There are many types of biological networks. Here's a small portion of a large metabolic network.
Contacts between proteins define protein interaction networks

X-ray structure of ATP synthase  
Schematic version  
Network representation

Total set = protein complex  
Sum of direct + indirect interactions
Let’s look at some of the types of interaction data in more detail.

Some of these capture physical interactions, some genetic, some informational or logical.

**Pairwise protein interactions**

In general, purifying proteins one at a time, mixing them, and assaying for interactions is far too slow & laborious. We need something faster! Hence, high-throughput screens, e.g. yeast two-hybrid assays
High-throughput yeast two-hybrid assays

Haploid yeast cells expressing activation domain-prey fusion proteins

Diploid yeast probed with DNA-binding domain-Pcf11 bait fusion protein


Protein complexes

High-throughput complex mapping by mass spectrometry

Tag → Bait → Affinity column → SDS-page → protein 1, protein 2, protein 3, protein 4, protein 5, protein 6

Trypsin digest, identify peptides by mass spectrometry
Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry


A variant: tandem affinity purification (TAP)

- Tag1 Tag2
- Bait
- Affinity column1
- Affinity column2
- + protease
- SDS-page
- Protein 1
- Protein 2
- Protein 3
- Protein 4
- Protein 5
- Protein 6
- Trypsin digest, identify peptides by mass spectrometry
~3,500 affinity purification experiments

~11K interactions / ~2.3K proteins

\( \rightarrow \) spans 556 complexes

Still daunting for the human proteome
The current state-of-the-art in animal PPI maps – co-fractionation/MS

>2,000 biochemical fractions, including replicates

>9,000 hours mass spec machine time

Genetic interactions

5.4 million gene-gene pairs assayed for synthetic genetic interactions in yeast

**Comparative genomics**

Functional relationships between genes impose subtle constraints upon genome sequences. Thus, genomes carry intrinsic information about the cellular systems and pathways they encode.

Linkages can be found from aspects of gene context, including:

- Distances between sequence elements
- Order of sequences
- Variation in order between organisms
- Regulatory sequences near genes
- Gene content of an organism
- Variation in gene content between organisms
- Fusions between genes from different organisms

**Phylogenetic profiles**

Organisms with e.g. a flagellum have the necessary genes; those without tend to lack them.

Specific trends of gene presence/absence thus inform about biological processes.

*PNAS 96, 4285-4288 (1999)*
Phylogenetic profiles

Grayscale indicates sequence similarity to closest homolog in that genome

Operons and evolutionary conservation of gene order

Prokaryotic operons tend to favor certain intergenic distances

Conserved gene neighbors also reveal functional relationships

Inferred operon in organism #1

Conserved gene neighbors in organism #2

Conserved gene neighbors in organism #3
Again, such observations can be turned into pairwise scores:

To summarize so far:

Data about gene interactions comes from many sources but is dominated by several major ones:

- mRNA co-expression. Historically microarrays & ESTs, increasingly RNAseq. Typically very high coverage data.

- Comparative genomics. Available for free for all organisms (typically phylogenetic profiles & operons)

- Protein interactions, especially co-complex interactions from mass spectrometry

- Genetic interactions (more so matching profiles of interaction partners than the interactions themselves)

- Transfer from other species
More abstractly, we might consider all of these as indicating “functional linkages” between genes

- Protein-protein interactions
- Participating in consecutive metabolic reactions
- Sharing genetic interactors
- Forming the same protein complex
- Giving rise to similar mutational phenotypes
- Exhibiting similar biological function
  and so on...

These sorts of data can be combined into functional gene networks

Including measurements of...

Gene expression (RNA-seq/arrays)

Protein expression and interactions (Mass spectrometry)

Gene organization (Genome sequences)

Gene-gene interactions (Genetic assays)

Yeast

Bacteria

C. elegans nematodes

Mouse

Xenopus frogs

Humans
These sorts of data can be combined into functional gene networks

Gene expression (RNA-seq/arrays)

Protein expression and interactions (Mass spectrometry)

Gene organization (Genome sequences)

Gene-gene interactions (Genetic assays)

These sorts of data can be combined into functional gene networks

Gene expression (RNA-seq/arrays)

Protein expression and interactions (Mass spectrometry)

Gene organization (Genome sequences)

Gene-gene interactions (Genetic assays)

These sorts of data can be combined into functional gene networks

Gene expression (RNA-seq/arrays)

Protein expression and interactions (Mass spectrometry)

Gene organization (Genome sequences)

Gene-gene interactions (Genetic assays)

Adapted from Fraser & Marcotte, Nature Genetics (2004)

Edward Marcotte/Univ. of Texas/BIO337/Spring 2014
In more detail: Constructing a functional gene network

- mRNA co-expression
- Protein physical interactions
- Synthetic lethal interactions
- Likelihood of genes’ functional association
- Other types of functional genomics data that may imply functional coupling between genes

(Infer associations by regression models in supervised Bayesian framework)

e.g., inferring functional linkages from mRNA co-expression across a given set of conditions

![Graphs showing co-expression](image)

Frequencies in the dataset (D) of gene pairs sharing pathway annotations (I) and not sharing annotations (~I), calculated per bin

\[ \text{LLS} = \ln \left( \frac{P(I|D)P(-I)}{P(I)P(-I)} \right) \]

Background frequencies of gene pairs sharing and not sharing annotations

Fit regression model, score all expressed gene pairs (annotated + unannotated)

Repeat for other datasets

Integrate scores for each link

Assess performance by cross-validation or bootstrapping

Typically calculated for many different sets of experiments sampling many different conditions...

![Graphs showing co-expression](image)

Each represents a different set of mRNA abundance profiling experiments (here, DNA microarrays) interrogating a given set of conditions.

Different gene pairs may be correlated in each condition.

... and so on

21 evidence types contribute to the HumanNet human gene network

Note: many are not human!

Table SI. Twenty-one different lines of evidence supporting HumanNet linkages.

<table>
<thead>
<tr>
<th>Data set</th>
<th>Description</th>
<th># genes</th>
<th># gene pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>CI-CC</td>
<td>Co-citation of worm genes</td>
<td>1,370</td>
<td>12,928</td>
</tr>
<tr>
<td>CI-CX</td>
<td>Co-expression among worm genes</td>
<td>2,633</td>
<td>41,645</td>
</tr>
<tr>
<td>CI-GT</td>
<td>Gene-rich interactions</td>
<td>1,040</td>
<td>2,430</td>
</tr>
<tr>
<td>CI-NC</td>
<td>Literature-curated worm protein physical interactions</td>
<td>1,402</td>
<td>2,640</td>
</tr>
<tr>
<td>CI-YH</td>
<td>High-throughput yeast 2-hybrid assays among worm genes</td>
<td>1,561</td>
<td>3,254</td>
</tr>
<tr>
<td>DS-PI</td>
<td>Putative protein interactions</td>
<td>6,153</td>
<td>15,738</td>
</tr>
<tr>
<td>HS-CC</td>
<td>Co-citation of human genes</td>
<td>3,423</td>
<td>6,152</td>
</tr>
<tr>
<td>HS-CX</td>
<td>Co-expression among human genes</td>
<td>11,050</td>
<td>150,315</td>
</tr>
<tr>
<td>HS-DG</td>
<td>Co-occurrence of domains among human proteins</td>
<td>1,357</td>
<td>38,707</td>
</tr>
<tr>
<td>HS-GN</td>
<td>Gene neighborhoods of bacterial and archaeal orthologs of human genes</td>
<td>3,504</td>
<td>36,457</td>
</tr>
<tr>
<td>HS-LC</td>
<td>Literature-curated linkages from protein-protein interaction DBs (HPID, BIND, BioGRID, IntAct, MINT) and Raafl et al., and pathway DBs (Reactome)</td>
<td>8,783</td>
<td>56,929</td>
</tr>
<tr>
<td>HS-MC</td>
<td>Human protein complexes from affinity purification/mass spectrometry</td>
<td>1,485</td>
<td>3,575</td>
</tr>
<tr>
<td>HS-PG</td>
<td>Co-inheritance of bacterial and archaeal orthologs of human genes</td>
<td>1,170</td>
<td>16,888</td>
</tr>
<tr>
<td>HS-YH</td>
<td>High-throughput yeast 2-hybrid assays among human genes</td>
<td>1,358</td>
<td>1,365</td>
</tr>
<tr>
<td>SC-CC</td>
<td>Co-citation of yeast genes</td>
<td>2,798</td>
<td>31,315</td>
</tr>
<tr>
<td>SC-CX</td>
<td>Co-expression among yeast genes</td>
<td>2,051</td>
<td>48,433</td>
</tr>
<tr>
<td>SC-GT</td>
<td>Yeast genetic interactions</td>
<td>2,584</td>
<td>17,678</td>
</tr>
<tr>
<td>SC-LC</td>
<td>Literature-curated yeast protein physical interactions</td>
<td>2,601</td>
<td>17,280</td>
</tr>
<tr>
<td>SC-MS</td>
<td>Yeast protein complexes from affinity purification/mass spectrometry</td>
<td>2,382</td>
<td>65,986</td>
</tr>
<tr>
<td>SC-TP</td>
<td>Yeast protein interactions inferred from tertiary structures of complexes</td>
<td>859</td>
<td>6,270</td>
</tr>
<tr>
<td>SC-YH</td>
<td>High-throughput yeast 2-hybrid assays among yeast genes</td>
<td>1,292</td>
<td>1,801</td>
</tr>
</tbody>
</table>

Evolutionary information is usually a key predictor—e.g., predictions for plant-specific traits often use fungal & animal data

For example, new seedling pigmentation genes...

...were predicted from both plant and animal data:
These networks are hypothesis generators. Given a gene, what other genes does it function with? What do they do?

Guilt-by-association in the gene network

Genes already linked to a disease or function

New candidate genes for that process

We can propagate annotations across the graph to infer new annotations for genes (network “guilt-by-association”, or GBA). Measuring how well this works on hidden, but known, functions gives us an idea how predictive it will be for new cases.

Query with genes already linked to a disease or function, e.g. the red or blue function

Assess the network’s predictive ability for that function using cross-validated ROC or recall/precision analysis

Infer new candidate genes for that process (e.g. predicting the green genes for the red function)

Calculating ROC curves

<table>
<thead>
<tr>
<th>Actual</th>
<th>Prediction</th>
<th>P</th>
<th>P'</th>
<th>N</th>
<th>N'</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>True Positive</td>
<td>False Positive</td>
<td>True Positive</td>
<td>False Negative</td>
<td>True Negative</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Basic idea: sort predictions from best to worst, plot TPR vs. FPR as you traverse the ranked list

TPR = TP / P = TP / (TP + FN)  
= True Positive Rate  
= Sensitivity, Recall

FPR = FP / N = FP / (FP + TN)  
= False Positive Rate  
= 1 - Specificity

Also useful to plot Precision \[ = TP / (TP + FP) \] vs. Recall \( = \text{TPR} \)

For example, predicting genes linked with worm phenotypes in genome-wide RNAi screens

Some very poorly predicted pathways:

ROC analysis indicates the likely predictive power of the network for a system of interest.

A poor ROC \( \rightarrow \) no better than random guessing.
A variety of algorithms have been developed for GBA

**GeneMANIA** Mostafavi et al., **Genome Biol.** (2008)

A variety of algorithms have been developed for GBA. Related to Google’s PageRank Ramakrishnan et al., Bioinformatics (2009)

Wang & Marcotte, **J. Proteomics** (2010)

The score of a gene is the sum of LLS edge weights to the query genes.

The score of a gene is the combination of the initial seeds and the weighted average of scores of the protein’s neighbors, defined iteratively.

Predicting genes for 318 *C. elegans* RNAi phenotypes.

Remarkably, this strategy works quite well

Some examples of network-guided predictions:

**In Arabidopsis:**
New genes regulating root formation

Lee, Ambaru et al., Nature Biotech (2010)

**In yeast:**
New mitochondrial biogenesis genes

Hess et al., PLoS Genetics (2009)

**In worms:**
Genes that can reverse ‘tumors’ in a nematode model of tumorigenesis

Lee, Lehner et al., Nature Genetics (2008)

**In mice/frogs:**
Functions for a birth defect gene

Gray et al., Nature Cell Biology (2009)

**In worms:**
Predicting tissue specific gene expression

Chikina et al., PLoS Comp Biology (2009)

Applicable to non-model organisms:
In rice: Identifying genes regulating resistance to *Xanthomonas oryzae* infection ...

Rice is the primary food source for >2 billion people worldwide
- >500 million tons rice/year are grown
- Bacterial rice blight destroys **up to 10-50% / year** in Africa/Asia

Summary of the major themes

- Gene networks serve as general frameworks for studying gene function
- Functional gene networks can be (re)constructed based upon millions of experimental observations via integrating these data into statistical models of functional connectivity among genes
- Guilt-by-association in such a network allows association of genes with functions, and genes with phenotypic traits, even highly polygenic ones
- Functional gene networks have been constructed for yeast, *C. elegans*, *Arabidopsis*, rice, mice, humans, many prokaryotes, and many other organisms...
Live demo of functional networks and Cytoscape