7.91 Amy Keating

The Protein Interactome

A critical framework underlying systems biology

1. Overview - the many levels of systems biology
2. Experimental methods for measuring protein-protein interactions, and their limitations
3. Data sources for information about proteins and their interactions
4. Computational methods for assessing and predicting protein-protein interactions.

Spectrum of Systems Biology

detailed models - describe rates, concentrations, structure

low-resolution models - describe information flow, logic, mechanism

circuitry logic/control - positive and negative regulation

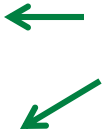
connectivity/topology - who talks to who? interaction scaffold

parts list - protein and DNA sequences (& structures)

Recommended reading: Ideker & Lauffenburger, TRENDS in Biotechnology (2003) 21, 255-262

Spectrum of Systems Biology

detailed models - differential equations



allow simulation & comparison with data

low-resolution models - Boolean & Markov models

circuitry logic/control - Bayesian networks

connectivity/topology - graph theory

parts list - databases

Recommended reading: Ideker & Lauffenburger, TRENDS in Biotechnology (2003) 21, 255-262

Spectrum of Systems Biology

detailed models - rates of individual reactions, protein concentrations in the cell, extent of phosphorylation, diffusion rates

low-resolution models - which elements are most crucial? combinatorial dependencies.

circuitry logic/control - Expression profiling, post-translational modifications in response to different stimuli. Identify pathways and clusters; does an interaction activate or repress; are multiple components required?

connectivity/topology - protein-protein, protein-DNA and protein-small molecule interactions

parts list - genome sequencing projects, gene finding algorithms, EST libraries

Recommended reading: Ideker & Lauffenburger, TRENDS in Biotechnology (2003) 21, 255-262

Spectrum of Systems Biology

detailed models **YAFFE**

low-resolution models - not covering this topic much this year

circuitry logic/control - **BURGE**

connectivity/topology - **KEATING (today)**

parts list **BURGE**

Recommended reading: Ideker & Lauffenburger, TRENDS in Biotechnology (2003) 21, 255-262

Protein-protein and protein-DNA interactions at the genomic level

Saccharomyces cerevisiae as a model organism.

A very simple eukaryote - "yeast as a model for human"

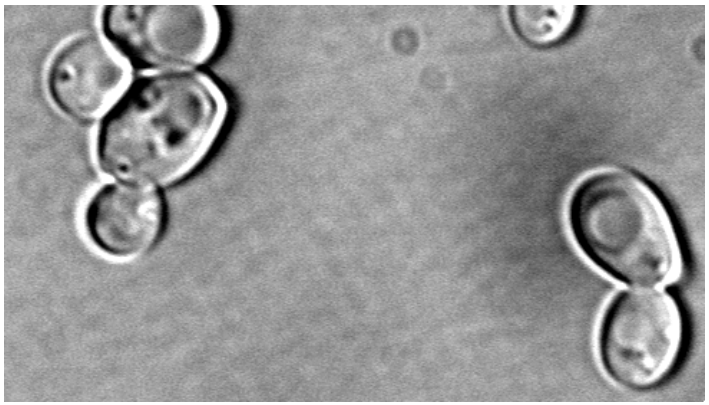
Genome 12,053 kb sequenced in 1996.

~5800 protein-coding genes.

Easy to do genetics in yeast.

Many regulatory and metabolic pathways are at least partly conserved between yeast and higher eukaryotes.

Many human disease genes have yeast orthologs.



Saccharomyces cerevisiae image from SGD™, provided by Peter Hollenhorst and Catherine Fox. Used with permission.

Small-scale interaction experiments

Protein-protein interactions

pull-down (GST, Ni affinity, co-immunoprecipitation)
cross-linking
more biophysical & quantitative: fluorescence, CD, calorimetry, surface plasmon resonance

Protein-DNA interactions

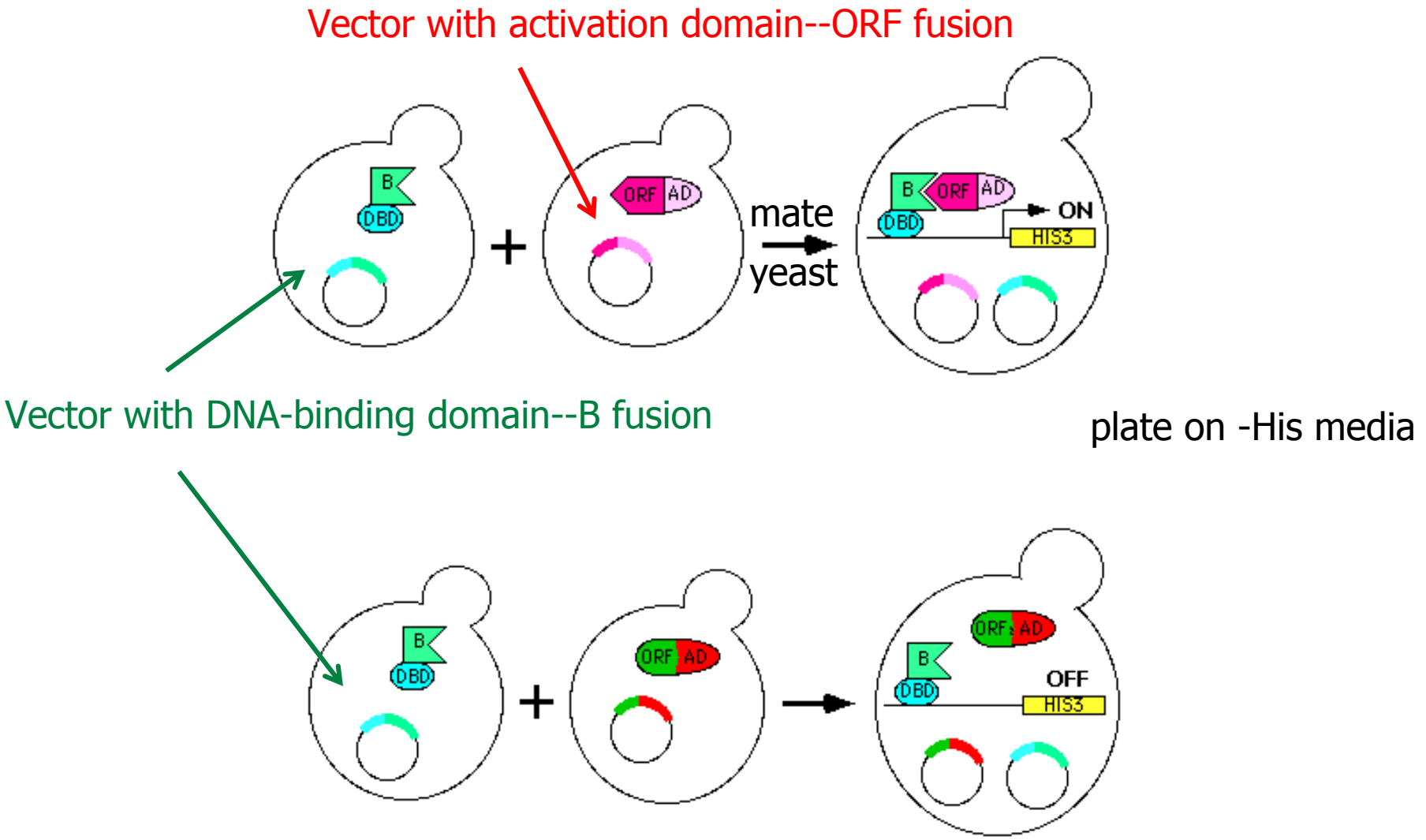
mostly by gel shift assay

Many, many thousands of such experiments have been done and reported in the literature, but how do you get the information out? This is hard, and an important problem in modern biology.

PreBIND is a machine learning application that can extract information about whether two proteins interact from the literature automatically.
http://www.blueprint.org/products/prebind/prebind.html

Small-scale experiment are generally the most reliable, though still rife with false negatives and false positives.

Yeast 2-hybrid assay



Images: http://depts.washington.edu/sfields/yp_interactions/YPLM.html

Yeast 2-hybrid assay

Pros

easy/fast

no purification required

in vivo conditions

can be adapted for
high-throughput screens

can detect transient interactions

Cons

prone to false negatives
- protein doesn't fold
- protein doesn't localize to nucleus
- interference from endogenous protein
- fusion protein doesn't interact like native
- protein fusion may be toxic to cell

prone to false positives
- auto-activation
- indirect ineractions

not quantitative

no control over post-translational modification

only test binary interactions

not quantitative

Yeast 2-hybrid assay for an entire genome

Uetz et al. Nature (2000) 403, 623-627
Two strategies:

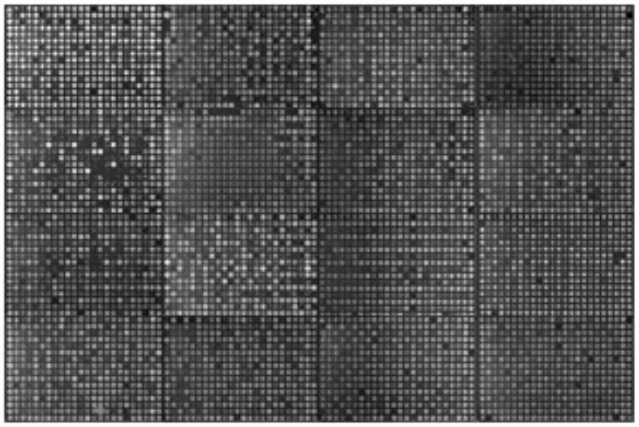
1. "array" approach: ~6,000 activation domain hybrid transformants mated to 192 DNA binding domain fusion transformants
   only 20% of interactions (281) reproducible (many auto-activate)
   3.3 positives per interaction-competent protein

2. "high-throughput screen" approach: 5,345 ORFs cloned separately into DNA-binding and activation domain plasmids (2 reporter genes); DBD fusions pooled and mated to AD fusions; 12 clones per pool sequenced, gave 692 unique interactions (472 seen more than once)
   1.8 positives per interaction-competent protein

Ito et al. PNAS (2001) 98, 4569-4574
For both DBD and AD, make 62 pools of ~96 proteins. Mate all pools against all. Gave 4,549 interactions; 841 observed ≥ 3 times (= core data).

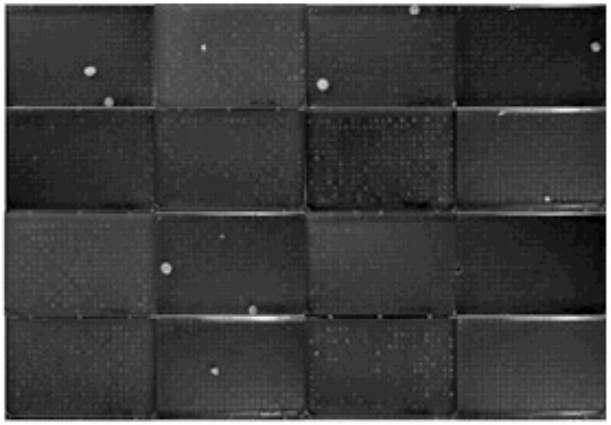
The potential number of interactions is huge, and the number of real interactions is probably very large (>10,000); these studies only characterize a tiny fraction (low coverage).

Example: Screen of the AD-Array with RPC19 as a bait



This is the activation domain array consisting of 6144 yeast colonies each expressing a different fusion of the GAL4 activation domain and one of the yeast ORFs.

The array is mated to the RPC19 bait and diploids are pinned to selective plates.



This is a set of selective plates (-Leu -Trp -His -Ade) on which only colonies grow that contain RPC19-interacting proteins:

Stan Field's web site
http://depts.washington.edu/sfields/images/RPC19.html

Additional "cons" when you do a large scale 2-hybrid screen

PCR amplification gives mutations - generally don't sequence everything to confirm!

Cloning & transformation inefficiencies

If baits are pooled, slow-growing cells will lose to faster ones, giving false negatives.

All vs. all assay contains many implausible interactions - proteins that aren't co-localized or expressed at the same time.

Can only sequence a small fraction of the positive clones.

High-throughput Y2H screens miss as many as 90% of Y2H interactions observed in focused, small-scale studies!

Affinity Purification

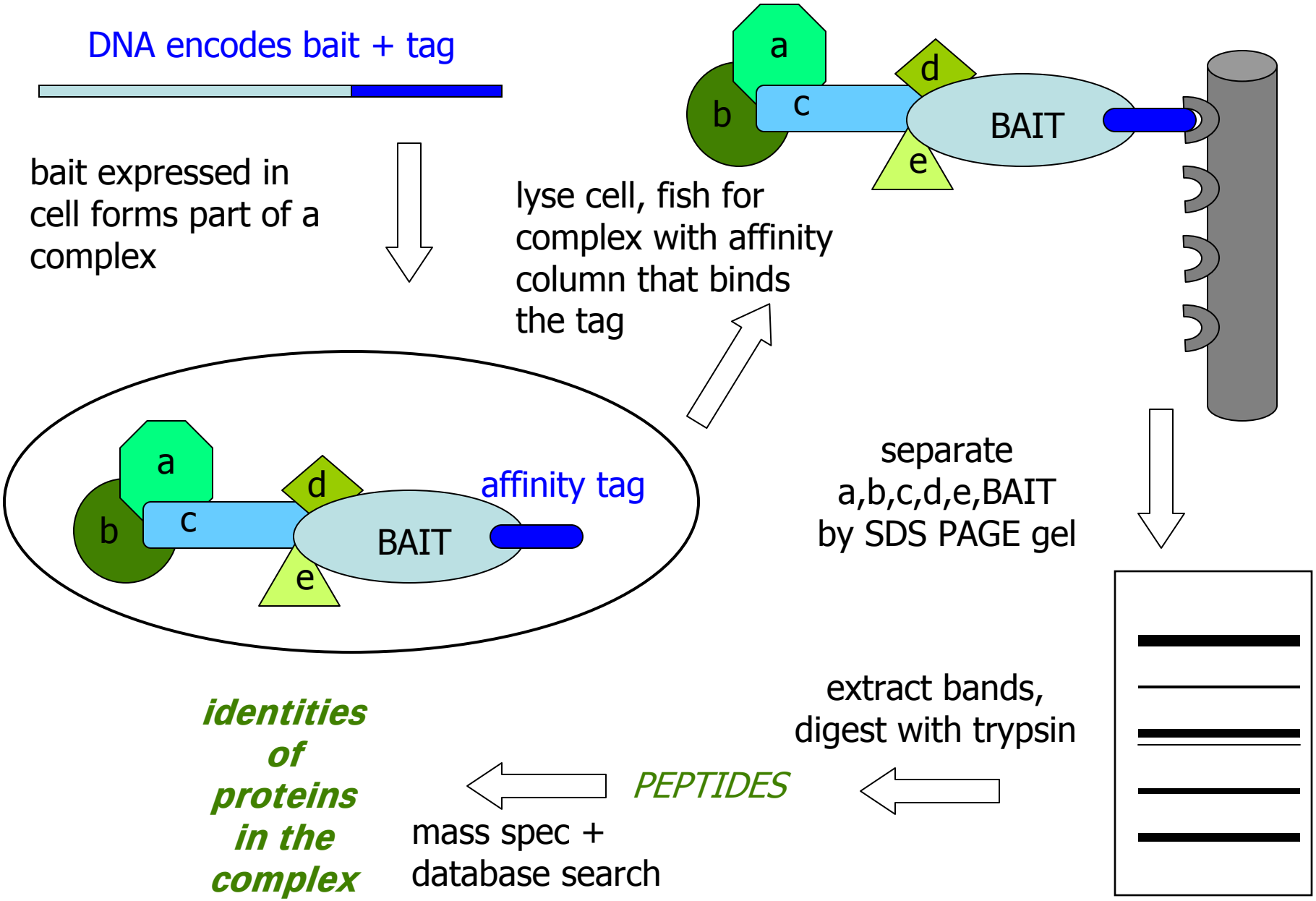
What do you mean by an "interaction"?

Most proteins interact with several other proteins (estimate 2-10).

Many proteins in the cell are found in complexes. For some purposes, knowing the identities of the members of the clusters is as useful, or more useful, than knowing the directly interacting partners.

Affinity purification is a method for characterizing the clusters directly, rather than one interaction at a time.

Affinity Purification/Mass spectrometry



Affinity purification/mass spectrometry for an entire genome

Gavin et al. *Nature* (2002) *415*, 141-147; Cellzome
1,167 bait proteins
TAP tag inserted at 3' end of gene; proteins under endogenous promoter
2 rounds of purification
232 distinct complexes with 2 to 83 proteins per complex
new cellular role proposed for 344 proteins

To assess confidence:
Repeat the experiment - only 70% reproducible using the same bait
Use different proteins in the complex as the bait, see if you recover the same proteins in the complex.

Ho et al. *Nature* (2002) *415*, 180-183; MDS Proteomics
725 bait proteins; 1,578 interacting proteins
FLAG tag, proteins transiently overexpressed

To assess confidence:
74% of interactions reproducible in small scale co-IP/blot

Affinity / ms assay

Pros

get the whole complex

proteins that purify together are likely to share a function

very sensitive - can detect ~15 copies per cell

in vivo conditions

can be adapted for high-throughput screens

Cons

doesn't determine direct interactions

not reliable for small proteins (< 15 kD)

affinity tag may interfere with interactions or with the function of essential proteins

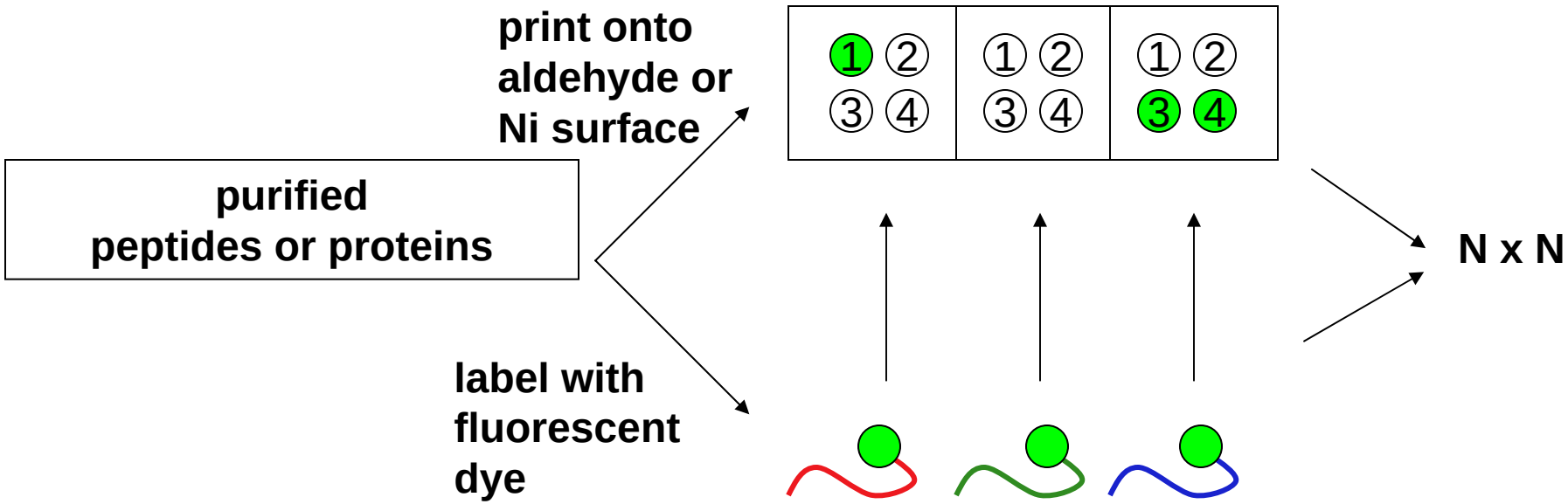
prone to false positives, e.g. "sticky" proteins

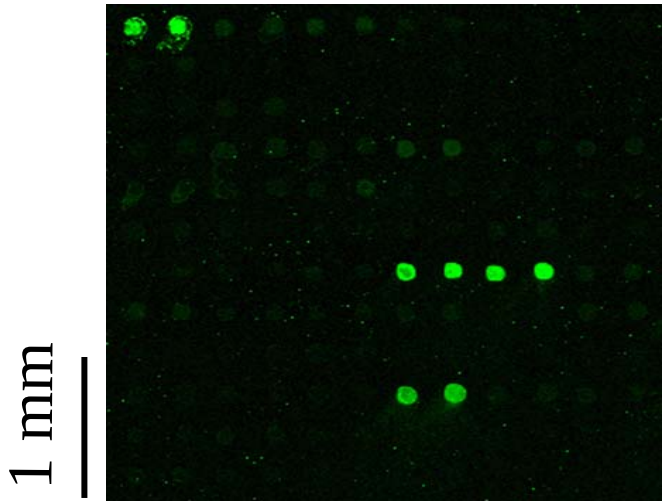
prone to false negatives

- won't get every protein every time
- complex must survive purification

not quantitative

Array Detection of Protein-Protein Interactions





Highly purified proteins were denatured using GdnHCl and printed onto aldehyde-derivatized glass slides using a commercial split pin arrayer. GdnHCl was to prevent homodimerization on the surface.

49 human proteins plus 3 duplicates plus 10 yeast proteins were printed in quadruplicate 62 times.

The 62 proteins were independently labeled with Cy-3 dye and denatured with GdnHCl.

Peptides were diluted from GdnHCl as they were added to the arrays. Following a brief incubation, slides were washed, dried and scanned, yielding NxN measurements, in quadruplicate of cc interactions.

The assay was repeated at concentration ranging from 160 pM to XXX nM.

Array Detection of Protein-Protein Interactions

MacBeath & Schreiber Science 2000

proof-of-principle for three types of interactions

protein-protein: protein G with IgG, FRAP with FKBP12, p50 with IκBα

protein-small molecule: biotin with steptavidin, Ab with DIG steroid ligand

enzyme-substrate: kinases PKA, Erk2

Zhu et al. Science 2001

assay of 5,800 yeast genes with calmodulin, phospholipids

Newman & Keating Science 2003

assay of ~48 x 48 human bZIP transcription factor coiled coils (plus 10 x 10 yeast)

Protein microarrays

Pros

N x N interactions at once

direct interaction assay

reagents can be well characterized

solution conditions are controlled

can be quantitative

requires very little protein

can be adapted for high-throughput screens

few false positives

Cons

tedious purification required, or else interactions may not be direct

surface may perturb folding or interactions

doesn't mimic in vivo conditions

not yet a mature technology - possibly not a good general approach

Overlap of high-throughput interaction studies is LOW

	Ito Y2H	Uetz Y2H	Gavin TAP/ms	Ho FLAG/ms
Ito 2-hybrid	4363	186	54	63
Uetz 2-hybrid		1403	54	56
Gavin affinity			3222	198
Ho affinity				3596
Small scale	442	415	528	391

data from Salwinski & Eisenberg, Current Opinion in Structural Biology (2003) 13, 377-382

Lesson:
Lots of protein-protein interaction data is now available for yeast, but it is not very reliable and it is not nearly comprehensive.

Nevertheless, these data have inspired the development of many computational methods.

To facilitate computational analysis, need to disseminate the data in a usable form!
This is often a rate limiting step in systems biology.

Databases that store interaction data

Database of Interacting Proteins (DIP)
Biomolecular Interaction Network Database (BIND)
Molecular Interactions Database (MINT)
INTERACT

MIPS contains interaction data (both direct and clusters) for yeast

Database of Interacting Proteins

[SEARCH:TOP][LOGIN]

Help
News
Register
Statistics
Satellites
Services
Articles
Search
Links
Files

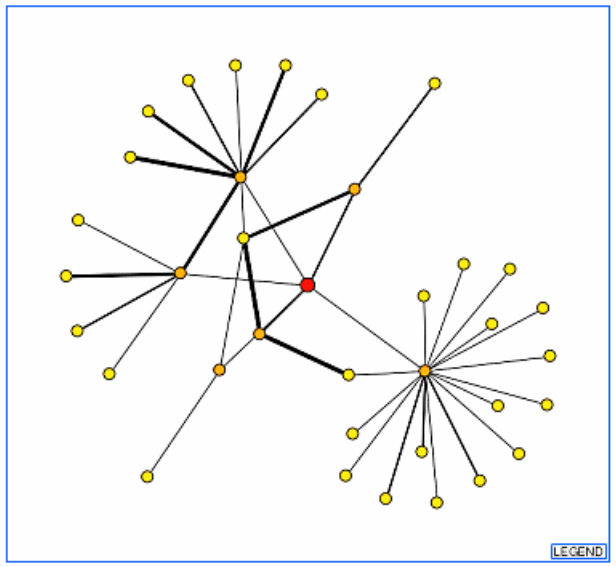
DIP 369N

BROWSE LINKS

Protein: cellular tumor antigen p53

Binary Complex Functional

DIP			Cross Reference			Protein Name/Description
Interaction	Interactor(s)	Links	PIR	SWISSPROT	GENBANK	
DIP:480E	DIP:1048N		TVHUF6	RAF1_HUMAN	gi:66762	protein kinase raf-1
DIP:522E	DIP:1074N		TVVPT4	---	gi:73275	large T antigen
DIP:40078E	DIP:24169N		---	Q64364	gi:6753390	p19ARF tumor suppressor protein
DIP:40079E	DIP:24196N		---	MDM2_MOUSE	gi:1209699	Ubiquitin-protein ligase E3 Mdm2
DIP:40140E	DIP:5978N		I38604	P531_HUMAN	gi:8928568	p53-binding protein 1
DIP:40141E	DIP:24266N		---	ASP2_HUMAN	gi:16197705	(Bbp)



partners of murine p53

Lab of David Eisenberg
http://dip.doe-mbi.ucla.edu

DIP interaction details

DIP 522E			
DIP 1074N	PIR TVVPT4	SwissProt	GenBank gi:73275
	Name/Description large T antigen		
DIP 369N	PIR DNMS53	SwissProt P53_MOUSE	GenBank gi:2144761
	Name/Description cellular tumor antigen p53		

Evidence Help

Type	Method	Details	Source
E	Two hybrid test	---	PMID:9043710
V	SMSC(1)	---	---

DIP 40078E			
DIP 369N	PIR DNMS53	SwissProt P53_MOUSE	GenBank gi:2144761
	Name/Description cellular tumor antigen p53		
DIP 24169N	PIR	SwissProt Q64364	GenBank gi:6753390
	Name/Description p19ARF tumor suppressor protein		

Evidence Help

Type	Method	Details	Source
E	Immunoprecipitation	---	PMID:9653180
V	SMSC(1)	---	---

http://dip.doe-mbi.ucla.edu

DIP interaction statistics

as of May, 2004

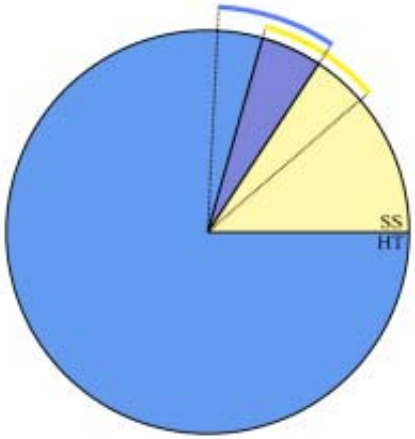
DATABASE STATISTICS	
Number of proteins	17048
Number of organisms	107
Number of interactions	44349
Number of distinct experiments describing an interaction	49104
Number of data sources (articles)	2694
Number of data sources (other)	34

ORGANISM	PROTEINS	INTERACTIONS	EXPERIMENTS	Details
Drosophila melanogaster (fruit fly)	7052	20988	21012	
Saccharomyces cerevisiae (baker's yeast)	4749	15658	19143	

http://dip.doe-mbi.ucla.edu

Saccharomyces cerevisiae
(baker's yeast)

PROTEINS	INTERACTIONS	#Exp	#Int
4749	15658	1	13636
		2	1270
		3	402
		4	165
		5	81
		6+	98



Yeast interactions by experiment type:

SS - small-scale experiments

HT - high-throughput experiments

SS/HT overlap - *purple*

Bars mark interactions that were indentified in more than one experiment.

http://dip.doe-mbi.ucla.edu

BIND

Designed to hold direct interaction, cluster and pathway data
81,000 interactions
written in ASN.1 (Abstract Syntax Notation) for computational efficiency

Interaction 13118 Mus musculus Full BIND Record **Launch Viewer:** Select Below

Molecule	Description	Molecular Function	Cellular Component	Biological Process	Experiment(s)	Links	
P53 • Trp53;TP53	Transformation related protein 53. Tumour suppressor protein with DNA binding and transcription factor function. Role in cell cycle; mutations involve [more...]	• DNA binding • transcription factor activity • protein binding	• nucleus • cytoplasm • cytosol	• protein-nucleus import\, translocation • transcription • regulation of transcription\, DNA-dependent • apoptosis • DNA damage response\, signal transduction by p53 class mediator • negative regulation of cell cycle	• Immunoprecipitation	NCBI SeqHound	[2 Pubmed Abstracts] [Other BIND data]
MDM2 • Mdm-2	Transformed Mmouse 3T3 cell Double Minute 2; nuclear phosphoprotein; LocusID:17246	• ubiquitin-protein ligase activity • protein binding • ATP binding • ligase activity	• nucleus	• start control point of mitotic cell cycle • cell growth and/or maintenance • protein ubiquitination • protein catabolism		NCBI SeqHound	

Bader GD, Betel D, Hogue CW. (2003)
BIND: the Biomolecular Interaction Network Database. Nucleic Acids Res. 31(1):248-50

Gene Ontology (GO) - an organizational framework for storing interaction and function data

http://www.geneontology.org

What is the *function* of a protein?

Not an easy question to answer!

There are many aspects to function, and different people might (and do!) describe the function of one protein many different ways.

In GO gene products (mostly proteins) are described using descriptors in three categories:

Molecular function - the activity carried out by the gene product at the molecular level: histidine kinase, alcohol dehydrogenase,

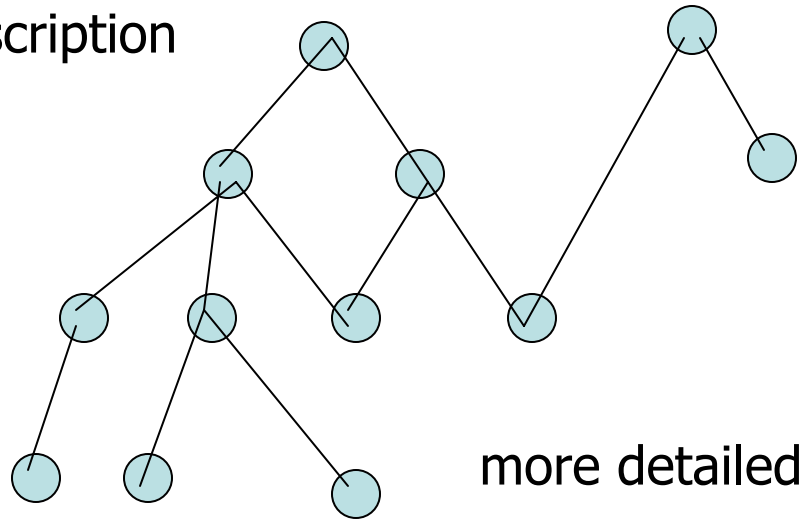
Biological process - a multi-step process, such as cell division, DNA replication, signal transduction

Cellular component - part of a cell that is a part of a larger structure; ribosome, spindle pole, kinetochore

The hierarchical structure of GO

Molecular function, biological process and the cellular compartment where a gene product is found can be described at many levels of detail.
GO uses hierarchical description where each protein gets a set of terms at different levels. The ontology structure is that of an acyclic directed graph.

less detailed description



more detailed description

MAP kinase activity

Accession:GO:0004707

Synonyms:

MAPK

mitogen activated kinase

Definition:

Catalysis of the phosphorylation of proteins. Mitogen-activated protein kinase; a family of protein kinases that perform a crucial step in relaying signals from the plasma membrane to the nucleus. They are activated by a wide range of proliferation- or differentiation-inducing signals; activation is strong with agonists such as polypeptide growth factors and tumor-promoting phorbol esters, but weak (in most cell backgrounds) by stress stimuli.

Term Lineage Graph view.

GO:0003673 : Gene_Ontology (146200)
- Ⓟ GO:0003674 : molecular_function (97507)
  - ① GO:0003824 : catalytic activity (32256)
    - ① GO:0016301 : kinase activity (5076)
      - ① GO:0004672 : protein kinase activity (3030)
        - ① GO:0004674 : protein serine/threonine kinase activity (1854)
          - ① GO:0004702 : receptor signaling protein serine/threonine kinase activity (257)
            - ① **GO:0004707 : MAP kinase activity (90)**
    - ① GO:0016740 : transferase activity (9890)
      - ① GO:0016772 : transferase activity, transferring phosphorus-containing groups (5162)
        - ① GO:0016773 : phosphotransferase activity, alcohol group as acceptor (3682)
          - ① GO:0004672 : protein kinase activity (3030)
            - ① GO:0004674 : protein serine/threonine kinase activity (1854)
              - ① GO:0004702 : receptor signaling protein serine/threonine kinase activity (257)
                - ① **GO:0004707 : MAP kinase activity (90)**
  - ① GO:0004871 : signal transducer activity (6386)
    - ① GO:0005057 : receptor signaling protein activity (598)
      - ① GO:0004702 : receptor signaling protein serine/threonine kinase activity (257)
        - ① **GO:0004707 : MAP kinase activity (90)**

p38 (fly)

regulation of innate immune response

Accession:GO:0045088
Synonyms: None.
Definition:

Any process that modulates the frequency, rate or extent of the innate immune response, the organism's first line of defense against infection.

Term Lineage Graph view.

GO:0003673 : Gene_Ontology (146200)
- Ⓟ GO:0008150 : biological_process (96312)
  - ① GO:0007582 : physiological process (60310)
    - ① GO:0050874 : organismal physiological process (4807)
      - ① GO:0006955 : immune response (1944)
        - ① GO:0045087 : innate immune response (520)
          - Ⓟ **GO:0045088 : regulation of innate immune response (37)**
        - Ⓟ GO:0050776 : regulation of immune response (103)
          - ① **GO:0045088 : regulation of innate immune response (37)**
      - ① GO:0050896 : response to stimulus (8829)
        - ① GO:0009605 : response to external stimulus (6919)
          - ① GO:0009607 : response to biotic stimulus (3692)
            - ① GO:0006952 : defense response (2995)
              - ① GO:0006955 : immune response (1944)
                - ① GO:0045087 : innate immune response (520)
                  - Ⓟ **GO:0045088 : regulation of innate immune response (37)**
                - Ⓟ GO:0050776 : regulation of immune response (103)
                  - ① **GO:0045088 : regulation of innate immune response (37)**

p38 (fly)
CELLULAR COMPONENT

nucleus

Accession:GO:0005634
Synonyms: None.
Definition:

A membrane-bounded organelle of eukaryotic cells that contains the chromosomes. It is the primary site of DNA replication and RNA synthesis in the cell.

Term Lineage

GO:0003673 : Gene_Ontology (146200)
Ⓟ GO:0005575 : cellular_component (79199)
① GO:0005623 : cell (56534)
Ⓟ GO:0005622 : intracellular (46101)
Ⓟ **GO:0005634 : nucleus (11723)**

Where do the GO terms come from?

A team of experts is responsible for assigning GO annotation.
Every term comes with an evidence code describing where the annotation came from.

Sample evidence codes:
IDA - inferred from direct assay (enzyme assay, cell fractionation)
IPI - inferred from physical interaction (2-hybrid)
IGI - inferred from genetic interaction (suppressor, synthetic lethal)
IEP - inferred from expression pattern (microarray)
IMP - inferred from mutant phenotype
ISS - inferred from sequence or structure similarlity
TAS - traceable author statement
NAS - non-traceable author statement

Advantages of GO

Controlled vocabulary! Everyone can communicate using the same terms!

Designed to apply across species.

Linked to genomic databases like TIGR, FlyBase, SGD (yeast), WormBase, MGD (mouse), RGD (rat), TAIR (Arabidopsis), ZFIN (zebrafish) and more…

Makes it possible to *compute* on protein function and localization. Can define formal relationships between proteins based on their GO annotation.

Flexible. GO is always growing and changing. New terms can be added to the hierarchy.

Free and open for the use of the community.

Computational methods for improving the quality of interaction data.

1. Assessment and validation (improve accuracy)

2. Prediction (improve coverage)

Assessing and filtering interaction data

1. Promiscuity criteria

In most high-throughput interaction studies, a few proteins are observed to interact promiscuously. Generally these are removed from the analysis. Problem: some interactions may be real!

Examples:

Affinity purification/ms

Even with no bait, 17 proteins were found in pull-downs by Gavin et al. 49 other proteins found to have a similar frequency of interaction to these false positives were thrown out.

Yeast 2-hybrid

Proteins observed to make many interactions in many screens usually discarded as probably false positives

Assessing and filtering interaction data

2. Overlap criteria

   A. with other interaction data - intersection is low!

   In 2001, ~2,000 high-throughput measurements were confirmed by small scale experiments.

   B. with non-interaction data, e.g. annotations in YPD = yeast protein databank

      YPD now proprietary at Incyte :-(

Please see figures 1 and 2 of

Deane, Charlotte M., Lukasz Salwinski, Ioannis Xenarios, and David Eisenberg. "Protein Interactions: Two Methods for Assessment of the Reliability of High Throughput Observations." *Mol. Cell. Proteomics* 1 (May 2002): 349-356.

Overlap with expression data: Expression Profile Reliability (EPR)

Please see figure 4 of

Deane, Charlotte M., Lukasz Salwinski, Ioannis Xenarios, and David Eisenberg. "Protein Interactions: Two Methods for Assessment of the Reliability of High Throughput Observations." *Mol. Cell. Proteomics* 1 (May 2002): 349-356.

Note: proteins involved in "true" protein-protein interactions have more similar mRNA expression profiles than random pairs. Use this to assess how good an experimental set of interactions is.

Assessing and filtering interaction data
Expression Profile Reliability (EPR)

Please see figure 4 of

Deane, Charlotte M., Lukasz Salwinski, Ioannis Xenarios, and David Eisenberg. "Protein Interactions: Two Methods for Assessment of the Reliability of High Throughput Observations." *Mol. Cell. Proteomics 1* (May 2002): 349-356.

Assume the observed distribution observed results from the true interactions and false positive interactions. The observed distribution is expressed as a weighted sum of these contributions.

Estimate the distribution for non interactions using all protein-protein pairs (assume interactions are rare).
Estimate distribution for true interactions using small-scale experiments.
Fit a parameter a_{EPR} to estimate how many high-throughput interactions are true positive vs. false positive.

$$F\_exp(d^2) = a_{EPR} \bullet F\_int(d^2) + (1-a_{EPR}) \bullet F\_no\_int(d^2)$$

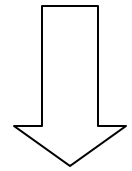
Best fit a_{EPR} = 31% -> ~70% of high-throughput pairs are false positives!
But method doesn't tell you which interactions these are.
Other methods have estimated that ~50% of yeast 2-hybrid pairs are true positives.

Assessing and filtering interaction data

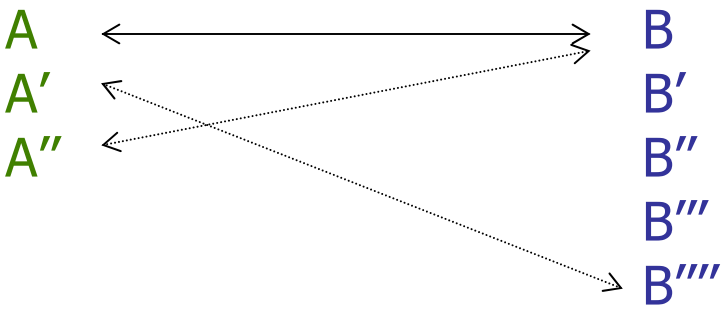
Homology methods - Paralogous Verification (PVM)

Sequence A *candidate interaction* Sequence B

PSI-BLAST
w/in genome



list of paralogs



PVM score = 2
(# non A-B interactions)

Deane et al. Mol. & Cell. Proteomics (2002) 1.5, 349-356

PVM is very specific, but not very sensitive

three different high-confidence interaction datasets
WP indicates proetin paris with ≥ 1 paralog for A or B

Please see figure 5 of

Deane, Charlotte M., Lukasz Salwinski, Ioannis Xenarios, and David Eisenberg. "Protein Interactions: Two Methods for Assessment of the Reliability of High Throughput Observations." *Mol. Cell. Proteomics 1* (May 2002): 349-356.

Points on this plot come from using different PVM score cutoffs to designate a true interaction. It is an example of a receiver-operator characteristic (ROC) curve, which is commonly used to illustrate the tradeoff between sensitivity vs. specificity.

PVM is very *selective* - if a pair scores by PVM it is almost certainly a true positive (x-axis -> low false positive rate).

However, PVM does not achieve good coverage - it is not *sensitive* (y-axis). At most, PVM can confirm ~50% of high-confidence examples. This is at least partly because many examples of paralogous complexes are sparse.

Assessing and filtering interaction data

DIP_CORE is a set of 3,003 interactions considered higher confidence.

DIP_CORE interactions either:

1. Have been observed in a small-scale experiment (2,246)
2. Have been observed in more than one experiment (1,179)
3. Have been confirmed by PVM (1,428)

DIP 40078E			
DIP 369N	**PIR** DNMS53	**SwissProt** P53_MOUSE	**GenBank** gi:2144761
	Name/Description cellular tumor antigen p53		
DIP 24169N	**PIR**	**SwissProt** Q64364	**GenBank** gi:6753390
	Name/Description p19ARF tumor suppressor protein		

Evidence Help

Type	Method	Details	Source
E	Immunoprecipitation	---	PMID:9653180
V	SMSC(1)	---	---

verification field indicates that one (1) small-scale experiment supports this interaction

Deane et al. Mol. & Cell. Proteomics (2002) 1.5, 349-356

Assessing and filtering interaction data

3. Topology criteria use information about the observed vs. expected interaction network.

We will discuss the paper:

Bader et al. "Gaining confidence in high-throughput protein interaction networks"

Nature Biotechnology (2004) 22, 78-85

Predicting protein-protein interactions

1. Sequence methods

How can you predict that an interaction might occur between two proteins based purely on sequence data?

Review: Valencia & Pazos, Current Opinion in Structural Biology (2002) 12, 368-373

Predicting protein-protein interactions

1. **Sequence methods**

   phylogenetic profiles - based on the joint presence/absence of a pair of proteins in a large number of genomes; recall the first literature discussion class (Pellegrini et al.).

   co-evolution - as assessed by similarity of phylogenetic trees. "mirrortree" method compares the distance matrices for generating trees; requires lots of sequences and a good alignment!

   gene fusions - genes encoding interacting proteins in one organism are sometimes fused into a single gene in another. Look for these occurrences.

   gene neighborhood - for bacteria, the arrangement of genes in operons means that interacting proteins are often encoded in adjacent sites in the genome

Review: Valencia & Pazos, Current Opinion in Structural Biology (2002) 12, 368-373

Predicting protein-protein interactions

1. Sequence methods

correlated mutations - the idea is that interacting positions on different proteins should co-evolve so as to maintain the interface. Look for correlation between sequence changes at one position and those at another position in a multiple sequence alignment.

Recall Süel et al. "Evolutionarily conserved networks of residues mediate allosteric communication in proteins"

$\Delta\Delta G_{i,j} = \Delta G_j - \Delta G_{j|i}$ where $\Delta G = kT \bullet \ln(P(x \text{ at } j)/P_{MSA}(x))$

Pazos & Valencia "In silico two-hybrid systems for the selection of physically interacting protein pairs." PROTEINS (2002) 47, 219-227

Pearson coefficient $r_{ij} = \Sigma(S_{i,k,l} - \langle S_i \rangle)(S_{j,k,l} - \langle S_j \rangle)/\text{normalization}$ describes the correlation between amino acid positions i and j in two proteins. Here $S_{i,k,l}$ is a measure of the similarity of the aa at position i in sequences k and l, and $\langle S_i \rangle$ is the average of these values. k and l are sequences taken from a MSA that has the same number of sequences, from the same species, for sites i and j.

Problem: need lots of sequences, and the method is very sensitive to the alignment used.

Review: Valencia & Pazos, Current Opinion in Structural Biology (2002) 12, 368-373

Predicting protein-protein interactions

2. Structure-based methods

Docking is a large field in and of itself, which involves predicting how two known structures will interact. It even has its own prediction contest - CAPRI, like CASP.

The main issues in docking are, as always when modeling structure, (1) sampling the conformational space and (2) selecting the correct solution

Docking approaches require structures of both interacting components.

Frequently, conformational changes accompany protein interactions. Docking methods generally require a structure of the *bound* conformation to predict interactions correctly. Modeling conformational flexibility is hard.

We don't have enough structures or good enough docking methods to make high-throughput prediction of protein-protein interactions practical at this point.

Predicting protein-protein interactions

2. Structure-based methods

What do you do when you don't have a structure?

Homology modeling methods (Aloy & Russell PNAS (2002) 99, 5896-5901)

For target proteins that have homologs that form a complex of known structure:

(1) Identify pairs of positions that form interactions in the known structure
(2) align the target proteins to the template proteins and score the interacting residue pairs identified in step (1) with a knowledge-based potential.
(3) Normalize using the scores for pairs of random sequences
(4) Z-scores above a certain cutoff indicate that a complex is likely.

~65% accuracy when assessing whether different fibroblast growth factors bind to various receptors (4 structures available, 252 possible pairings evaluated). Not practical to apply at the genome level due to lack of homologous complexes with structures.

Predicting protein-protein interactions

2. Structure-based methods

What do you do when you don't have a structure?

Threading methods (Lu et al., MULTIPROSCPECTOR)

Phase I: Thread each target sequence onto a library of folds using a permissive cutoff

Phase II: Take pairs of fold assignments and thread the targets onto complexes of these folds (complexes of known structure)

Evaluate an interfacial score to determine how complementary the fit is

S_interface = -log (N_obs_in_PDB(i,j)/N_expect_by_chance(i,j))

Used library of 768 complexes, predicted 7,321 interactions for yeast proteins.

Hard to assess performance. One way is to look at some property that you believe should correlate with interactions, e.g. co-localization or function.

Lu et al., PROTEINS (2002) 49, 350-364, Genome Research (2003) 13, 1146-1154

2. Structure-based methods

Threading methods Lu, L, H Lu, and J Skolnick. "MULTIPROSPECTOR: An Algorithm for The Prediction of Protein-protein Interactions by Multimeric Threading." *Proteins* 49, no. 3 (15 November 2002): 350-64.

Co-localization:
Are the proteins found
in the same part of the cell?

Please see figure 2 of

Lu, Long, Adrian K. Arakaki, Hui Lu, and Jeffrey Skolnick. "Multimeric Threading-Based Prediction of Protein–Protein Interactions on a Genomic Scale: Application to the Saccharomyces Cerevisiae Proteome." *Genome Res.*13 (June 2003): 1146-1154.

Predicting protein-protein interactions

3. Methods based on data

    Jansen et al. - next class

Next class: literature about protein-protein interaction assessment and prediction

Bader et al. "Gaining confidence in high-throughput protein interaction networks" *Nature Biotechnology* (2004) *22*, 78-85

Jansen et al. "A Bayesian networks approach for predicting protein-protein interactions from genomic data" *Science* (2003) *320*, 449-453.

Focus on:

1. What are they trying to do?
2. What do they use as a set of positive and negative examples?
3. What is their basis for deciding if an interaction is good or not?
4. How well do the methods work? How can you tell?
5. Do they learn anything new or exciting about interactions in the proteome?