V4 In silico studies to predict protein protein contacts

The computational side of studying protein interactions can be split into two areas of activity:

(1) analysis on the macro level:
   map networks of protein interactions

(2) analysis on the micro level:
   understand structural mechanisms of interaction to predict interaction sites

Growth of genome data has stimulated a lot of research in area (1). Fewer studies have addressed area (2).

However, constructing detailed models of the protein-protein interfaces is important for comprehensive understanding of molecular processes, for drug design and for prediction the arrangement into macromolecular complexes.
Bioinformatic identification of interface patches

Statistical analysis of interfaces in crystal structures of protein-protein complexes shows that residues at interfaces

1 have a **different amino acid composition** than the rest of the protein.
   → can one predict protein-protein interaction sites from local sequence information?

2 are **evolutionary slightly more conserved** than other regions on the protein surface
   → identify conserved regions on protein surfaces

3 that are in contact and belong to different proteins may show **correlated mutations**
   → identify correlated mutations in multiple sequence alignments of various organisms

4 The interface often contains a central hydrophobic patch surrounded by a ring of polar or charged residues.
   → identify suitable patches on protein surface if 3D structure is known
Association pathway for protein-protein interaction

Steps involved in protein-protein association for a pair of proteins that electrostatically attract each other (not the case for all pairs):

- random diffusion (1)
- electrostatic steering (2)
- formation of encounter complex (3)
- dissociation or formation of final complex via TS (4)

Association pathway depends on:

- forces between the proteins
- solvent properties like temperature, ionic strength

Spar & Helms, JCTC (2005)
Example: prototypic binding of redox partners

Typical properties of interaction patches of electron transfer pairs:

- Electrostatic complementarity → fast association

- Inner ring of hydrophobic residues to promote binding affinity.

- Surrounding charged residues often do not form salt bridges across interface to allow fast dissociation (RC:c$_2$)

---

1 Analysis of interfaces

1812 non-redundant protein complexes from PDB (less than 25% identity).

Results don’t change significantly if NMR structures, theoretical models, or structures at lower resolution (altogether 50%) are excluded.

Most interesting are the results for transiently formed complexes.

<table>
<thead>
<tr>
<th>Type of interface</th>
<th>Number of contacts</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Internal</strong></td>
<td></td>
</tr>
<tr>
<td>Intra-domain</td>
<td>3,340,485</td>
</tr>
<tr>
<td>Domain–domain</td>
<td>255,144</td>
</tr>
<tr>
<td><strong>External</strong></td>
<td></td>
</tr>
<tr>
<td>Homo-obligomers</td>
<td>218,104</td>
</tr>
<tr>
<td>Homo-complexes</td>
<td>3077</td>
</tr>
<tr>
<td>Hetero-obligomers</td>
<td>18,886</td>
</tr>
<tr>
<td>Hetero-complexes</td>
<td>166,412</td>
</tr>
</tbody>
</table>

Contacts: residues were defined as in contact if the separation between the two closest atoms was ≤ 6 Å. We separated the following six types of interfaces (see Methods for details). (1) Intra-domain: contacts between residues in the same structural domains (according to the domain definition of PrISM[43]). (2) Domain–domain: contacts between residues in different structural domains in the same chain. (3) Homo-obligomers: contacts between residues on two different chains that have identical sequence and are permanent in the sense that we have no evidence for any biological interaction of the monomer. (4) Homo-complexes: contacts between residues on two different chains that have identical sequence and are transient in the sense that another chain of that sequence is observed in the cell as functional monomers. (5) Hetero-obligomers: contacts between two non-identical chains from the same protein (transient). (6) Hetero-complexes: contacts between two non-identical chains from two different proteins (permanent).

1 Amino acid composition of interface types

The frequencies of all residues found in SWISS-PROT were used as background when the frequency of an amino acid is similar to its frequency in SWISS-PROT, the height of the bar is close to zero. Over-representation results in a positive bar, and under-representation results in a negative bar.

1 Pairing frequencies at interfaces

red square: interaction occurs more frequently than expected;
blue square: it occurs less frequently than expected.

(A) Intra-domain: hydrophobic core is clear
(B) domain–domain,
(C) obligatory homo-oligomers,
(D) transient homo-oligomers,
(E) obligatory hetero-oligomers, and
(F) transient hetero-oligomers.

The amino acid residues are ordered according to hydrophobicity, with isoleucine as the most hydrophobic and arginine as the least hydrophobic.

→ propensities have been successfully used to score protein-protein docking runs.

2 NOXClass: Distinguish Permanent / Transient Complexes

Aim:
(1) distinguish different types of biological interactions (X-ray structures of protein-protein complexes).
(2) develop automatic classification scheme.

Zhu, Domingues, Sommer, Lengauer, BMC Bioinformatics 7, 27 (2006),
### Dataset

**Obligate Interactions (75)**

<table>
<thead>
<tr>
<th>1ahj A B</th>
<th>1b34 A B</th>
<th>1dce A B</th>
<th>1efv A B</th>
<th>1gux A B</th>
<th>1h2a L S</th>
<th>1luc A B</th>
<th>1lpxk A B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1i0q A B</td>
<td>1c0i A B</td>
<td>2aai A B</td>
<td>1aOOf A B</td>
<td>1a4i A B</td>
<td>1afw A B</td>
<td>1aj8 A B</td>
<td>1ajx A B</td>
</tr>
<tr>
<td>1aqm A B</td>
<td>1a3q A B</td>
<td>1b3e A B</td>
<td>1b5e A B</td>
<td>1b7b A C</td>
<td>1b8a A B</td>
<td>1b8j A B</td>
<td></td>
</tr>
<tr>
<td>1bmn A B</td>
<td>1bmn A B</td>
<td>1blm A B</td>
<td>1bfy A B</td>
<td>1byk A B</td>
<td>1c7n A B</td>
<td>1cli A B</td>
<td></td>
</tr>
<tr>
<td>1hj A C</td>
<td>1hss A B</td>
<td>1iss A B</td>
<td>1jkm A B</td>
<td>1kpe A B</td>
<td>1msp A B</td>
<td>1nse A B</td>
<td>1one A B</td>
</tr>
<tr>
<td>1pp2 L R</td>
<td>1qae A B</td>
<td>1qpx A B</td>
<td>1qrx A B</td>
<td>1qfe A B</td>
<td>1qfh A B</td>
<td>1qor A B</td>
<td>1qu7 A B</td>
</tr>
<tr>
<td>1smnt A B</td>
<td>1smnt A B</td>
<td>1spu A B</td>
<td>1spa A B</td>
<td>1vIt A B</td>
<td>1vok A B</td>
<td>1wgi A B</td>
<td>1xik A B</td>
</tr>
<tr>
<td>2u8g A B</td>
<td>3umk A B</td>
<td>4mdn A B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Non-obligate Interactions (62)**

| 1jav A C  | 1awv A B  | 1bvn T P  | 1cse E E  | 1eal C A  | 1f34 A B  | 1fss A B  | 1gdx F G  |
| 1lqx H A  | 1lsp A A  | 1lat B E  | 1lm2 B Z  | 1lms A B  | 1lq S A   | 1lqcd A D | 1d09 A B  |
| 1lqo A B  | 1fin A B  | 1dhk A B  | 1lql R G  | 1rrp A B  | 1ccO A E  | 1eg9 A B  |           |
| 1laz B C  | 1frv A B  | 3hhr A B  | 1lyc A B  | 1lar A L  | 1cmx A B  | 1bmn A C  |           |
| 2pce A B  | 3fGO A B  | 1lsf E E  | 1emw A B  | 1tua A B  | 1qbk B C  | 1hlu A P  | 1ibb A B  |
| 1tet A B  | 1jd A B  | 1lfd A B  | 1dnl A B  | 1tnm A B  | 1a9y A B  |           |           |

**Crystal Packing Contacts (106)**

| 1ks55     | 1ual     | 1mrx     | 1j9b     | 1e9g     | 1liup     | 1lis3     | 1gy7     | 1jz1     | 1jke     |
| 1km1      | 1hhr     | 2bttc    | 1eq9     | 1qi8     | 1k8u      | 1m7g      | 1p5z     | 1e19     | 1k75     |
| 1lac      | 1m96f    | 1h9c     | 1lqs     | 1b8z     | 1l5c      | 1gs5      | 1lve     | 1n20     | 1h4u     |
| 1k9u      | 1ls8     | 1e9s     | 1lqm     | 1j8b     | 1ldi      | 1lwy      | 1j24     | 1h1y     | 1ly      |
| 1lxxq     | 1lw6     | 1m7y     | 1n3i     | 1lms     | 1peO      | 1f6b      | 1l3p     | 1k3p     | 1j79     |
| 1lxni     | 1my7     | 1k4i     | 1lzt     | 1lt1m    | 1jdO      | 1nrv      | 1mvo     | 1m2d     | 1f7z     |
| 1gxo      | 1f6b     | 1b67     | 1lk2k    | 1lkm     | 1k94      | 1l0r      | 1euv     | 1lqO     | 1g2y     |
| 1m9h      | 1ed9     | 1dtd     | 1ld8     | 1ljt     | 1lc4      | 1n3z      | 1l6c     | 1l2m     | 1lcp     |
| 1h7r      | 1n2e     | 1ll2     | 1lubk    | 1g8q     | 1e87      | 1lj0      | 1lr8     | 1lqip    | 1ln9     |
| 1g60      | 1uaq     | 1lazu    | 1dnhl    | 1leye    | 1l52      | 1ljf      | 1b16     | 1e4m     | 3lyn     |

---

4. Lecture WS 2007/08

Bioinformatics III

9
Interface properties considered in NOX-Class

- Amino Acid Composition of the interface, Area-based
- Amino Acid Composition of the interface, Number-based
- CORrelation between amino acid compositions of interface and surface, Area-based
- CORrelation between amino acid compositions of interface and surface, Number-based
- Conservation Score of the interface, Area-based
- Conservation Score of the interface, Number-based
- ΔνDISTance between amino acid compositions of the interfaces, Number-based
- ΔνDISTance between amino acid compositions of the interfaces, Area-based
- Gap Volume
- Gap Volume Index
- Interface Area
- Interface Area Ratio
- Solvent Accessible Surface Area

![Diagram](image)

**Figure 1**
Schematic plot of the two-stage SVM. If an interaction is classified as crystal packing by SVM1, it will not be considered by SVM2; otherwise it is classified by SVM2 as either obligate or non-obligate interaction.

Zhu, Domingues, Sommer, Lengauer, BMC Bioinformatics 7, 27 (2006),
Crystal packing contacts have very small interfaces.

→ Obligate interfaces are on average larger than non-obligate interfaces.

Zhu, Domingues, Sommer, Lengauer, BMC Bioinformatics 7, 27 (2006),
The distributions of obligate and non-obligate interfaces are quite similar, but very different from crystal packing contacts.

Zhu, Domingues, Sommer, Lengauer, BMC Bioinformatics 7, 27 (2006),

\[
\text{Interface Area Ratio} = \frac{\text{Interface Area}}{\min(SASA_a, SASA_b)}
\]
Hydrophobic residues (FLIV) contribute twice as much to obligate interfaces as to crystal packing contacts.

Aromatic residues (FWY) tend to be more abundant in biological interfaces.

Zhu, Domingues, Sommer, Lengauer, BMC Bioinformatics 7, 27 (2006),

Figure 5
Area-based Amino Acid Composition for three types of interactions in the BNCP-C5 dataset.
### Good Performance

**Table 5: Performance of the two-stage SVM classifier**

<table>
<thead>
<tr>
<th>Interaction Type</th>
<th>OB(^b)</th>
<th>NO(^c)</th>
<th>CP(^d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>85.2%</td>
<td>88.1%</td>
<td>99.0%</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>92.0%</td>
<td>83.9%</td>
<td>96.2%</td>
</tr>
<tr>
<td>Specificity</td>
<td>96.3%</td>
<td>94.6%</td>
<td>97.1%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SVM stage</th>
<th>Stage 1</th>
<th>Stage 2</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>97.9%</td>
<td>86.4%</td>
<td>91.8%</td>
</tr>
</tbody>
</table>

\(^a\)Three out of the six properties (IA, IAR, and AACa) are used in the SVM classification for the BNCP-CS dataset;  
\(^b\)OB: Obligate interactions;  
\(^c\)NO: Non-obligate interactions;  
\(^d\)CP: Crystal packing contacts.
3 Multimeric threading: Fit pair A, B to complex database

Phase 1: single-chain threading.

Each sequence is independently threaded and assigned to a list of possible candidate structures according to the Z-scores of the alignments.

The Z-score for the $k$-th structure having energy $E_k$ is given by:

$$Z_k = \frac{E_k - \langle E \rangle}{\sigma}$$

where $\langle E \rangle$ and $\sigma$ are the mean and standard deviation values of the energy of the probe in all templates of the structural database.

For the assignment of energies, statistical potentials of residue pairing frequencies are used.

Library of 3405 protein folds where the pairwise sequence identity is < 35%.

Lu, ..., Skolnick, Genome Res 13, 1146 (2003)
Phase 2: a set of probe sequences, each at least weakly assigned to a monomer template structure that is part of a complex, is then threaded in the presence of each other in the associated quarternary structure.

If the interfacial energy and Z-scores are sufficiently favorable, the sequences are assigned this quarternary structure.

Library contains 768 dimer complexes (617 homodimers, 151 heterodimers).

Interfacial statistical potentials

Interfacial pair potentials \( P(i, j) \) \((i = 1...20, j = 1 \ldots 20)\) are calculated by examining each interface of the selected dimers in the database by:

\[
P(i, j) = -\log\left(\frac{N_{\text{obs}}(i, j)}{N_{\text{exp}}(i, j)}\right)
\]

where

\( N_{\text{obs}}(i, j) \) is the observed number of interacting pairs of \( i, j \) between two chains.

\( N_{\text{exp}}(i, j) \) is the expected number of interacting pairs of \( i, j \) between two chains if there are no preferential interactions among them.

\( N_{\text{exp}}(i, j) \) is computed as

\[
N_{\text{exp}}(i, j) = X_i \times X_j \times X_{\text{total}}
\]

where \( X_i \) is the mole fraction of residue \( i \) among the total surface residues.

\( X_{\text{total}} \) is the number of total interacting pairs.

Lu, Skolnick, Proteins 49, 350 (2002),
Dimer Template Structures from MULTIPROSPECTOR

2-stage protocol:
In phase I, both sequences X and Y are independently threaded using a set of suitable templates A and B.

Start phase II with decision whether the template structure pair $A_iB_j$ is part of a known complex. If $A_iB_j$ forms a complex continue multimeric threading to rethread on the partners in the complex and incorporate the protein-protein interfacial energies.

This step uses double-chain threading. It first fixes the alignment of X to the template A and adjusts the alignment of Y to the template B, and then it fixes the alignment of Y to the template B and adjusts the alignment of X to the template A.

Finally, the algorithm gives the template $A_iB_j$ that has the highest Z-score as a possible solution. At the same time, the algorithm provides the total energy of the complex as well as the interfacial energy.

Lu, Skolnick, Proteins 49, 350 (2002),
Genomic-scale prediction of protein-protein interactions

Out of 6298 unique ORFs encoded by *S. cerevisae*, 1836 can be assigned to a protein fold by a medium-confidence Z-score.

Result: 7321 predicted interactions between 1256 different proteins.

(Use this set for analysis).

Lu, ..., Skolnick, Genome Res 13, 1146 (2003)
Subcellular localization

Distribution of subcellular localization of yeast proteome (obtained from the YPD datatase at MIPS, Munich) compared with proteins involved in the predicted interactions

→ prediction is somehow biased towards the cytoplasmic compartment and against unknown locations.

Lu, ..., Skolnick, Genome Res 13, 1146 (2003)
Co-localization of interaction partners

Use localization data to assess the quality of prediction because two predicted interacting partners sharing the same subcellular location are more likely to form a true interaction.

Comparison of co-localization index (defined as the ratio of the number of protein pairs in which both partners have the same subcellular localization to the number of protein pairs where both partners have any sub-cellular localization annotation).

**Finding**: Multithreading predictions (MTA) are less reliable than high-confidence inter-actions, but score quite well amongst predictions + HTS screens.

Lu, ..., Skolnick, Genome Res 13, 1146 (2003)
Do partners have the same function?

Proteins from different groups of biological functions may interact with each other.

However, the degree to which interacting proteins are annotated to the same functional category is a measure of quality for predicted interactions.

Here, the predictions cluster fairly well along the diagonal.

Lu, ..., Skolnick, Genome Res 13, 1146 (2003)
Cofunctionality index

Cofunctionality index is defined as the ratio of the average protein interaction density for homofunctional interactions (diagonal of the matrix in $A$) to the average protein interaction density for heterofunctional interactions.

MTA method ranks third.

Lu, ..., Skolnick, Genome Res 13, 1146 (2003)
Correlation with mRNA abundance

Correlation between predicted interactions and mRNA abundance. The yeast proteome is divided into ten groups of equal size according to their mRNA expression levels and is arranged in an increasing abundance order from 1–10.

In contrast to other methods, MTA predictions are not correlated with abundance of mRNA expression. Method seems more capable of revealing interactions with low abundance.

Lu, ..., Skolnick, Genome Res 13, 1146 (2003)
Overlap between Large-Scale Studies

Unfortunately, the overlap of identified interactions by different methods is still very small.

<table>
<thead>
<tr>
<th></th>
<th>MTA</th>
<th>TAP</th>
<th>HMS</th>
<th>Y2H</th>
<th>RNA</th>
<th>SIL</th>
<th>SYN</th>
</tr>
</thead>
<tbody>
<tr>
<td>MTA</td>
<td>7321</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TAP</td>
<td>103</td>
<td>18027</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HMS</td>
<td>166</td>
<td>1728</td>
<td>33013</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Y2H</td>
<td>57</td>
<td>156</td>
<td>146</td>
<td>5125</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RNA</td>
<td>44</td>
<td>192</td>
<td>124</td>
<td>8</td>
<td>16496</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SIL</td>
<td>21</td>
<td>124</td>
<td>57</td>
<td>7</td>
<td>98</td>
<td>7446</td>
<td></td>
</tr>
<tr>
<td>SYN</td>
<td>37</td>
<td>55</td>
<td>37</td>
<td>17</td>
<td>2</td>
<td>5</td>
<td>886</td>
</tr>
</tbody>
</table>

*a* Our Multimeric Threading Algorithm.  
*b* TAP complexes.  
*c* HMS-PCI complexes.  
*d* Yeast two-hybrid.  
*e* Correlated mRNA expression.  
*f* Other *in silico* methods.  
*g* Synthetic lethality.  
*h* Number of interactions determined by the corresponding large-scale study.  
*i* Number of overlapping interactions between the corresponding large-scale studies.

Lu, ..., Skolnick, Genome Res 13, 1146 (2003)
4 Correlated mutations at interface


correlation information is sufficient for selecting the correct structural arrangement of known heterodimers and protein domains because the correlated pairs between the monomers tend to accumulate at the contact interface.

Use same idea to identify interacting protein pairs.
Correlated mutations at interface

Correlated mutations evaluate the similarity in variation patterns between positions in a multiple sequence alignment. Similarity of those variation patterns is thought to be related to compensatory mutations. Calculate for each positions $i$ and $j$ in the sequence a rank correlation coefficient ($r_{ij}$):

$$
r_{ij} = \frac{\sum_{k,l} (S_{ikl} - \bar{S}_i)(S_{jkl} - \bar{S}_j)}{\sqrt{\sum_{k,l} (S_{ikl} - \bar{S}_i)^2} \sqrt{\sum_{k,l} (S_{jkl} - \bar{S}_j)^2}}
$$

where the summations run over every possible pair of proteins $k$ and $l$ in the multiple sequence alignment.

$S_{ikl}$ is the ranked similarity between residue $i$ in protein $k$ and residue $i$ in protein $l$.

$S_{jkl}$ is the same for residue $j$.

$\bar{S}_i$ and $\bar{S}_j$ are the means of $S_{ikl}$ and $S_{jkl}$.

i2h method

Schematic representation of the i2h method.

A: Family alignments are collected for two different proteins, 1 and 2, including corresponding sequences from different species (a, b, c, ...).

B: A virtual alignment is constructed, concatenating the sequences of the probable orthologous sequences of the two proteins. Correlated mutations are calculated.

C: The distributions of the correlation values are recorded. We used 10 correlation levels. The corresponding distributions are represented for the pairs of residues internal to the two proteins (P11 and P22) and for the pairs composed of one residue from each of the two proteins (P12).

Predictions from correlated mutations

Results obtained by i2h in a set of 14 two domain proteins of known structure = proteins with two interacting domains. Treat the 2 domains as different proteins.

A: Interaction index for the 133 pairs with 11 or more sequences in common. The true positive hits are highlighted with filled squares.

B: Representation of i2h results, reminiscent of those obtained in the experimental yeast two-hybrid system. The diameter of the black circles is proportional to the interaction index; true pairs are highlighted with gray squares. Empty spaces correspond to those cases in which the i2h system could not be applied, because they contained <11 sequences from different species in common for the two domains.

In most cases, i2h scored the correct pair of protein domains above all other possible interactions.

Predicted interactions for *E. coli*

Number of predicted interactions for *E. coli*.

The bars represent the number of predicted interactions obtained from the 67,238 calculated pairs (having at least 11 homologous sequences of common species for the two proteins in each pair), depending on the interaction index cutoff established as a limit to consider interaction.

Among the high scoring pairs are many cases of known interacting proteins.

5 Construct complete network of gene association

Most network reconstructions focus on physical protein interaction and so represent only a subset of biologically important relations.

**Aim** here: construct a more extensive gene network by considering functional, rather than physical, associations.

**Idea:** each experiment, whether genetic, biochemical, or computational, adds evidence linking pairs of genes, with associated error rates and degree of coverage.

In this framework, gene-gene linkages are probabilistic summaries representing functional coupling between genes.

Only some of the links represent direct protein-protein interactions; the rest are associations not mediated by physical contact, such as regulatory, genetic, or metabolic coupling. All these represent functional constraints satisfied by the cell during the course of the experiments.

Method for integrating functional genomics data

- mRNA coexpression across 497 microarrays (CC)
- Gene-fusion network (P)
- Phylogenetic profiles (MI)
- Co-citation network (P)
- Protein interaction experiments

Unified scoring scheme (LLS)

Initial Integrated Network (IntNet)

Final Integrated Network (FinalNet)

Confident 34,000 linkages between 4,681 genes (ConfidentNet)

Hierarchical clustering of 627 gene modules (ModularNet)

Scoring scheme for linkages

Unified scoring scheme for linkages is based on a **Bayesian statistics** approach (see future lecture V8). Each experiment is evaluated for its ability to reconstruct known gene pathways and systems by measuring the likelihood that pairs of genes are functionally linked conditioned on the evidence, calculated as a log likelihood score:

\[
LLS = \ln \left( \frac{P(L|E)/\sim P(L|E)}{P(L)/\sim P(L)} \right)
\]

\(P(L|E)\) and \(\sim P(L|E)\): frequencies of linkages (L) observed in the given experiment (E) between annotated genes operating in the same pathway and in different pathways

\(P(L)\) and \(\sim P(L)\): the prior expectations (i.e., the total frequency of linkages between all annotated yeast genes operating in the same pathway and operating in different pathways).

Scores > 0 indicate that the experiment tends to link genes in the same pathway, with higher scores indicating more confident linkages.

Benchmarks

As scoring benchmarks, the method was tested against two primary annotation references:

(1) the Kyoto-based KEGG pathway database and
(2) the experimentally observed yeast protein subcellular locations determined by genome-wide green fluorescent protein (GFP)–tagging and microscopy.

KEGG scores were used for integrating linkages.
The other benchmark was withheld as an independent test of linkage accuracy.

Cross-validated benchmarks and benchmarks based on the Gene Ontology (GO) and COG gene annotations provided comparable results.
Functional inference from interaction networks

Benchmarked accuracy and extent of functional genomics data sets and the integrated networks. A critical point is the comparable performance of the networks on distinct benchmarks, which assess the tendencies for linked genes to share (A) KEGG pathway annotations or (B) protein subcellular locations. x axis: percentage of protein-encoding yeast genes provided with linkages by the plotted data; y axis: relative accuracy, measured as the of the linked genes’ annotations on that benchmark. The gold standards of accuracy (red star) for calibrating the benchmarks are smallscale protein-protein interaction data from DIP. Colored markers indicate experimental linkages; gray markers, computational. The initial integrated network (lower black line), trained using only the KEGG benchmark, has measurably higher accuracy than any individual data set on the subcellular localization benchmark; adding context-inferred linkages in the final network (upper black line) further improves the size and accuracy of the network.

Features of integrated networks

Portions of the final, confident gene network are shown for (C) DNA damage response and/or repair, where modularity gives rise to gene clusters, indicated by similar colors, and (D) chromatin remodeling, with several uncharacterized genes (red labels). Networks are visualized with Large Graph Layout (LGL).

Summary

The probabilistic gene network integrates evidence from diverse sources to reconstruct an accurate network, by estimating the functional coupling among yeast genes. These relations between yeast proteins are distinct from their physical interactions.

Applying this strategy to other organisms, such as human, is conceptually straightforward:
(i) assemble benchmarks for measuring the accuracy of linkages between human genes based on properties shared among genes in the same systems,
(ii) assemble gold standard sets of highly accurate interactions for calibrating the benchmarks, and
(iii) benchmark functional genomics data for their ability to correctly link human genes. Then integrate the data as described.

New data can be incorporated in a simple manner serving to reinforce the correct linkages. Thus, the gene network will ultimately converge by successive approximation to the correct structure simply by continued addition of functional genomics data in this framework.

Additional slides (not used)
Database of Dimer Template Structures

criteria:
1 The resolution of the two-chain PDB records should be < 2.5 Å.
2 The threshold for the number of interacting residues is set to be >30 to avoid crystallizing artifacts. Interacting residues are defined as a pair of residues from different chains that have at least one pair of heavy atoms within 4.5 Å of each other.
3 Each chain in the dimer database should have >30 amino acids to be considered as a domain.
4 Dimers in the database should not have >35% identity with each other.
5 The dimers should be confirmed in the literature as genuine dimers instead of crystallization artifacts.

This selection resulted in 768 dimer complexes (617 homodimers, 151 heterodimers)

Lu, Skolnick, Proteins 49, 350 (2002),
Which structural templates are used preferentially?

Structural groups of predicted interactions: the number of predictions assigned to the protein complexes in our dimer database. The 100 most populous complexes are shown. The *inset* is an enlargement for the top 10 complexes.

- 1KOB – twitchin kinase fragment
- 1IO9 – glycogen synthase kinase-3 beta
- 1AD5 – src family tyrosine kinase
- 1CKI – casein kinase I delta
- 1HCl – rod domain alpha-actinin
- 1CDO – liver class I alcohol dehydrogenase
- 1QBK – nuclear transport complex
- 1J7D – ubiquitin conjugating enzyme complex
- 1BLX – cyclin-dependent kinase CDK6/inhibitor
- 1QOR – quinone oxidoreductase

Lu, ..., Skolnick, Genome Res 13, 1146 (2003)
Features of integrated networks

At an intermediate degree of clustering that maximizes cluster size and functional coherence, 564 (of 627) modules are shown connected by the 950 strongest intermodule linkages. Module colors and shapes indicate associated functions, as defined by Munich Information Center for Protein Sequencing (MIPS), with sizes proportional to the number of genes, and connections inversely proportional to the fraction of genes linking the clusters.

Features of integrated networks

Adding context-inferred linkages increased clustering of genes, which produced a highly modular gene network with well-defined subnetworks.

We expected these gene clusters to reflect gene systems and modules. We could therefore generate a simplified view of the major trends in the network (Fig. 3B) by clustering genes of ConfidentNet according to their connectivities. Of the 4681 genes, 3285 (70.2%) were grouped into 627 clusters, reflecting the high degree of modularity.

Genes' functions within each cluster are highly coherent, and with 2 to 154 genes per cluster (ca. 5 genes per cluster on average), the clusters effectively capture typical gene pathways and/or systems.

5 Coevolutionary Analysis

Idea: if co-evolution is relevant, a ligand-receptor pair should occupy related positions in phylogenetic trees.

Goh & Cohen, 2002 showed that within correlated phylogenetic trees, the protein pairs that bind have a higher correlation between their phylogenetic distance matrices than other homologs drawn from the ligand and receptor families that do not bind.

Other Idea: analyze occurrence of proteins that can functionally substitute for another in various organisms.

Detect analogous enzymes in thiamin biosynthesis
Detect analogous enzymes in thiamin biosynthesis

Gene names are applied according to the first gene described from a group of orthologs.

**Solid black arrows** represent known or proposed reaction steps and **dashed black arrows** indicate unknown reactions. In addition, significant anticorrelations in the occurrence of genes across species (red arrows), and relevant *in silico* predicted protein-protein interactions (blue dashed arrows) are illustrated.

Distinct precursors have been proposed for different species3-5 (indicated in gray). Genes with orthologous sequences35 in eukaryotes and prokaryotes are in green; genes assumed to be prokaryote-specific are black. Interestingly, significant 'one-to-one' anticorrelations usually involve a prokaryote-specific and a 'ubiquitous' gene.

Abbreviations: AIR, 5-aminoimidazole ribonucleotide; Cys, cysteine; Gly, glycine; His, histidine; HMP, 2-methyl-4-amino-5-hydroxymethylpyrimidine; THZ, 4-methyl-5- -hydroxyethylthiazole; Tyr, tyrosine; Vit. B6, Vitamin B6.

THI-PP biosynthesis pathway: analogous genes

Negatively correlating gene occurrences are highlighted using the same colors. Species having at least two genes with a role unique to THI-PP biosynthesis are predicted to possess the functional pathway. The column 'STRING score' shows the most significant interaction for each gene, predicted using the STRING server. Predicted interaction partners are listed in the column 'Interact. with'. COG id: „id in groups of orthologous proteins server“

(a) Essential THI-PP biosynthesis enzymes, which are unique to the pathway.

(b) Essential THI-PP biosynthesis enzymes, which have been implicated in more than one biological process. The thiO gene, suggested to play a role in the pathway, was also added to that list.

(c) Proteins predicted in silico to be involved in the pathway.

4 analogies detected:

thiE can be replaced by MTH861
thiL by THI80
thiG by THI80
thiC by tenA

Interpretation

Proteins that functionally substitute each other have anti-correlated distribution pattern across organisms.

→ allows discovery of non-obvious components of pathways and function prediction of uncharacterized proteins and prediction of novel interactions.