The function of proteins

- Structural: the organelles of the cell
- Signaling: pass information from the environment and between different parts of the cell; turn genes on & off.
- Catalyze reactions (act as enzymes).
Proteins are the Building Blocks of Life

Their shape is instrumental in determining their function.
• Central dogma: DNA → mRNA → Proteins
• Proteins are building blocks of many cellular processes
• Conservation ⇒ functional importance

• Whole-genome (noisy) protein-protein interaction networks and other networks becoming available:
  - function annotation
  - combining graphs, assigning confidence, predicting edges, eliminating noise
  - comparing, searching graphs
  - figuring out how they evolved

• Start with experimental techniques for generating the graphs; then move on to network clustering.
• Central dogma: DNA → mRNA → Proteins
• Proteins are building blocks of many cellular processes

• Networks:

<table>
<thead>
<tr>
<th>Network</th>
<th>Nodes</th>
<th>Edges</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transcription (aka regulatory)</td>
<td>proteins/genes</td>
<td>A “regulates” B</td>
</tr>
<tr>
<td>Metabolic</td>
<td>Metabolites / small molecules</td>
<td>Reactions</td>
</tr>
<tr>
<td>Protein-Protein</td>
<td>Proteins</td>
<td>Physical Interactions</td>
</tr>
</tbody>
</table>
“Why” proteins interact:

Bring chains of enzymes together

“Tethered” Signal Transduction

From “Analysis of Biological Networks” Junker and Schreiber, eds
Experimental Techniques to Determine Protein Interactions

- Slow, accurate, costly:
  - X-ray crystallography
  - NMR

- High throughput, but noisy:
  - Yeast Two-Hybrid
  - TAP-MS (tandem affinity purification / mass spec)
Determining protein structure:

- X-ray crystallography
- NMR

  If you can determine structure of a complex, you know the position of each of its atoms.

- Slow, costly techniques.
- Don’t always work.

- More recent: high-throughput techniques
A Resource for Studying Biological Macromolecules

The PDB archive contains information about experimentally-determined structures of proteins, nucleic acids, and complex assemblies. As a member of the wwPDB, the RCSB PDB curates and annotates PDB data according to agreed upon standards.

The RCSB PDB also provides a variety of tools and resources. Users can perform simple and advanced searches based on annotations relating to sequence, structure and function. These molecules are visualized, downloaded, and analyzed by users who range from students to specialized scientists.

Molecule of the Month: Xanthine Oxidoreductase

Our diet includes a wide variety of different molecules. Many of these molecules are broken down completely and used to generate the metabolic energy that powers our cells. Others are disassembled piece-by-piece and recycled to build our own proteins and nucleic acids. The ones that are left over are broken down and discarded. Xanthine oxidoreductase, shown here...
the hemagglutinin structure of an avian H1N1 influenza A virus

3hto

DOI:10.2210/pdb3hto/pdb

Primary Citation

Molecular Description
Classification: Viral Protein
Structure Weight: 55101.73

Molecule: Hemagglutinin HA1 chain
Polymer: 1 Type: polypeptide(L) Length: 324
Chains: A

Molecule: Hemagglutinin HA2 chain
Polymer: 2 Type: polypeptide(L) Length: 160
Chains: B

Source
Polymer: 1
Scientific Name: Influenza a virus

Polymer: 2
Scientific Name: Influenza a virus

Polymer: 3

Oligomeric State: TRIMERIC
Other View of a Protein

AVIAN H5 HAEMAGGLUTININ

“Cartoon” drawing, showing major features such as alpha helices and beta sheets.
“Domain” = functional, evolutionary conserved unit of a protein
Yeast Two-Hybrid

Inside Yeast

Proteins of Interest

DNA Binding Domain

Transcription Activation Domain

Expressed
Scaling Up (Ito et al, 2001)

96-well plates
Each well contains a yeast strain with a different hybrid

~ 6,000 genes / 96
= 62 plates
= 3,844 crosses between plates
Mixed together and allowed to mate

96 x 96 combinations all mixed together

Gal4 activates 4 genes in the hybrids:
- ADE2 => adenine
- HIS3 => histidine
- URA3 => uracil
- MEL1

Kill off all strains that don’t express all 4 genes.

Sequence remaining hybrids
Ito et al., 2001 Results:

~ 18 million gene pairs; (prey,bait) & (bait, prey)

Involve 3278 proteins out of ~ 6,000

# of colonies that passed all 4 tests

Comparable overlaps between known interactions

**Table 1. Summary of the comprehensive two-hybrid screening**

<table>
<thead>
<tr>
<th>Interaction Type</th>
<th>Total</th>
<th>Known</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mating reactions</td>
<td>3,844</td>
<td></td>
</tr>
<tr>
<td>Combinations to be examined</td>
<td>~3.5 × 10^7</td>
<td></td>
</tr>
<tr>
<td>Positive colonies</td>
<td>15,523</td>
<td></td>
</tr>
<tr>
<td>ISTs</td>
<td>13,754</td>
<td></td>
</tr>
<tr>
<td>Independent two-hybrid interactions</td>
<td>4,549</td>
<td></td>
</tr>
<tr>
<td>More than 2 IST hits</td>
<td>1,533</td>
<td></td>
</tr>
<tr>
<td>More than 3 IST hits (core data)</td>
<td>841</td>
<td></td>
</tr>
</tbody>
</table>

**Table 2. Comparison between the two genomewide IST projects**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Total interactions</th>
<th>Known interactions(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uetz et al. (11)</td>
<td>691(^t)</td>
<td>88 (12.7)</td>
</tr>
<tr>
<td>This study</td>
<td></td>
<td></td>
</tr>
<tr>
<td>More than 2 IST hits</td>
<td>1,533</td>
<td>128 (8.3)</td>
</tr>
<tr>
<td>More than 3 IST hits (core data)</td>
<td>841</td>
<td>105 (12.5)</td>
</tr>
</tbody>
</table>

*Those described in the YPD (14) as previously known to associate or to occur in the same complex.

\(^t\)In Uetz et al. (11), total number of interactions revealed by IST approach was claimed to be 692, whereas their list contained 691 interactions.
What could go wrong with Yeast 2-Hybrid?
Low overlap!

Why?
- different experimental protocols
- different ways of making the hybrid genes (some fold correctly in Ito et al, but not in Uetz et al)
- actual randomness in binding
- Test takes place in nucleus, so proteins that never enter the nucleus won’t be tested.
- Will always be using yeast: so required post-translational modifications might not happen.
- Triple interactions: A - X - B
- Both proteins may not meet in vivo.

- Transcription factors can be hard to test (b/c they may activate the reporter gene w/o binding)
- hydrophobic / membrane proteins may not fold correctly.
Fruit Fly (*Drosophila melanogaster*)
Fly (*Drosophila melanogaster*) (Giot et al, 2003)

7,048 proteins
20,405 interactions

High-confidence:
4,679 proteins
4,780 interactions
Colored and placed by sub-cellular location

(Giot et al, 2003)
4 of the 6 highly connected proteins in fact are predicted by other means to be in the nucleus.
General Topological Properties

A. Degree $k_i = \text{number of links connected to node } i$

B. Distance $d_{ij} = \text{shortest path length between node } i \text{ and } j$

C. Diameter $D = \max \{d_{ij} | i, j \in N\}$, \( N \) : all nodes in the network

D. Clustering Coefficient $c_i = \frac{2e_i}{k_i(k_i - 1)}$, \( e_i \) : number of existing links (labeled in red) among the \( k_i \) nodes that connect to node \( i \)

E. Betweenness $b_l = \sum_{ij} p_{ij}(l) / p_{ij}$, \( p_{ij} \) : number of shortest paths between \( i \text{ and } j \), \( p_{ij}(l) \) : number of shortest paths between \( i \text{ and } j \) going through node \( l \)

Characteristic: Huge number of low-degree nodes, with a few very high-degree nodes.
Giant connected component
    = 3659 edges
    = 3039 nodes

Avg shortest path = 9.4 links, longer than expected in a random network (7.7 links)
Node disjoint loops

Real network contains more triangles than random
C. elegans
Li et al, 2004 Results:

<table>
<thead>
<tr>
<th>Interaction Set</th>
<th># interactions</th>
<th># proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td>Core</td>
<td>2157</td>
<td>502 baits</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1039 preys</td>
</tr>
<tr>
<td>Non-core</td>
<td>1892</td>
<td>531 baits</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1395 preys</td>
</tr>
</tbody>
</table>

- “Core” means they observed an interaction ≥ 3 times.
- Asymmetric (bait, prey) vs (prey, bait)
- Out of 2157 core pairs, only 22 were observed in both orientations
- 108 interactions in WormPD involved the tested proteins
- Core contain 8 of these interactions; Non-core contained 2
- Coverage = (8+2)/108 ≈ 10%
Estimating Reliability

\[ D = TP \times I + (1-TP) \times R \]

Assuming FP are ‘random-like’

\[ TP = \frac{(D - R)}{(I - R)} \]

Sprinzak et al, 2003
Estimating TP

\[ TP = \frac{(D - R)}{(I - R)} \]

- \( D \) = fraction co-localized predicted edges
- \( R \approx \) fraction of all pairs that are co-localized (~0.36)
- \( I \approx 1 \) or 0.95 [assumed to be very high]
### Table 1. Data sets of pairs of interacting proteins

<table>
<thead>
<tr>
<th>Experimental method category</th>
<th>Number of interacting pairs</th>
<th>Co-localization (%)</th>
<th>Co-cellular-role (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All: All methods</td>
<td>9347</td>
<td>64</td>
<td>49</td>
</tr>
<tr>
<td>A: Small scale Y2H</td>
<td>1861</td>
<td>73</td>
<td>62</td>
</tr>
<tr>
<td>A0: GY2H Uetz et al. (published results)</td>
<td>956</td>
<td>66</td>
<td>45</td>
</tr>
<tr>
<td>A1: GY2H Uetz et al. (unpublished results)</td>
<td>516</td>
<td>53</td>
<td>33</td>
</tr>
<tr>
<td>A2: GY2H Ito et al. (core)</td>
<td>798</td>
<td>64</td>
<td>40</td>
</tr>
<tr>
<td>A3: GY2H Ito et al. (all)</td>
<td>3655</td>
<td>41</td>
<td>15</td>
</tr>
<tr>
<td>B: Physical methods</td>
<td>71</td>
<td>98</td>
<td>95</td>
</tr>
<tr>
<td>C: Genetic methods</td>
<td>1052</td>
<td>77</td>
<td>75</td>
</tr>
<tr>
<td>D1: Biochemical, <em>in vitro</em></td>
<td>614</td>
<td>87</td>
<td>79</td>
</tr>
<tr>
<td>D2: Biochemical, chromatography</td>
<td>648</td>
<td>93</td>
<td>88</td>
</tr>
<tr>
<td>E1: Immunological, direct</td>
<td>1025</td>
<td>90</td>
<td>90</td>
</tr>
<tr>
<td>E2: Immunological, indirect</td>
<td>34</td>
<td>100</td>
<td>93</td>
</tr>
<tr>
<td>2M: Two different methods</td>
<td>2360</td>
<td>87</td>
<td>85</td>
</tr>
<tr>
<td>3M: Three different methods</td>
<td>1212</td>
<td>92</td>
<td>94</td>
</tr>
<tr>
<td>4M: Four different methods</td>
<td>570</td>
<td>95</td>
<td>93</td>
</tr>
</tbody>
</table>
Figure 1. True positive rates in various data sets that are distinguished by the experimental method for determining protein–protein interaction. Blue, percentage of TP based on co-localization ($l = 1$); red, green, percentage of TP based on shared cellular-role for $l = 1$ (red) and $l = 0.95$ (green). $l$ is the fraction of pairs with co-localized pair-mates in true interacting pairs (see the text). For the method categories see the legend to Table 1.
• High-throughput interaction detection
• Yeast two-hybrid - pairwise
  • organisms as machines to learn about organisms
  • yeast, worm, fly, human,...
  • low intersection between repeated experiments
  • *in vivo*, but takes place inside the nucleus.
  • Estimated 50% FP rate
  • statistics: shortest path distribution, degree distribution, # triangles, etc. show that Y2H graphs ≠ random graphs.

• TAP-MS (co-immunoprecipitation) - complexes: Simultaneous interactions between several proteins.
Need information? Your journey starts here.

Catalog | Worldcat
Books, journal titles, videos and other materials.

Research Port
Databases, journal articles, electronic journals.

Reference Shelf
Dictionaries

Ask us!
Communicate through

Library news
"Musical Milestones" on Flickr. If you missed this exhibit (about 100 Years of the University of Maryland Band) when it was on campus, see it on Flickr.

Celebrate Writing
Exhibit. "Celebrate Writing" (in McKeldin Library lobby) explores the evolution of writing, from
PubMed: Where nearly all the relevant papers can be found.


<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>Etymology</th>
<th>Amino Acid</th>
<th>Etymology</th>
</tr>
</thead>
<tbody>
<tr>
<td>glycine</td>
<td>greek, “Sweet” b/c it tastes sweet</td>
<td>asparagine</td>
<td>first found in asparagus</td>
</tr>
<tr>
<td>alanine</td>
<td>nonsense, euphonic</td>
<td>aspartic acid</td>
<td>similar to asparagine</td>
</tr>
<tr>
<td>leucine</td>
<td>greek, “white”, first isolated as white crystals</td>
<td>glutamine</td>
<td>first found in wheat gluten</td>
</tr>
<tr>
<td>isoleucine</td>
<td>isomer of leucine: same atoms, different arrangement</td>
<td>glutamic acid</td>
<td>similar to glutamine</td>
</tr>
<tr>
<td>proline</td>
<td>shorten “pyrrolidine”</td>
<td>lysine</td>
<td>greek, “a breaking up”, b/c first isolated in broken up molecules</td>
</tr>
<tr>
<td>phenylalanine</td>
<td>alanine + phenyl group</td>
<td>histidine</td>
<td>greek, “tissue” b/c first isolated from tissue protein</td>
</tr>
<tr>
<td>tyrosine</td>
<td>greek, “cheese” from which it was first isolated</td>
<td>arginine</td>
<td>latin “silver”, first isolated in combination with silver atom</td>
</tr>
<tr>
<td>tryptophan</td>
<td>greek, “trypsin-appearing” b/c first discovered in after action of trypsin</td>
<td>methionine</td>
<td>methyl group attached to sulfur atom (called theion in greek)</td>
</tr>
<tr>
<td>serine</td>
<td>latin, “silk”, from which it was first isolated</td>
<td>cystine</td>
<td>greek “bladder” b/c first isolated in bladderstone</td>
</tr>
<tr>
<td>threonine</td>
<td>related to sugar called ‘threose’</td>
<td>valine</td>
<td>related to valeric acid</td>
</tr>
</tbody>
</table>

Asimov, The Human Brain, 1965