Shakhnovich*‡

Fig. 1. Correlation between PPI networks. (a) The correlation between the network degree of a given protein in the Ito (13) and Uetz (12) data sets. Each point corresponds to a particular protein that exhibited interactions in both experiments. (b) A plot similar to (a) but comparing the ItoCore data set with Uetz.

Fig. 2. A physical model for PPI measurements. (a) A schematic of the model described in the text. Association free energies are largely the result of desolvation of the two protein surfaces. The overall burial of hydrophobic groups is represented by the sum of the contributions from each protein. (b) The distribution of surface hydrophobicities in yeast proteins. The fraction of surface residues that are hydrophobic (defined as residues AVILMFYW) is calculated according to the description in the supporting information. This distribution is taken from proteins in the Ito experiment (13). The red squares represent the model hydrophobicities sampled from a Gaussian distribution with the same mean and standard deviation as the Ito proteins themselves. (c) A degree distribution for the realization of the model used in (b). The cutoff was chosen such that the power-law fit gives an exponent of approximately -2.5, close to that of Ito graph. The degrees in this plot are shifted by +1 to allow for orphans (nodes of degree 0) to be displayed on a log–log plot. Note that the fraction of orphans in the graph is very high.
Network properties of genes harboring inherited disease mutations

Igor Feldman*, Andrey Rzhetsky†, and Dennis Vitkup*

PNAS | March 18, 2008
Network properties of genes harboring inherited disease mutations

Igor Feldman*, Andrey Rzhetsky††, and Dennis Vitkup*†

PNAS | March 18, 2008
degree centrality = degree

Closeness centrality (farness) =
   sum of its distances to all other nodes

Betweenness centrality of i = the number of shortest path
   between all other nodes which go through i
# Motifs

<table>
<thead>
<tr>
<th>Network</th>
<th>Nodes</th>
<th>Edges</th>
<th>$N_{\text{real}}$</th>
<th>$N_{\text{rand ± SD}}$</th>
<th>$Z$ score</th>
<th>$N_{\text{real}}$</th>
<th>$N_{\text{rand ± SD}}$</th>
<th>$Z$ score</th>
<th>$N_{\text{real}}$</th>
<th>$N_{\text{rand ± SD}}$</th>
<th>$Z$ score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene regulation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(transcription)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>E. coli</em></td>
<td>424</td>
<td>519</td>
<td>40</td>
<td>7 ± 3</td>
<td>10</td>
<td>203</td>
<td>47 ± 12</td>
<td>13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>S. cerevisiae</em></td>
<td>685</td>
<td>1,052</td>
<td>70</td>
<td>11 ± 4</td>
<td>14</td>
<td>1812</td>
<td>300 ± 40</td>
<td>41</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neurons</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>C. elegans</em>†</td>
<td>252</td>
<td>509</td>
<td>125</td>
<td>90 ± 10</td>
<td>3.7</td>
<td>127</td>
<td>55 ± 13</td>
<td>5.3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Food webs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Little Rock</td>
<td>92</td>
<td>984</td>
<td>3219</td>
<td>3120 ± 50</td>
<td>2.1</td>
<td>7295</td>
<td>2220 ± 210</td>
<td>25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ythan</td>
<td>83</td>
<td>391</td>
<td>1182</td>
<td>1020 ± 20</td>
<td>7.2</td>
<td>1357</td>
<td>230 ± 50</td>
<td>23</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>St. Martin</td>
<td>42</td>
<td>205</td>
<td>469</td>
<td>450 ± 10</td>
<td>NS</td>
<td>382</td>
<td>130 ± 20</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chesapeake</td>
<td>31</td>
<td>67</td>
<td>80</td>
<td>82 ± 4</td>
<td>NS</td>
<td>26</td>
<td>5 ± 2</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coachella</td>
<td>29</td>
<td>243</td>
<td>279</td>
<td>235 ± 12</td>
<td>3.6</td>
<td>181</td>
<td>80 ± 20</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Skipwith</td>
<td>25</td>
<td>189</td>
<td>184</td>
<td>150 ± 7</td>
<td>5.5</td>
<td>397</td>
<td>80 ± 25</td>
<td>13</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Motifs

Network Motifs: Simple Building Blocks of Complex Networks

R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, U. Alon

25 OCTOBER 2002  VOL 298  SCIENCE
Clusters in networks

Fig. 2. Fragment of the protein network. Nodes and interactions in discovered clusters are shown in bold. Nodes are colored by functional categories in

Protein complexes and functional modules in molecular networks

Victor Spirin and Leonid A. Mirny*

Harvard–MIT Division of Health Sciences and Technology, 16-343, Massachusetts Institute of Technology, 77 Massachusetts Avenue,
Extent of annotation of proteins in model species. For each species, the charts give the fractions and numbers of annotated and unannotated proteins, according to the three ontologies of the GO annotation. The numbers are based on the Entrez Gene and the WormBase databases as of September 2006.

FROM:
Network-based prediction of protein function
Roded Sharan, Igor Ulitsky & Ron Shamir
Figure 1. Illustration of the method. Subgraph of the protein interaction network of the yeast Saccharomyces cerevisiae. Proteins in gray boxes are unclassified (unknown function); the others are classified proteins (functions in brackets) and are labeled according to the following criteria: 1, cell growth; 2, budding, cell polarity and filament formation; 3, pheromone response, mating-type determination, sex-specific proteins; 4, cell cycle checkpoint proteins; 5, cytokinesis; 6, rRNA synthesis; 7, tRNA synthesis; 8, transcriptional control; 9, other transcription activities; 10, other pheromone response activities; 11, stress response; 12, nuclear organization. Given one of these proteins of unknown function, if we take as a prediction the function that appears more often in the neighbor proteins of known function, then we obtain the following classification (from top to bottom) YNL127W (2), YDR200C (3,4,10) and YLR238W (12). Our method, however, considers also the interactions among unclassified proteins. If we iterate once more the 'majority rule' by taking into account the interactions among the three unclassified proteins, we obtain the following classification: YNL127W (2,4), YDR200C (3,4,10) and YLR238W (12). This way we determined another possible function for YNL127W.
From network

FROM:
Network-based prediction of protein function
Roded Sharan, Igor Ulitsky & Ron Shamir
Alignment of networks

Cross-species analysis of biological networks by Bayesian alignment

Johannes Berg and Michael Lässig

Institut für Theoretische Physik, Universität zu Köln, Zölpicherstrasse 77, 50937 Cologne, Germany
Outline

1. Learning biological networks: experiments
2. Statistical properties of the networks
3. Understanding networks structure: motifs, modules, etc
4. From structure to function
5. Compare/align networks
6. Dynamics of networks
Evolution of power-law graphs

1. Growth
2. Preferential attachment

Albert and Barabasi 2000
Herbert A. Simon 1955
Yule 1925

<table>
<thead>
<tr>
<th>Phenomenon</th>
<th>Discovered</th>
<th>Re-discovered</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yule’s process</td>
<td>Yule (1925)</td>
<td>Fermi (1949)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Huberman and Adamic (1999)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Barabasi and Albert (1999)</td>
</tr>
<tr>
<td>Champernowne’s process</td>
<td>Champernowne (1953)</td>
<td>Levy and Solomon (1996)</td>
</tr>
<tr>
<td>Power law of word frequencies</td>
<td>Estoup (before 1916)</td>
<td>Condon (1928)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Zipf (1935)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Redner (1998)</td>
</tr>
</tbody>
</table>
Evolution of graphs

• Growth
  1. start with \( m_0 \) nodes
  2. add a node with \( m \) edges
  3. connect these edges to existing nodes
     at timestep \( t + m_0 \) nodes, \( tm \) edges
Evolution of graphs

• Preferential attachment

Probability $\Pi$ of connection to node $i$ depends on the degree $k_i$ of this node.

E.g. $\Pi(k_i) = \frac{k_i}{\sum_j k_j}$

“Rich gets richer”
Better evolution of graphs
A. Wagner, M. Lassig, A. Maritan etc

- Gene duplication
- Mutations
- Preferential attachment
More biological neutral evolution of graphs

A. Wagner, M. Lassig, A. Maritan, S. Redner etc.

• Gene duplication

• Mutations
More biological neutral evolution of graphs

A.Wagner, M.Lassig, A.Maritan, S.Redner etc.

- Gene duplication

- Mutations (rich gets richer)

=> Broad (not power-law) distribution!
More biological neutral evolution of graphs

- Gene duplication and re-wiring

Infinite-Order Percolation and Giant Fluctuations in a Protein Interaction Network

J. Kim¹, P. L. Krapivsky², B. Kahng¹, and S. Redner²

FIG. 1. Growth steps of the protein interaction network: The new node duplicates 2 out of the 3 links between the target node (shaded) and its neighbors. Each successful duplication occurs with probability $1 - \delta$ (solid lines). The new node also attaches to any other network node with probability $\beta/N$ (dotted lines). Thus 3 previously disconnected clusters are joined by the complete event.
Evidence for Network Evolution in an *Arabidopsis* Interactome Map

*Arabidopsis Interactome Mapping Consortium*†

Plants have unique features that evolved in response to their environments and ecosystems. A full account of the complex cellular networks that underlie plant-specific functions is still missing. We describe a proteome-wide binary protein-protein interaction map for the interactome network of the plant *Arabidopsis thaliana* containing about 6200 highly reliable interactions between about 2700 proteins. A global organization of plant biological processes emerges from community analyses of the resulting network, together with large numbers of novel hypothetical functional links between proteins and pathways. We observe a dynamic pattern of gene duplication events, providing evidence for a model of network evolution. This and future plant interactome maps should facilitate understanding plant biology and improving crops.

Classical genetic and molecular approaches have provided fundamental understanding of processes such as growth control or development and molecular descriptions of genotype-to-phenotype relationships for a variety of plant traits. However, the underlying protein-protein networks are largely unknown. We compiled a comprehensive interactome for *Arabidopsis* and assessed the biological organization of complex cellular networks. We show that the interactome is lacking evidence of a rapid whole-genome duplication (WGD), and that the process of WGD is occurring on a global scale.

![Diagram](image)

**Fig. 4.** Evidence for network evolution in Al-1MAD. (A) Interaction rewiring over time, according to the duplication-divergence model (24). (B) Average fraction of interactors shared between pairs of paralogous proteins with no (n = 4), low (n = 10), and high (n = 3) functional divergence (28). Error bars, mean ± SEM. P value, one-sided Kendall ranking correlation test (r, association) (3). (C) Average fraction of shared interactors, corrected for low experimental coverage (3), and average protein sequence identity between pairs of paralogous proteins as a function of the estimated time elapsed since duplication. Error bars, mean ± SEM (3). Dashed black line, corrected average fraction of shared interactors of nonparalogous pairs; myrs, million years. (D) Corrected average fraction of shared interactors (3) for pairs of paralogous proteins originating from polyploidy events (n = 109), as compared with other paralogous protein pairs of similar age (n = 147). Error bars, mean ± SEM (3). P values, Mann-Whitney U test. (E) Corrected average fraction of shared interactors (3) for pairs of paralogous proteins encoded by genes with high or low coexpression correlation (top and bottom tetrilc, respectively) as a function of phylogeny-based age group. Error bars, mean ± SEM (3). **, P < 0.05; **, P < 0.01; ***, P < 0.001.
Metabolic networks
Flux Balance Analysis

No accumulation of intermediates

\# of molecules in = \# of molecules out

Vin + Vout = 0

Example:

\[ 2A + B \rightarrow 3D \]
\[ D + C \rightarrow E \]

\[ (2V_A + V_B) / 3 = V_E \]
Flux Balance Analysis

Steady state Mass Balance

\[
\begin{align*}
A: & \quad -r_1 = -R_A \\
B: & \quad -r_1 + r_4 - r_2 - r_3 = 0 \\
C: & \quad +r_2 - r_5 - r_6 = +R_c \\
D: & \quad +r_3 + r_5 - r_4 - r_7 = +R_d \\
E: & \quad +r_6 + r_7 = +R_e
\end{align*}
\]

\[
\begin{bmatrix}
S \cdot v = b \\
\begin{bmatrix}
-r_1 & r_2 & r_3 & r_4 & r_5 & r_6 & r_7 \\
-1 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & -1 & -1 & 1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & -1 & -1 & 0 \\
0 & 0 & 1 & -1 & 1 & 0 & -1 \\
0 & 0 & 0 & 0 & 0 & 1 & 1
\end{bmatrix} \quad = \begin{bmatrix}
-R_A \\
0 \\
R_c \\
R_d \\
R_e
\end{bmatrix}
\end{bmatrix}
\]

Internal fluxes

Transportation fluxes
Flux Balance Analysis

Boundary

Steady state Mass Balance

\[ A: r_1 = -10 \]
\[ B: r_1 + r_4 - r_2 - r_3 = 0 \]
\[ C: r_2 - r_5 - r_6 - R_C = 0 \]
\[ D: r_3 + r_5 - r_4 - r_7 - R_D = 0 \]
\[ E: r_6 + r_7 - R_E = 0 \]

\[
\begin{bmatrix}
    r_1 & r_2 & r_3 & r_4 & r_5 & r_6 & r_7 & R_C & R_D & R_E \\
\end{bmatrix}
\begin{bmatrix}
    -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
    1 & -1 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
    0 & 1 & 0 & 0 & -1 & -1 & 0 & -1 & 0 & 0 \\
    0 & 0 & 1 & -1 & 1 & 0 & -1 & 0 & -1 & 0 \\
    0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\
\end{bmatrix}
\begin{bmatrix}
    0 \ & \ & \ & \ & \ & \ & \ & \ & -10 \\
\end{bmatrix}
\begin{bmatrix}
    0 \ & \ & \ & \ & \ & \ & \ & 0 \\
\end{bmatrix}
\]

10 Unknown fluxes

1 Known fluxes

\[ S' \cdot v' = b' \]
Total Number of fluxes = 11
Total number of known flux = 1
Total number of Metabolites = 5
Total number of d.f = 11-1+5=5
(i.e 5 possible solutions for this reaction network)
Flux Balance Analysis

• If cells optimize their growth rate then we need to find a solution that maximizes growth.
• Growth = biomass/time

+1BIOM-0.582GLY-0.0485MethylTHF-0.25GLN- 45.135ATP+44.96ADP+ 44.96Pi -
0.25GLU-0.176PHE-0.131TYR-0.205SER-0.054TRP-0.229ASP-0.229ASN-0.326LYS-
0.087CYS-0.146MET-0.241THR-0.276ILE-0.21PRO-0.281ARG-0.488ALA-0.402VAL-
0.428LEU-0.09HIS-0.203GTP-0.136UTP-0.126CTP-0.0247dATP-0.0254GTP-
0.0254dCTP-0.0247dTTP-0.00258PS -0.09675PE-0.02322PG-0.00645CL-
0.00785LPS-0.0276Pept-0.0341PTRSC-0.007SPRMD-0.154Glycogen;

Millimoles of metabolites present in 1 gm (dry wt.) of biomass
Flux Balance Analysis

- Input: stoichiometric matrix
  optimization function (biomass)
- Constrains

\[
\frac{dX}{dt} = S \cdot v - b = 0
\]

Maximize:
\[
Z = \sum c_i \cdot v_i = c \cdot v
\]
Flux Balance Analysis

- Linear programming
Flux Balance Analysis

- Linear programming
Flux Balance Analysis

• Effects of external conditions

• Effect of mutations

• Predictive cell physiology
Flux Balance Analysis

- Effect of C and N starvation
Flux Balance Analysis

• Effect of mutations and starvation
Predicting outcomes of knockouts

Using the Topology of Metabolic Networks to Predict Viability of Mutant Strains

Zeba Wunderlich* and Leonid A. Mirny†

*Biophysics Program, Harvard University, Cambridge, Massachusetts; and †Harvard-MIT Division of Health Sciences and Technology, Cambridge, Massachusetts
Predicting outcomes of knockouts

Using the Topology of Metabolic Networks to Predict Viability of Mutant Strains

Zeba Wunderlich* and Leonid A. Mirny†

*Biophysics Program, Harvard University, Cambridge, Massachusetts; and †Harvard-MIT Division of Health Sciences and Technology, Cambridge, Massachusetts
Networks

• Structure and dynamics of some biological network can be studied experimentally

• Networks don’t look like random graphs, more like power-law graphs.
  - results of neutral evolution
  - results of selection