# 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

mRNA

$\Sigma = \{A,C,G,U\}$

↓

protein

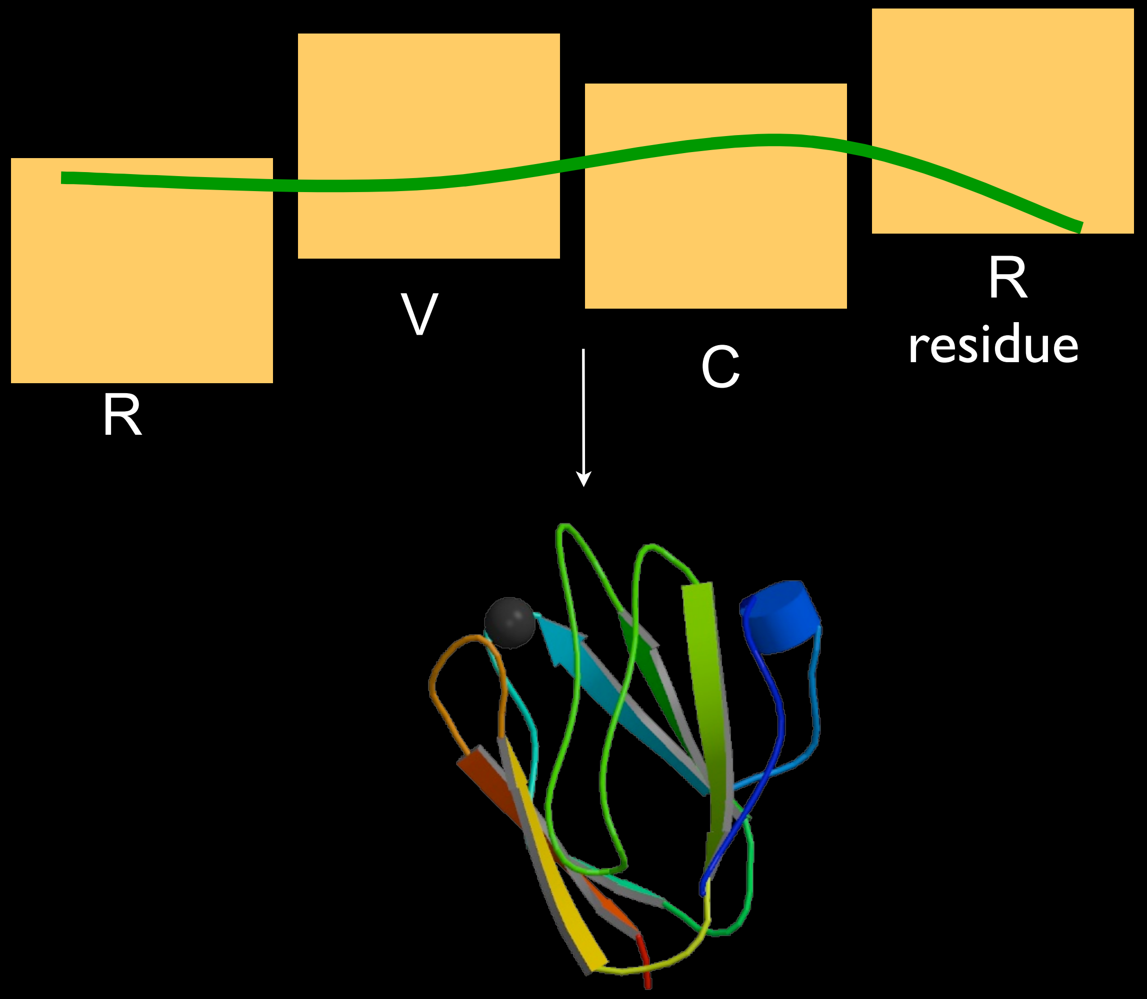
$|\Sigma| = 20$ amino acids

Amino acids with flexible side chains strung together on a backbone

**Proteins are the Building Blocks of Life**

Their shape is instrumental in determining their function.

| AGG | GUC | UGU | CGA |
|---|---|---|---|
| ↓ | ↓ | ↓ | ↓ |
| R | V | C | R |



- 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

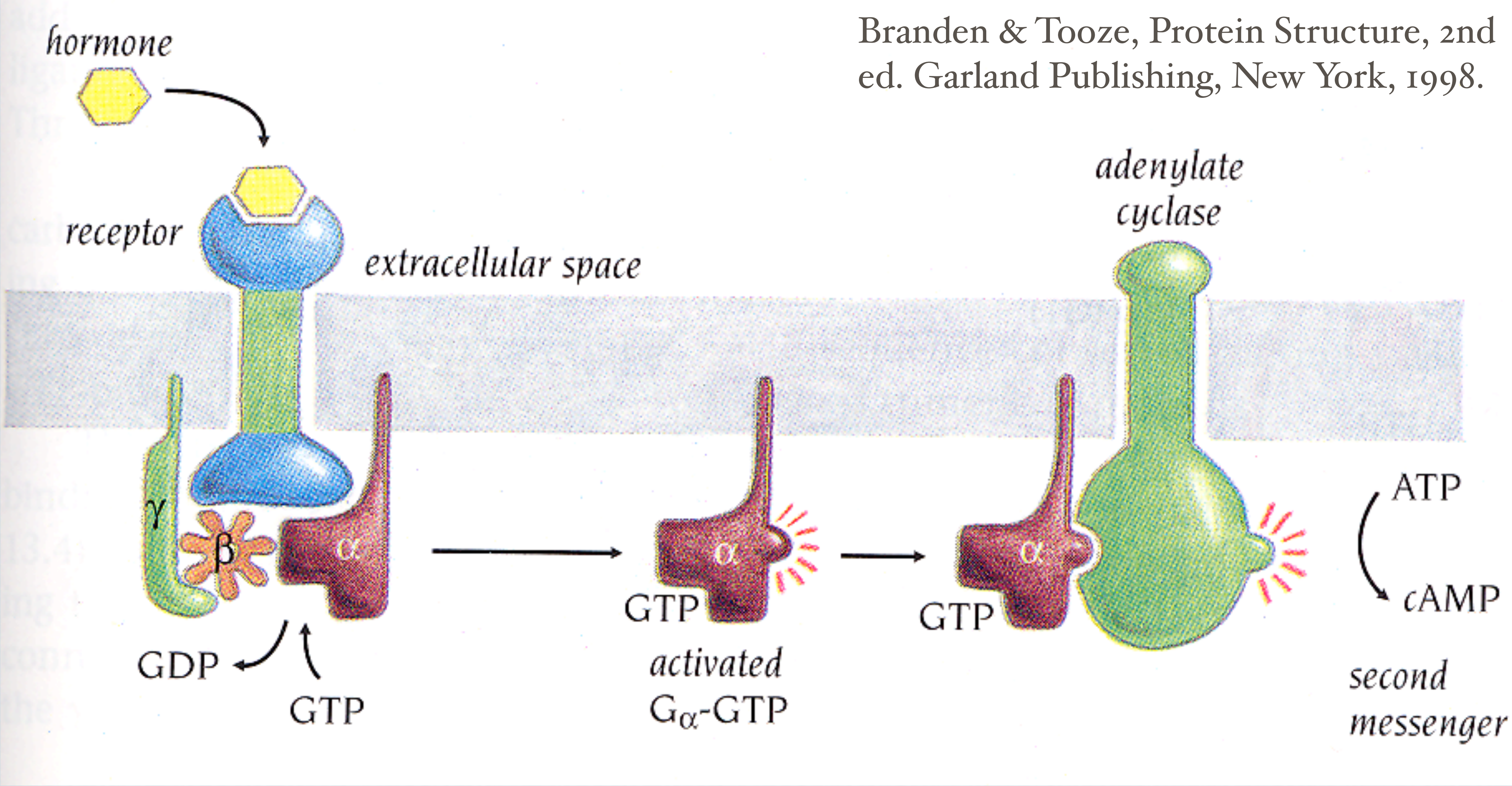
- Networks:

| Network | Nodes | Edges |
|---|---|---|
| Transcription (aka regulatory) | proteins/genes | A “regulates” B |
| Metabolic | Metabolites / small molecules | Reactions |
| **Protein-Protein** | **Proteins** | **Physical Interactions** |

# Experimental Techniques

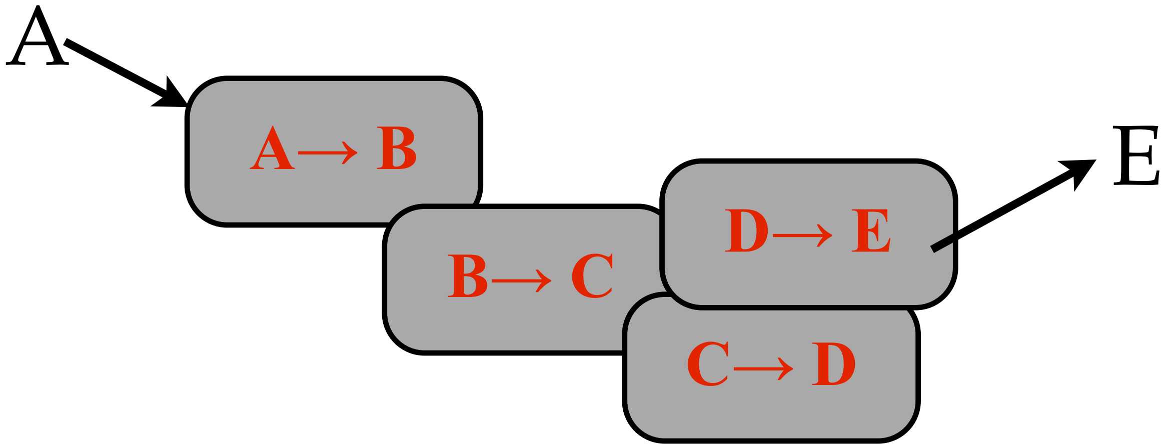
CMSC 858L

# Proteins Interact

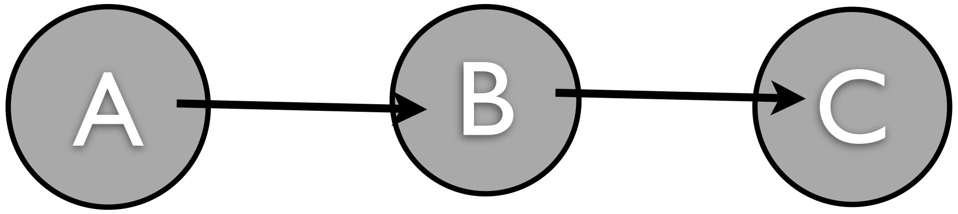


Branden & Tooze, Protein Structure, 2nd ed. Garland Publishing, New York, 1998.

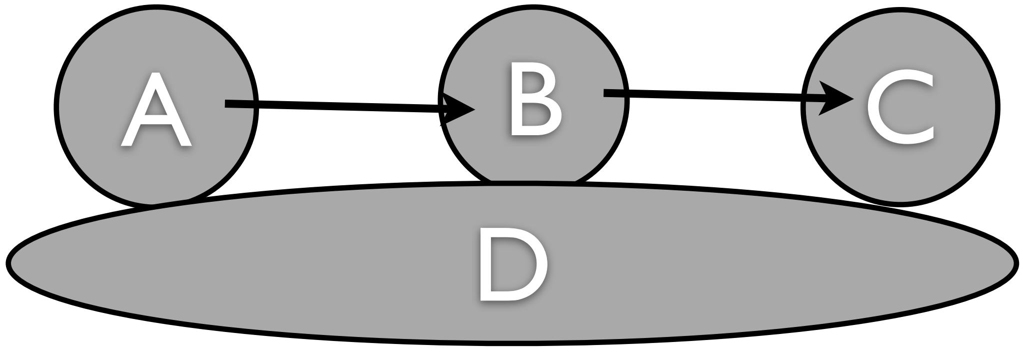
# “Why” proteins interact:



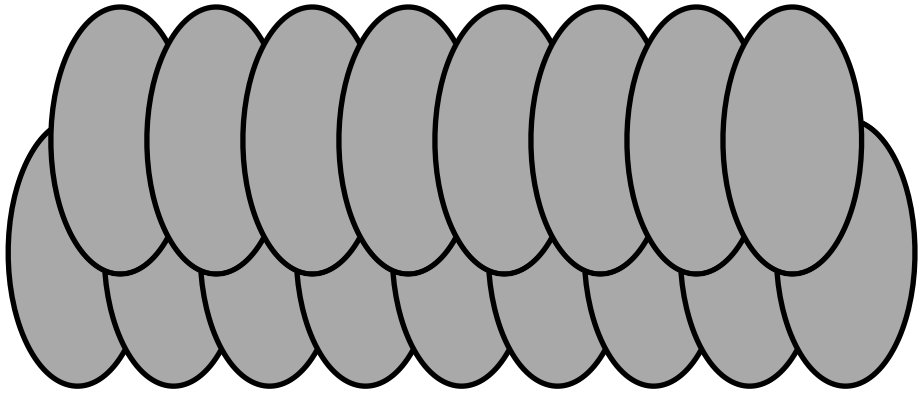
Bring chains of enzymes together



Signal Transduction



“Tethered” Signal Transduction



Form structures

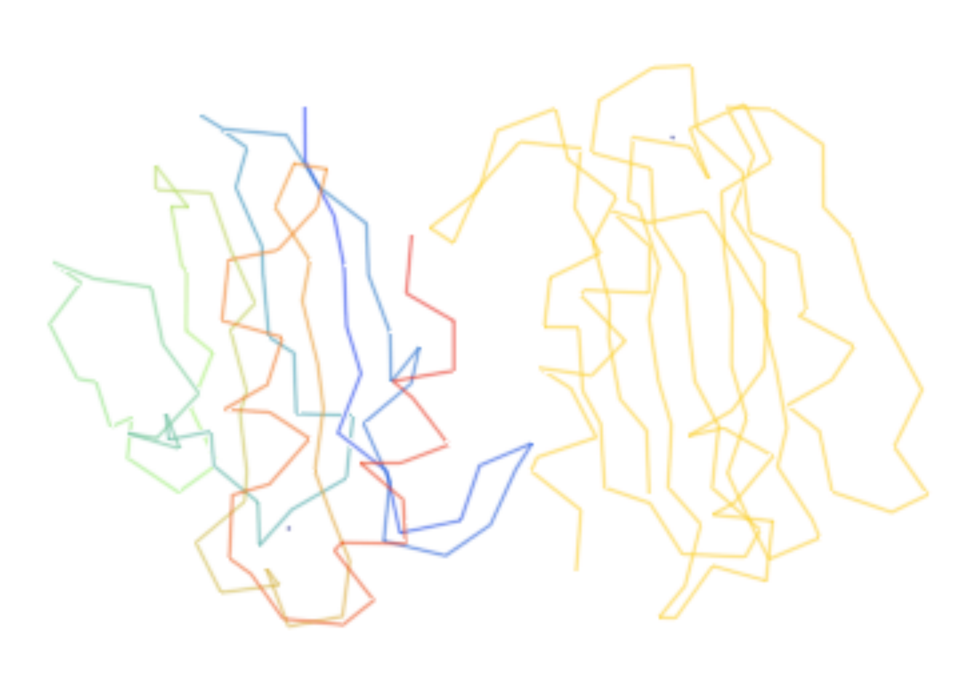
From “Analysis of Biological Networks” Junker and Schreiber, eds

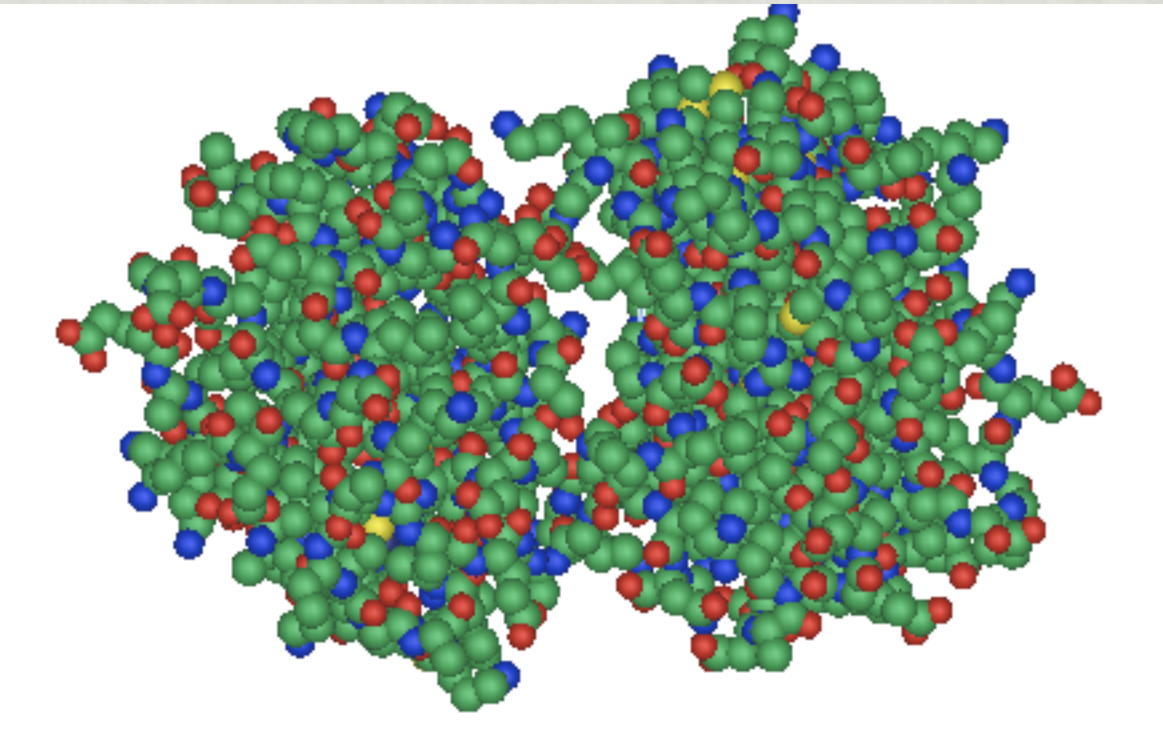
# Experimental Techniques to Determine Protein Interactions

- Slow, accurate, costly:
  - X-ray crystallography
  - NMR
- High throughput, but noisy:
  - Yeast Two-Hybrid
  - TAP-MS (tandem affinity purification / mass spec)

# Determining protein structure:

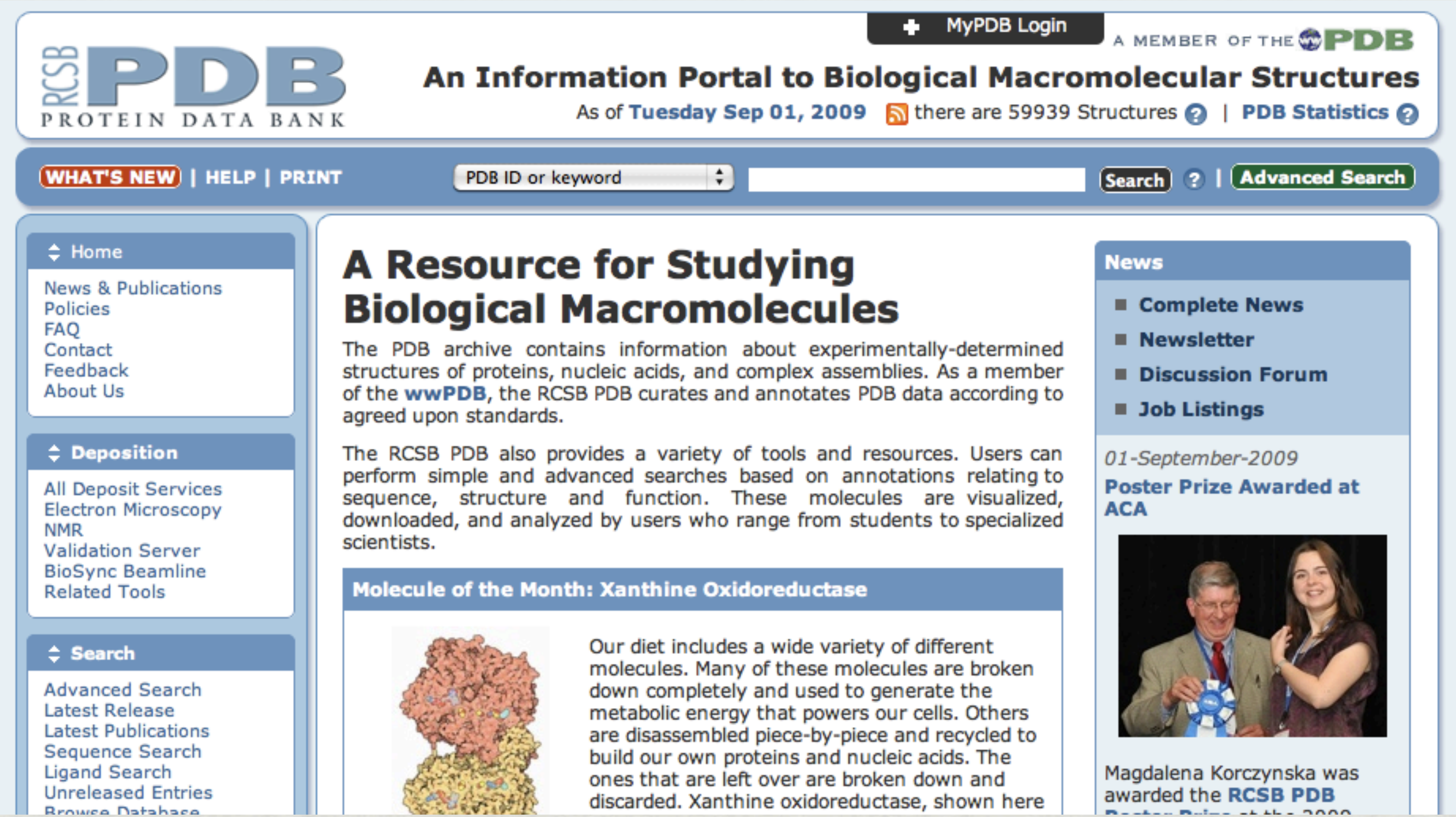
- X-ray crystallography
- NMR
  - If you **can** determine structure of a complex, you know the position of each of its atoms.
  - Slow, costly techniques.
  - Don’t always work.
- More recent: high-throughput techniques





# PDB

http://www.rcsb.org/pdb/home/home.do



An Information Portal to Biological Macromolecular Structures

As of Tuesday Sep 01, 2009 there are 59939 Structures | PDB Statistics

## 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.

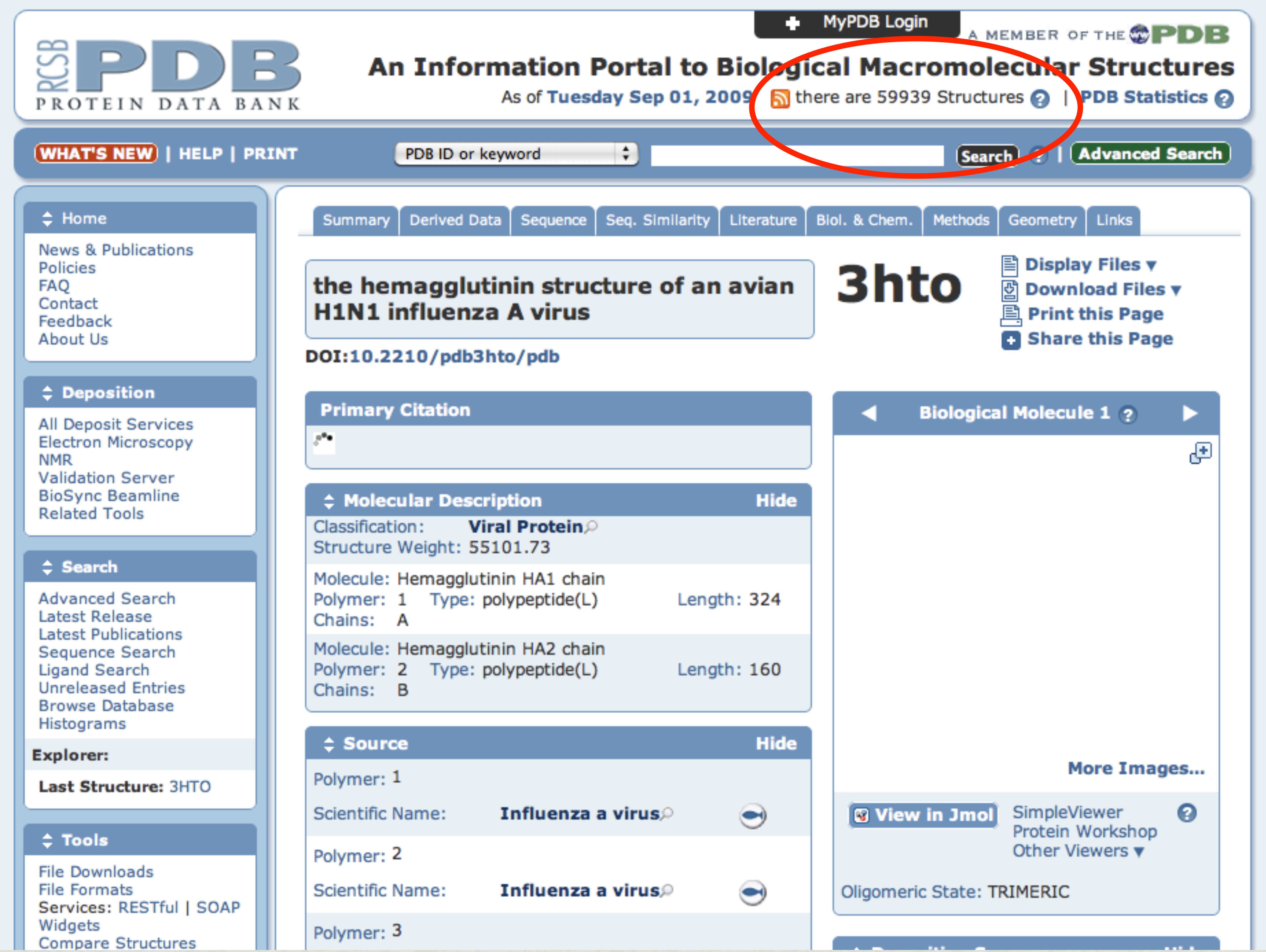
**Molecule of the Month: Xanthine Oxidoreductase**

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

*01-September-2009*

**Poster Prize Awarded at ACA**

Magdalena Korczynska was awarded the **RCSB PDB**



MyPDB Login

A MEMBER OF THE PDB

RCSB PDB PROTEIN DATA BANK

An Information Portal to Biological Macromolecular Structures

As of Tuesday Sep 01, 2009 there are 59939 Structures ? | PDB Statistics ?

WHAT'S NEW | HELP | PRINT

PDB ID or keyword

Search ? | Advanced Search

Home

News & Publications
Policies
FAQ
Contact
Feedback
About Us

Deposition

All Deposit Services
Electron Microscopy
NMR
Validation Server
BioSync Beamline
Related Tools

Search

Advanced Search
Latest Release
Latest Publications
Sequence Search
Ligand Search
Unreleased Entries
Browse Database
Histograms

**Explorer:**

**Last Structure:** 3HTO

Tools

File Downloads
File Formats
Services: RESTful | SOAP
Widgets
Compare Structures

Summary | Derived Data | Sequence | Seq. Similarity | Literature | Biol. & Chem. | Methods | Geometry | Links

**the hemagglutinin structure of an avian H1N1 influenza A virus**

**3hto**

Display Files ▾
Download Files ▾
Print this Page
Share this Page

DOI:10.2210/pdb3hto/pdb

Primary Citation

Molecular Description Hide

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 Hide

Polymer: 1
Scientific Name: **Influenza a virus**

Polymer: 2
Scientific Name: **Influenza a virus**

Polymer: 3

Biological Molecule 1 ?

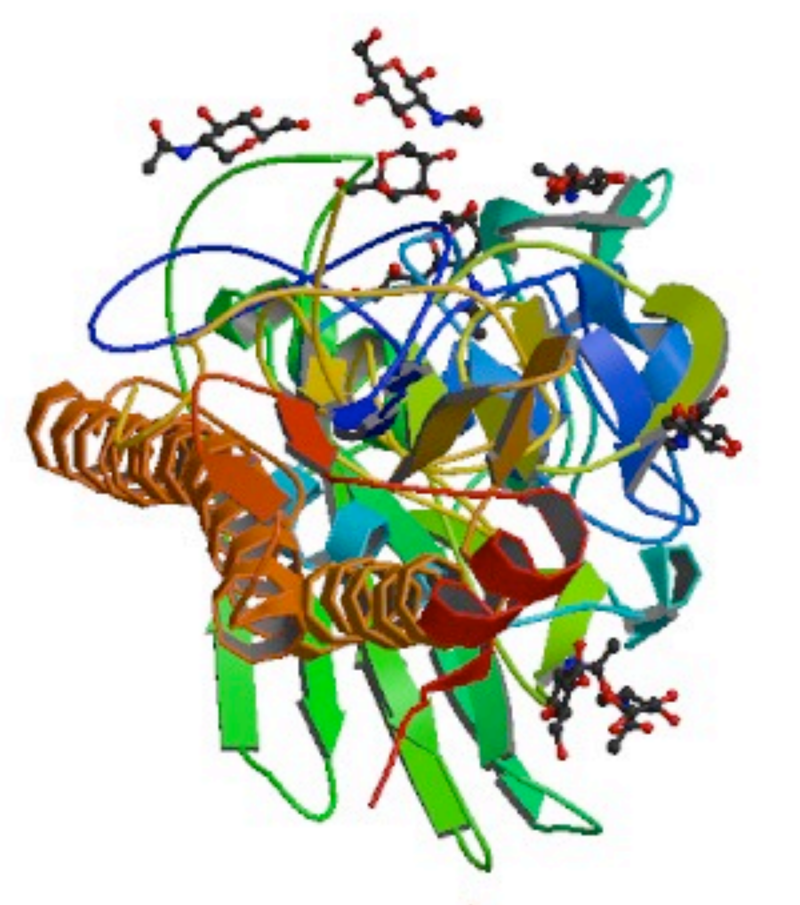
More Images...

View in Jmol
SimpleViewer
Protein Workshop
Other Viewers ▾

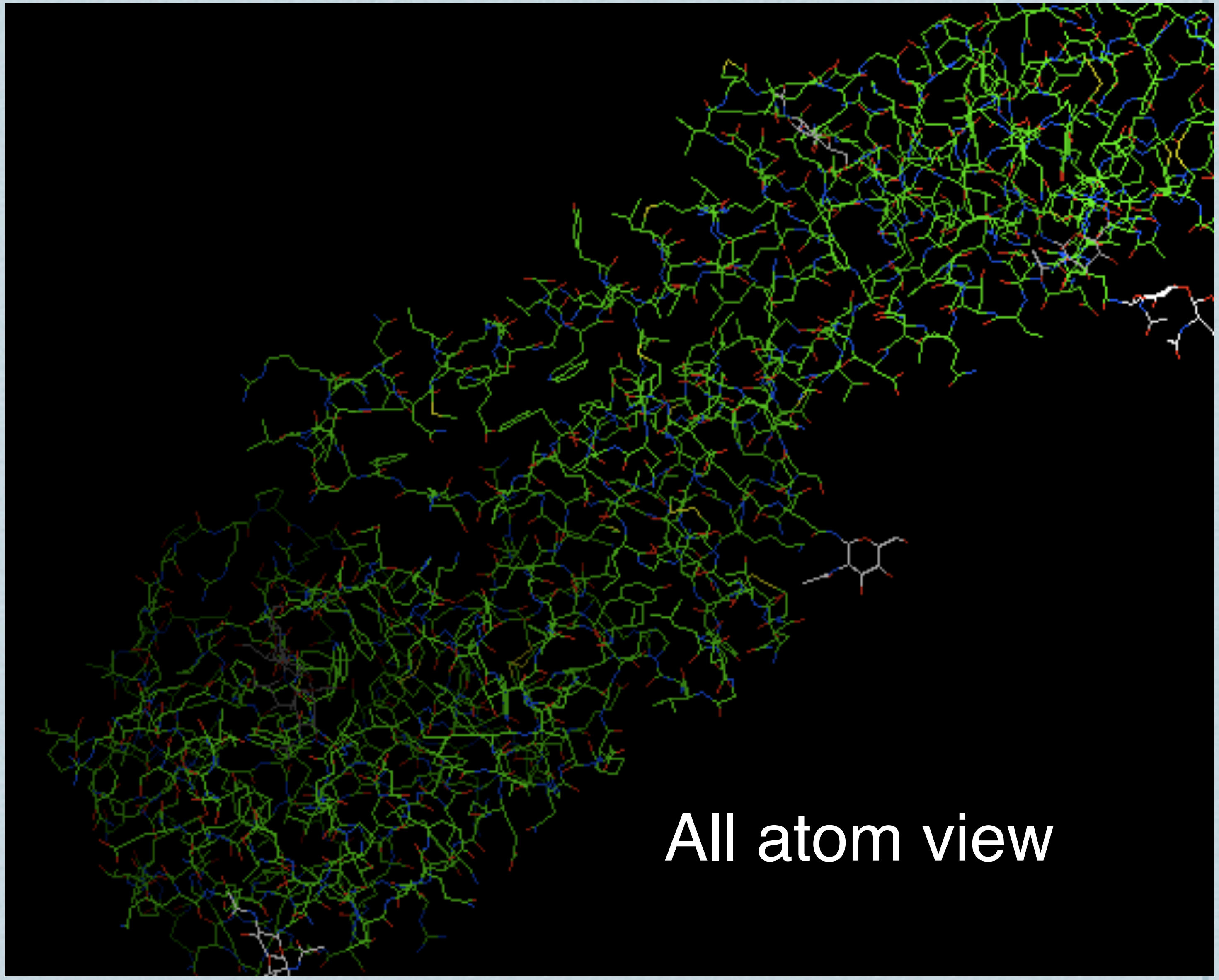
Oligomeric State: TRIMERIC

# Other View of a Protein

AVIAN H5 HAEMAGGLUTININ

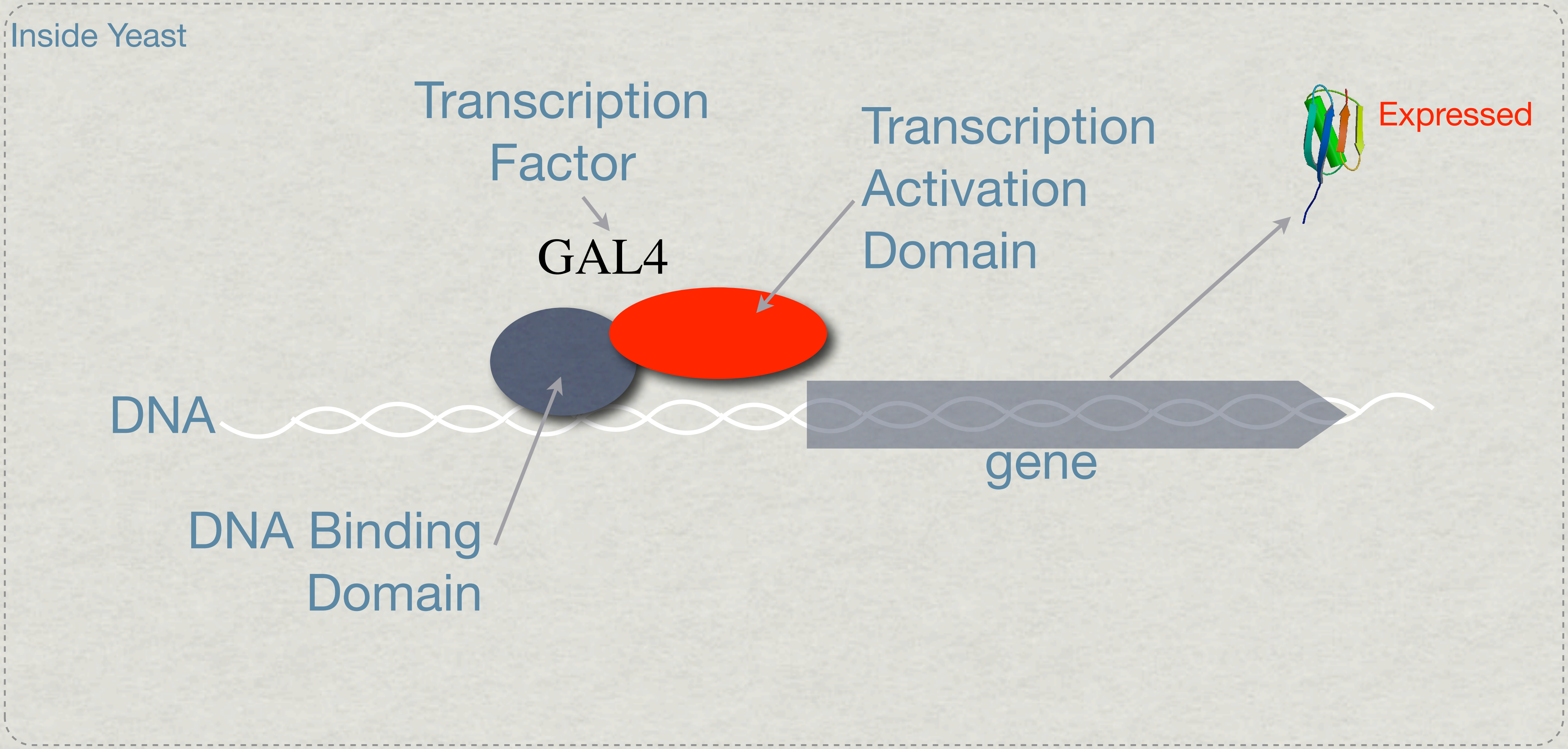


“Cartoon” drawing, showing major features such as alpha helices and beta sheets



All atom view

# Yeast Two-Hybrid



Inside Yeast

Transcription Factor

GAL4

Transcription Activation Domain

Expressed

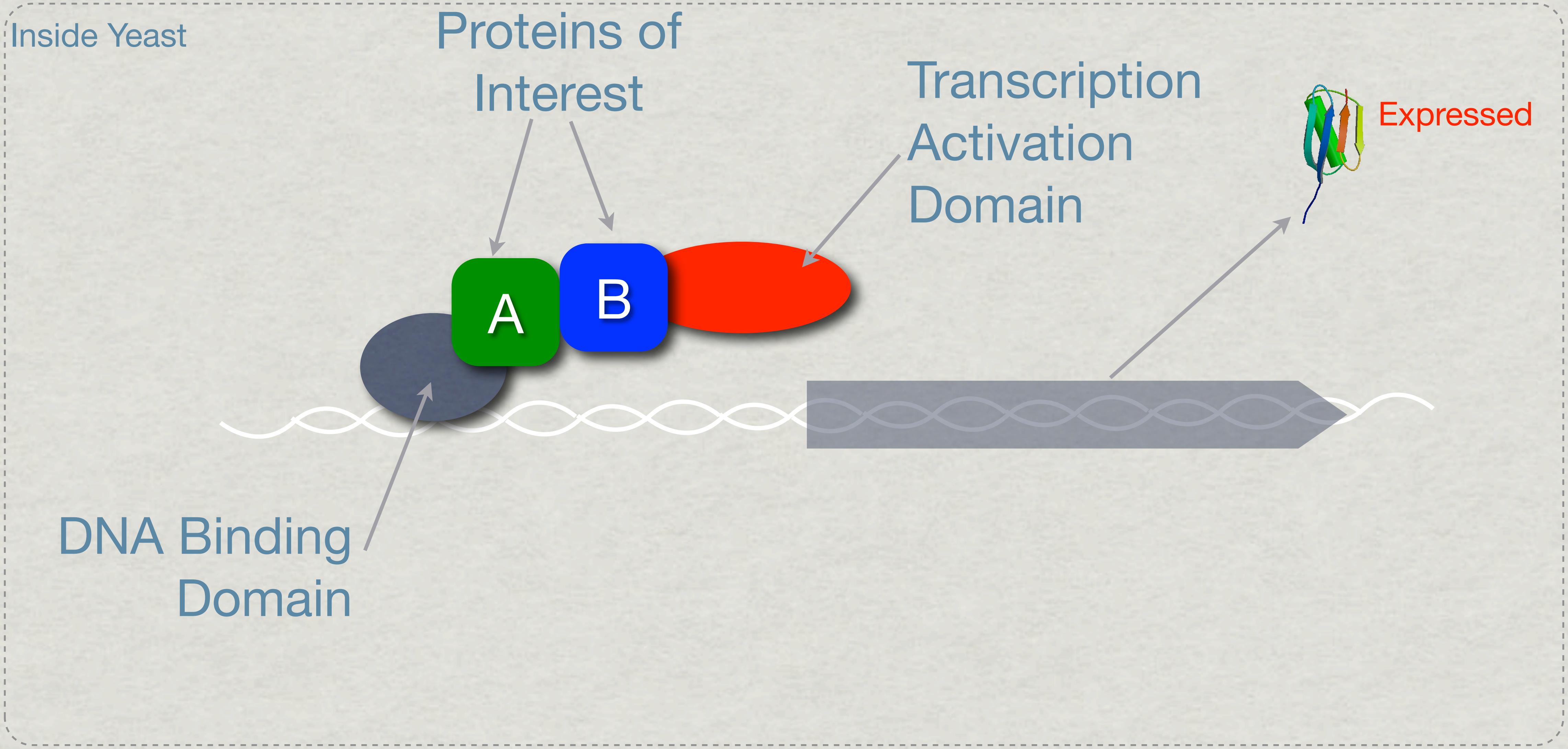
DNA

gene

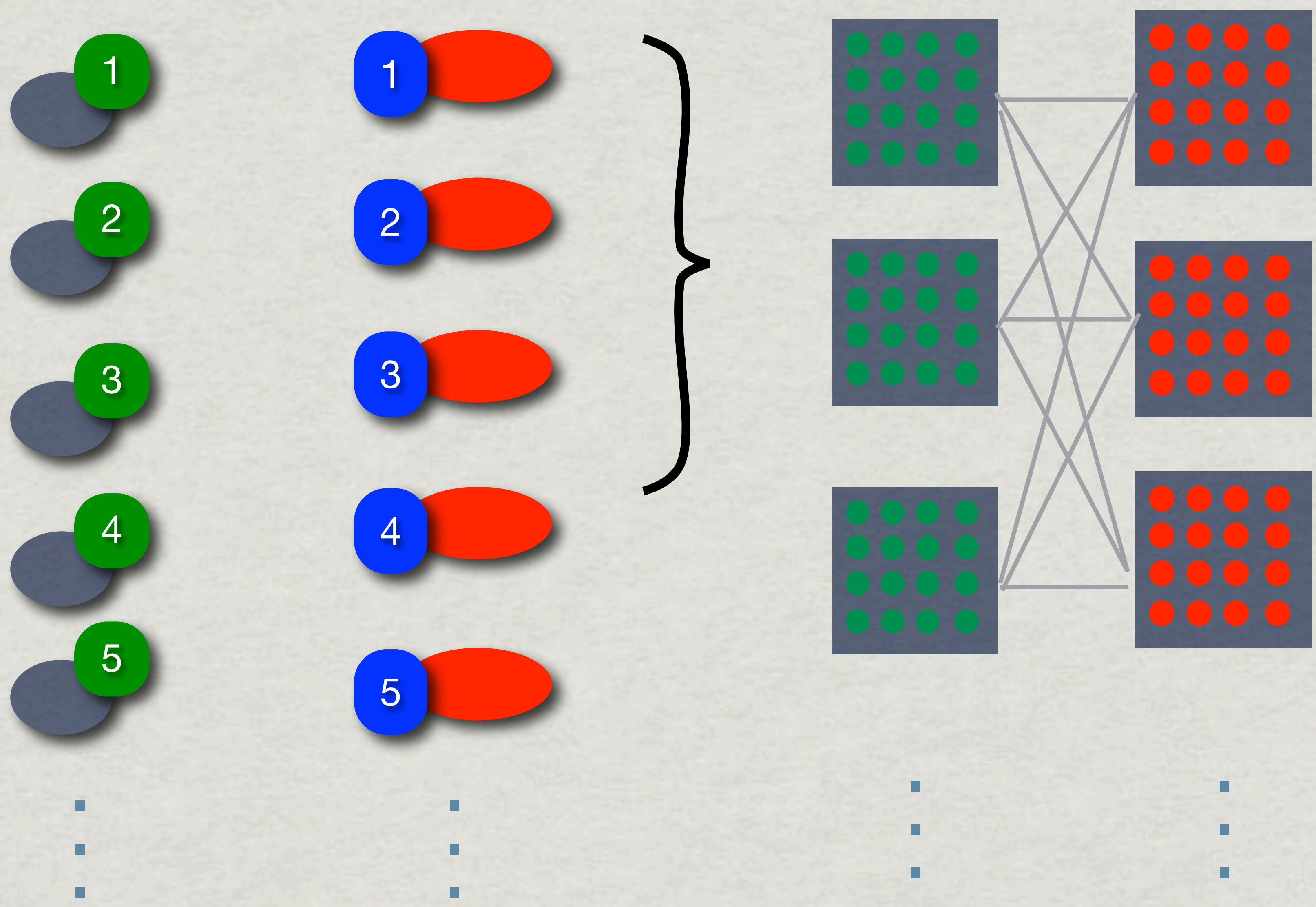
DNA Binding Domain

“Domain” = functional, evolutionary conserved unit of a protein

# Yeast Two-Hybrid



# Scaling Up (Ito et al, 2001)

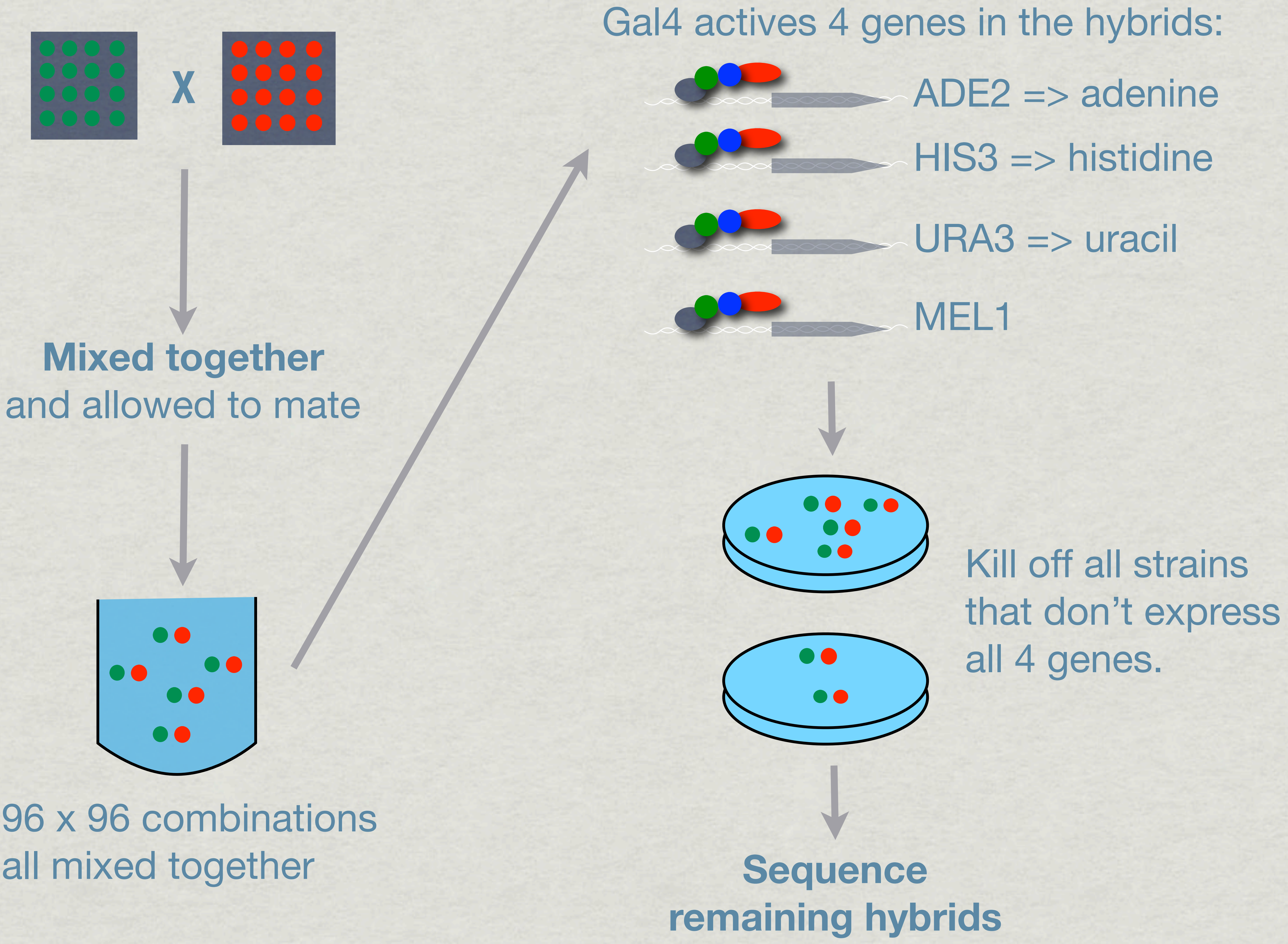


**BAIT**

**PREY**

**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 actives 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:

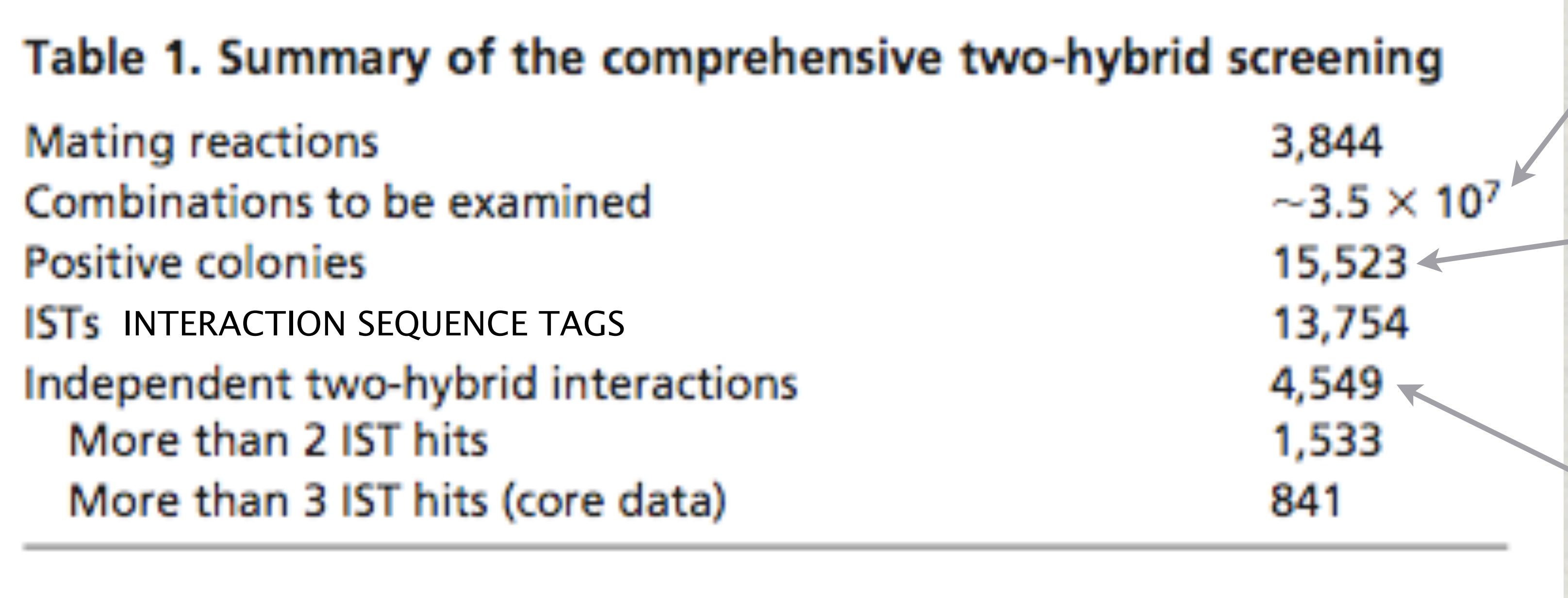


**Table 1. Summary of the comprehensive two-hybrid screening**

| | |
|---|---|
| Mating reactions | 3,844 |
| Combinations to be examined | $\sim 3.5 \times 10^7$ |
| Positive colonies | 15,523 |
| ISTs INTERACTION SEQUENCE TAGS | 13,754 |
| Independent two-hybrid interactions | 4,549 |
| More than 2 IST hits | 1,533 |
| More than 3 IST hits (core data) | 841 |

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

# of colonies that passed all 4 tests

Involve 3278 proteins out of ~ 6,000

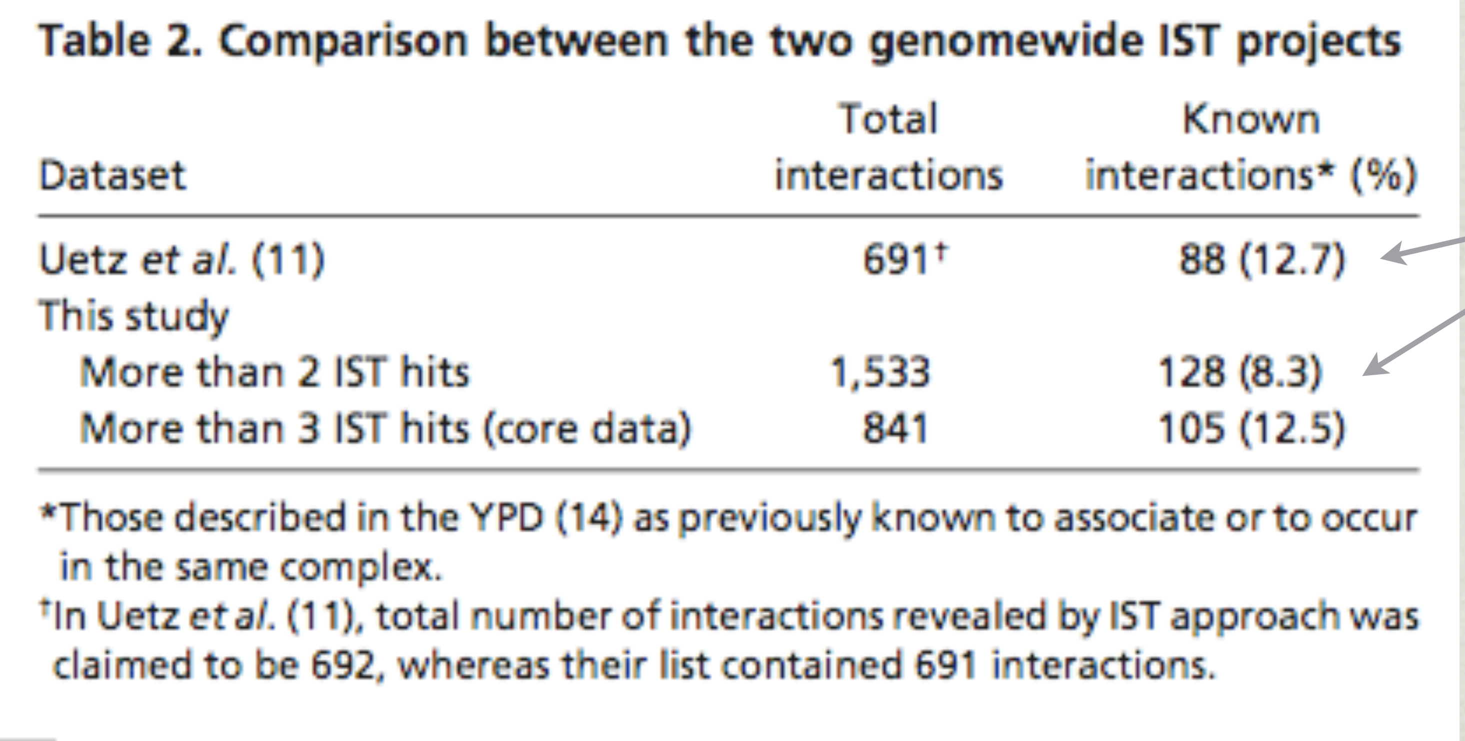


**Table 2. Comparison between the two genomewide IST projects**

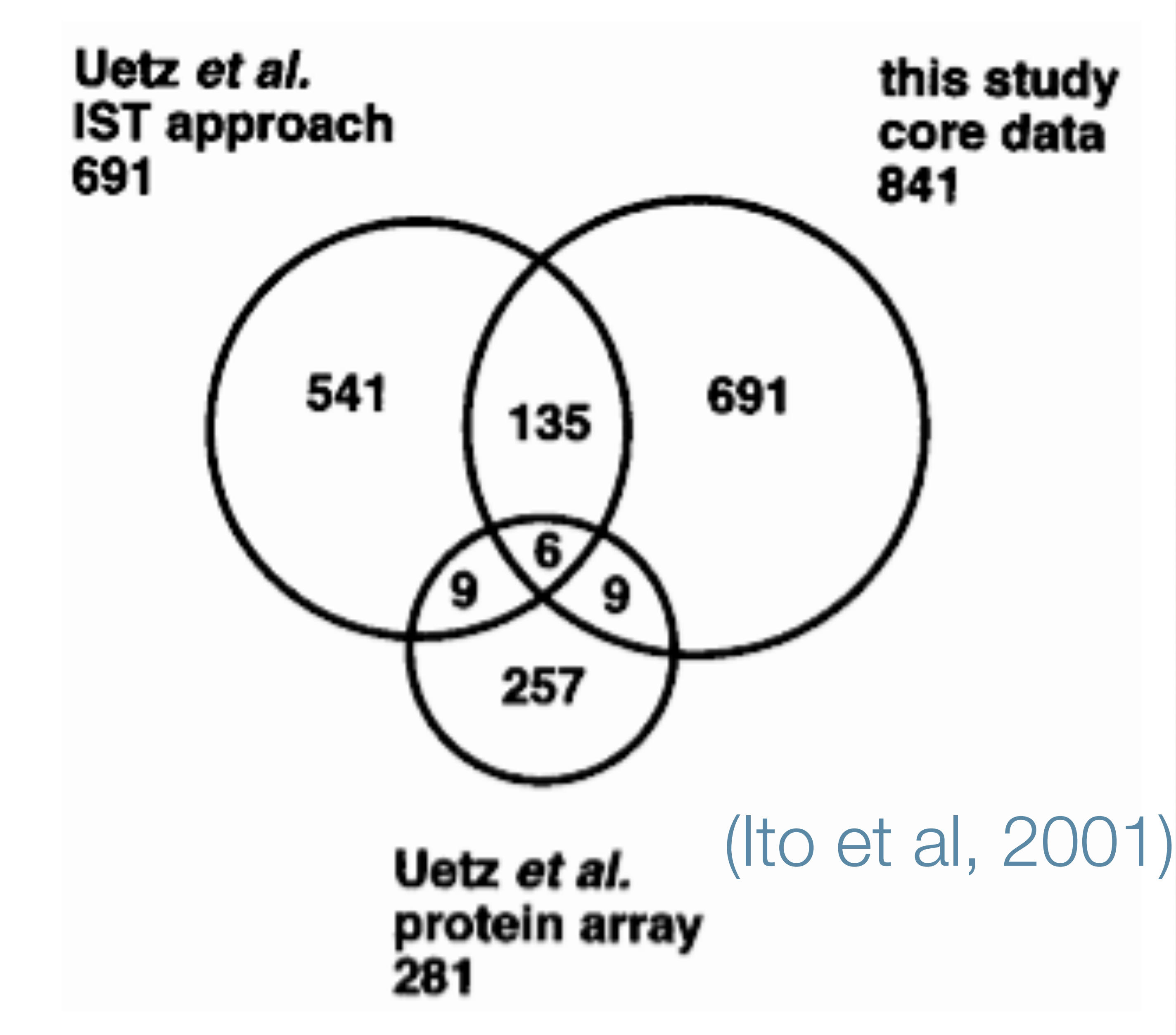
| Dataset | Total interactions | Known interactions[*] (%) |
|---|---|---|
| Uetz *et al.* (11) | 691[†] | 88 (12.7) |
| This study | | |
| More than 2 IST hits | 1,533 | 128 (8.3) |
| More than 3 IST hits (core data) | 841 | 105 (12.5) |

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

[†] In Uetz *et al.* (11), total number of interactions revealed by IST approach was claimed to be 692, whereas their list contained 691 interactions.

Comparable overlaps between known interactions

# What could go wrong with Yeast 2-Hybrid?



(Ito et al, 2001)

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)

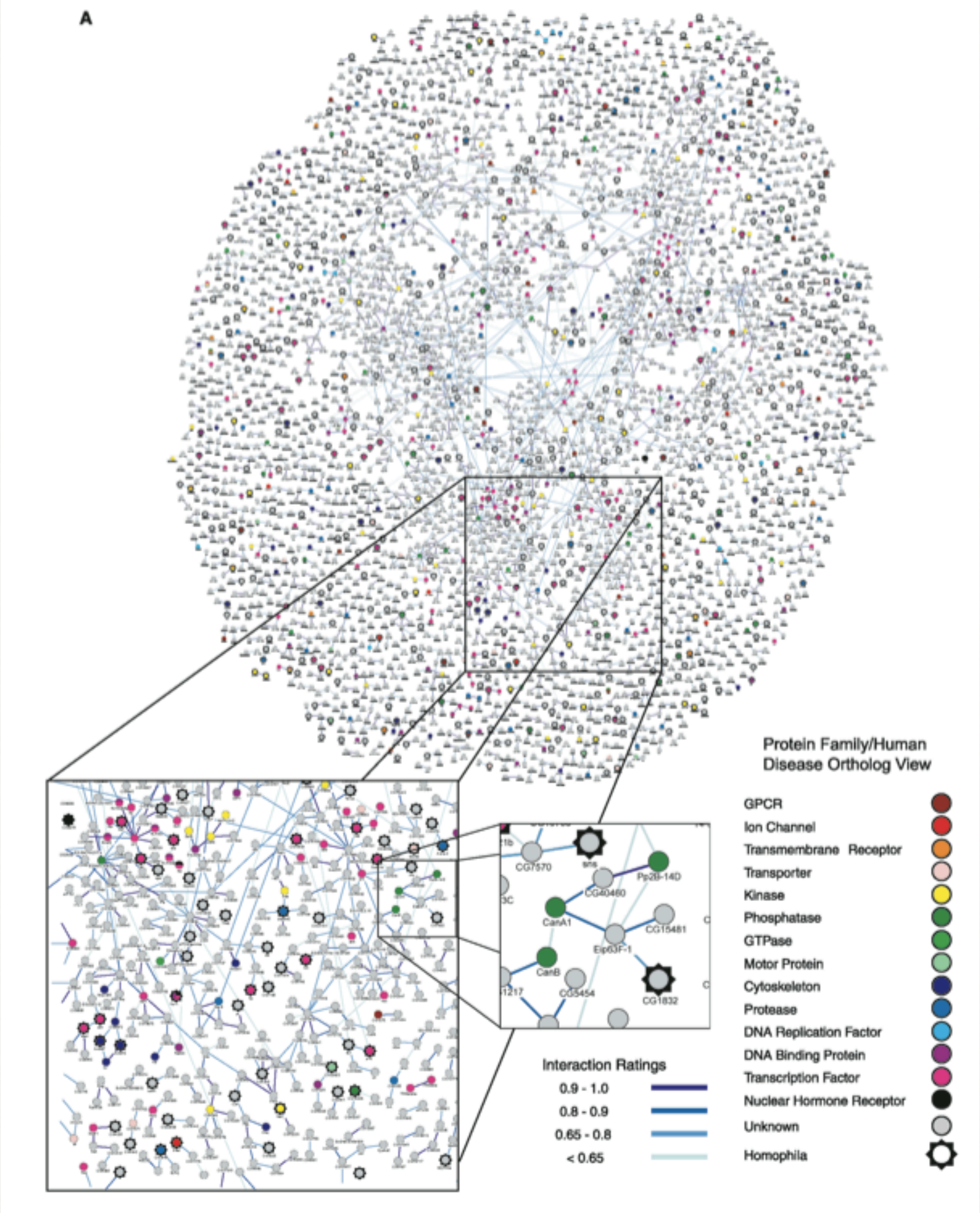
- hydrophobic / membrane proteins may not fold correctly.

# Fruit Fly (*Drosophila melanogaster*)

# Fly (*Drosophila melanogaster*) (Giot et al, 2003)

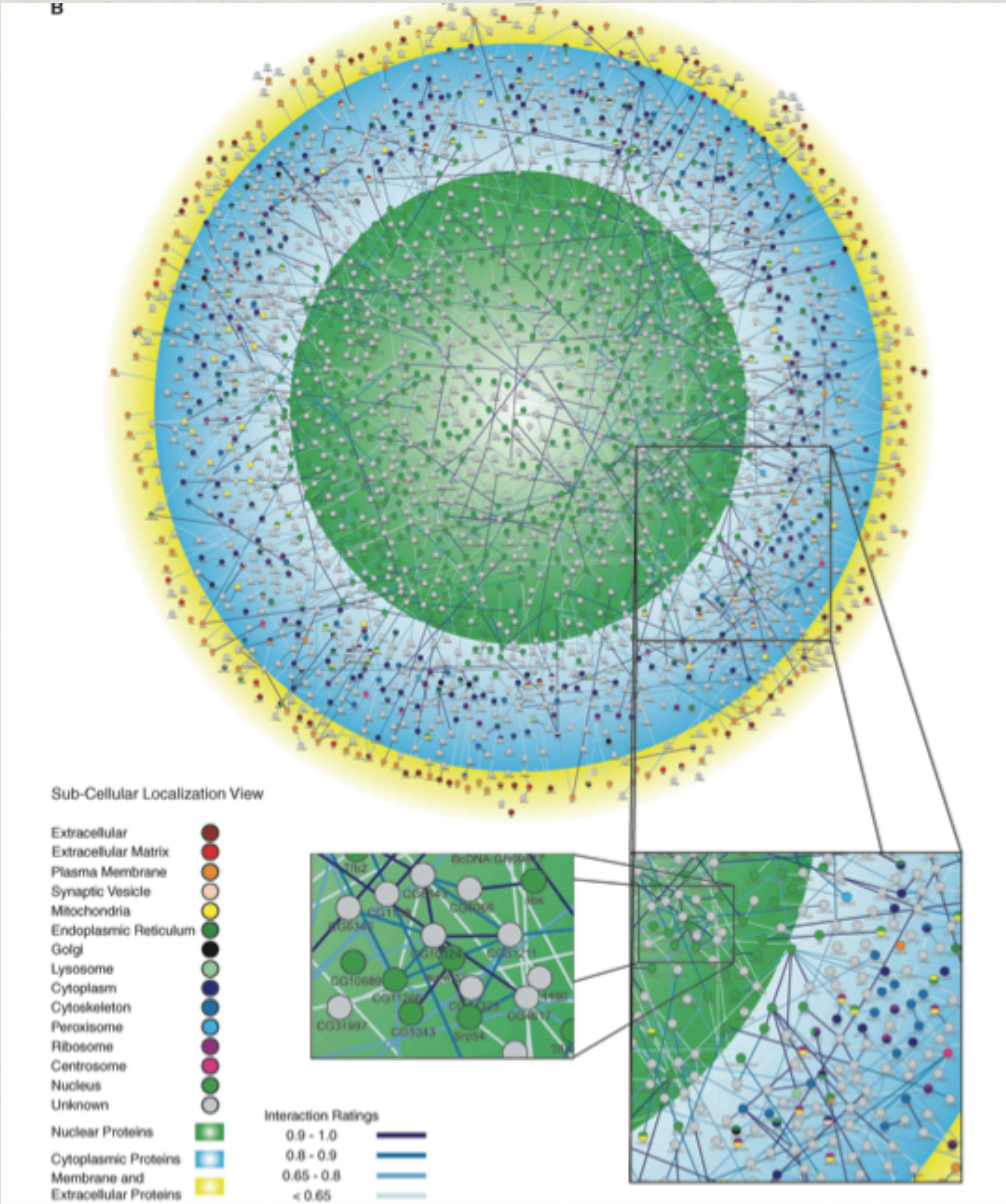
7,048 proteins
20,405 interactions

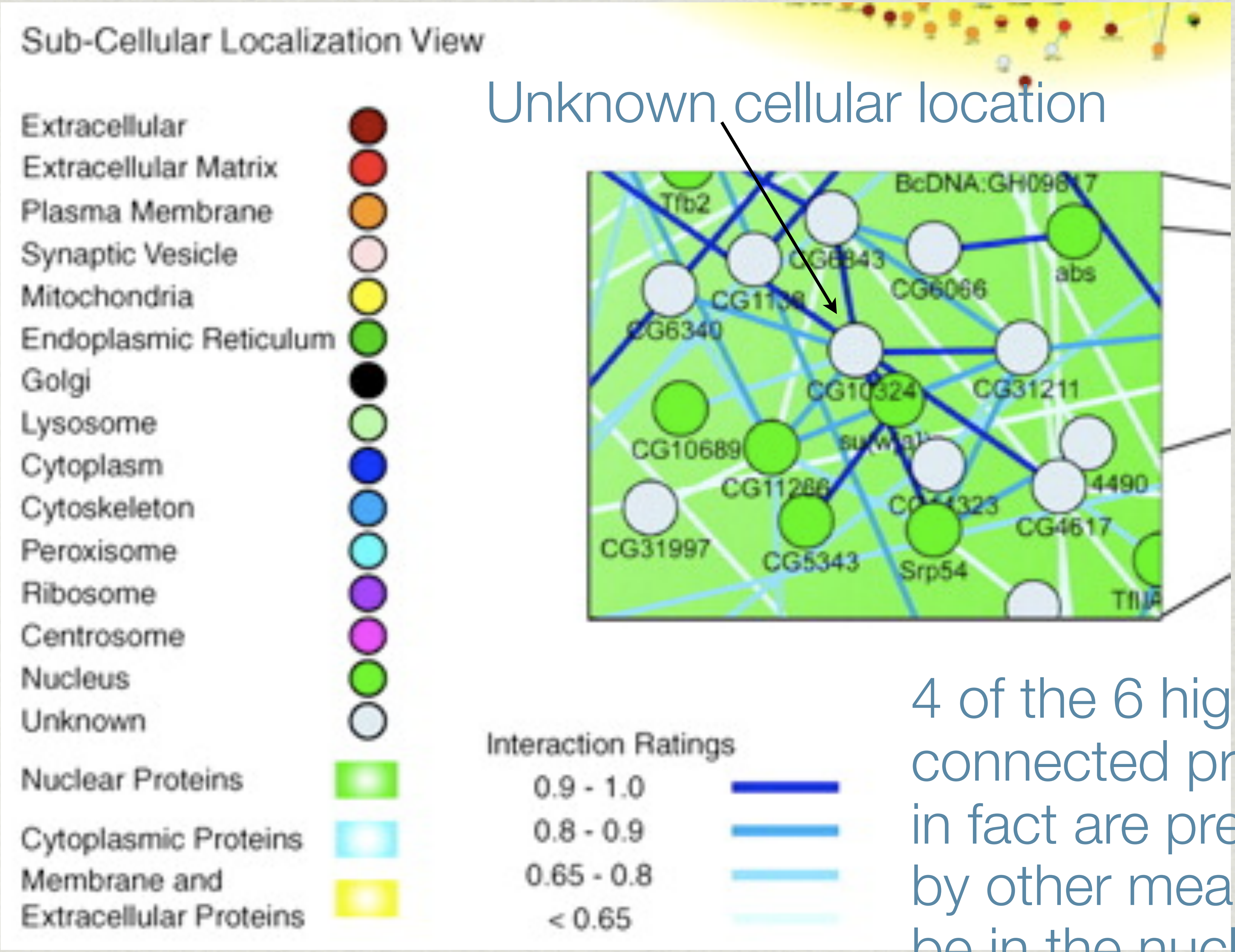
High-confidence:
4,679 proteins
4,780 interactions



# Colored and placed by sub-cellular location

(Giot et al, 2003)

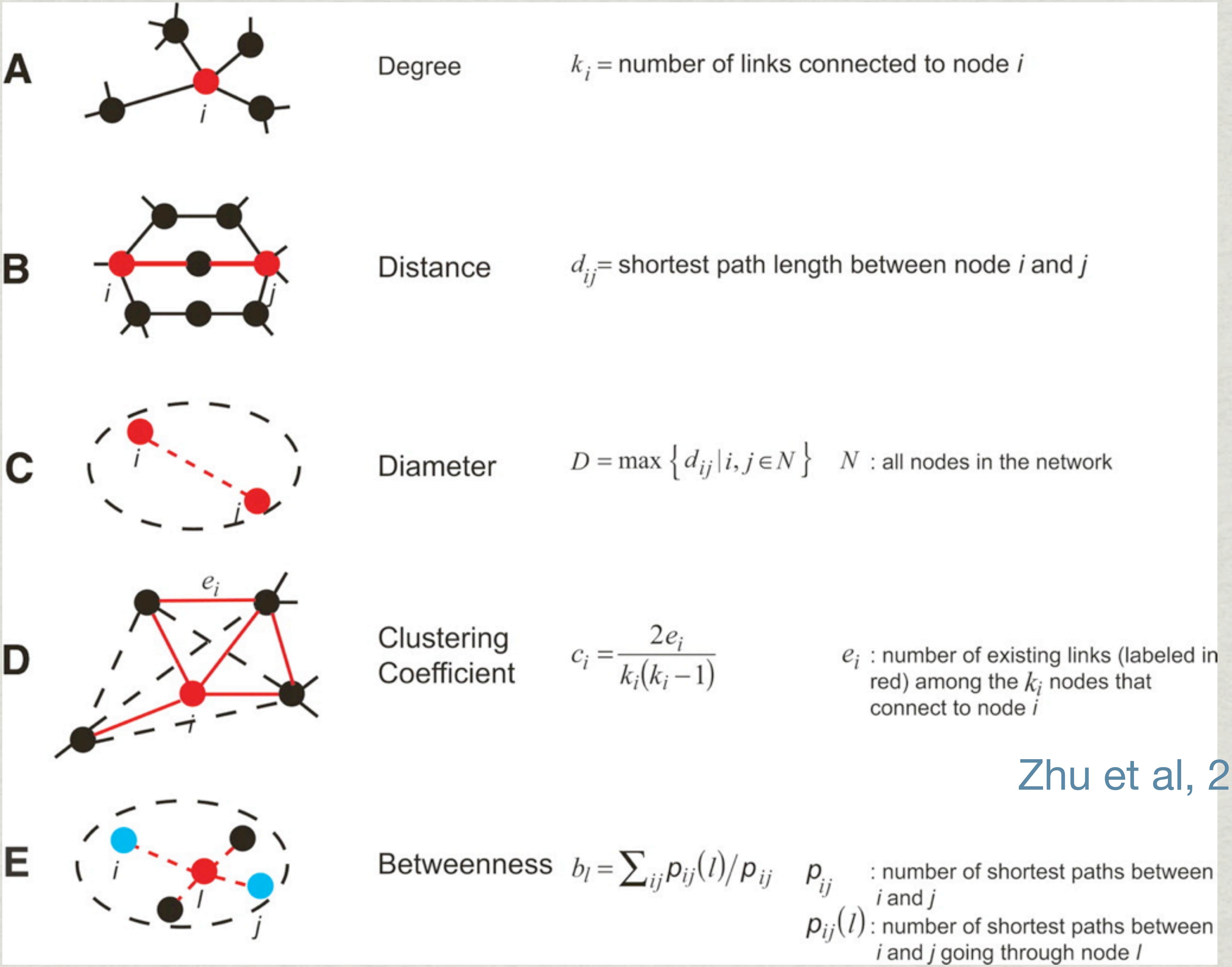




Unknown cellular location

4 of the 6 highly connected proteins in fact are predicted by other means to be in the nucleus.

# General Topological Properties



Degree $k_i$ = number of links connected to node *i*

Distance $d_{ij}$= shortest path length between node *i* and *j*

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

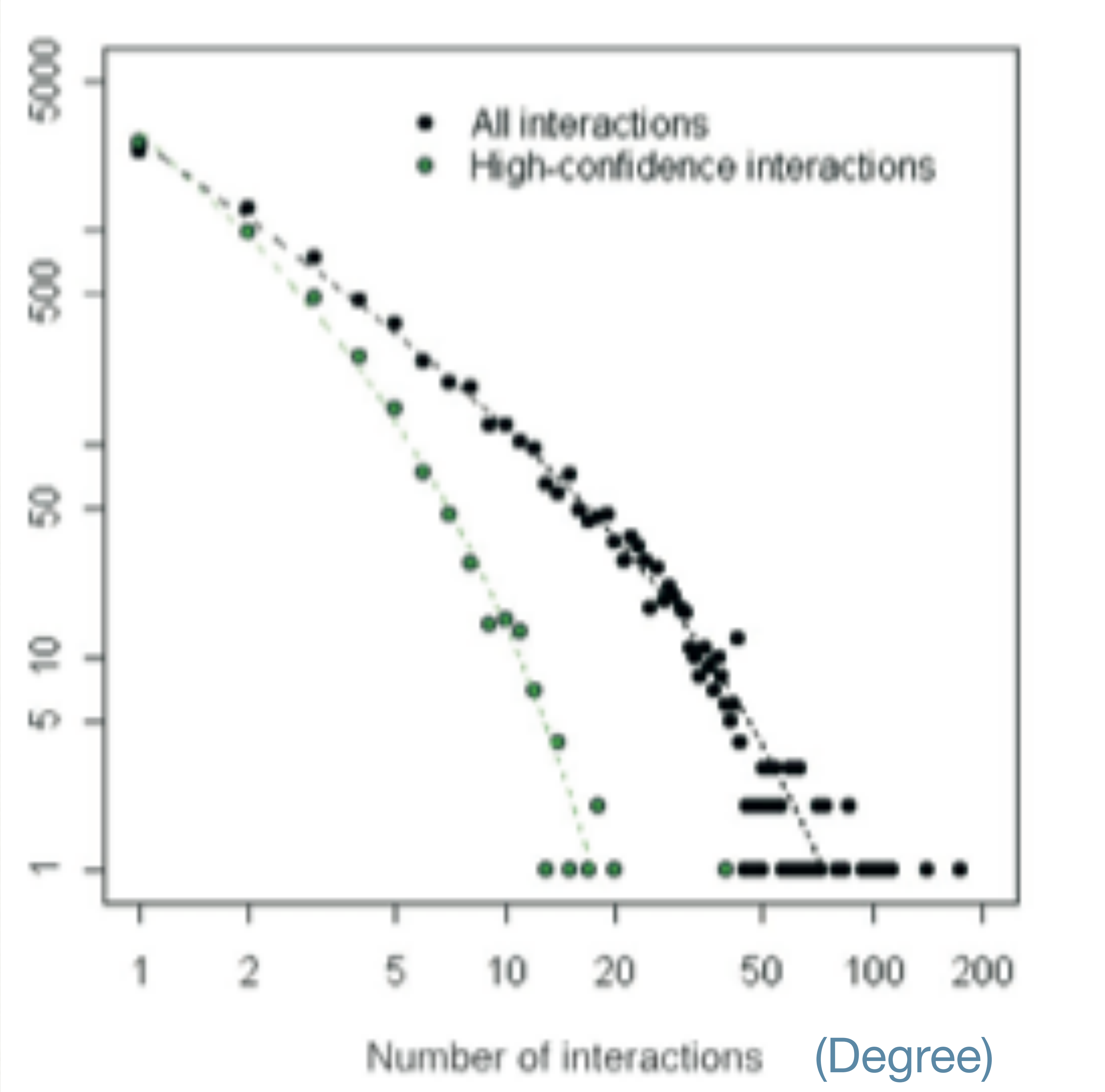
Clustering Coefficient $c_i = \dfrac{2e_i}{k_i(k_i-1)}$ $e_i$ : number of existing links (labeled in red) among the $k_i$ nodes that connect to node *i*

Betweenness $b_l = \sum_{ij} p_{ij}(l)/p_{ij}$

$p_{ij}$ : number of shortest paths between *i* and *j*

$p_{ij}(l)$ : number of shortest paths between *i* and *j* going through node *l*

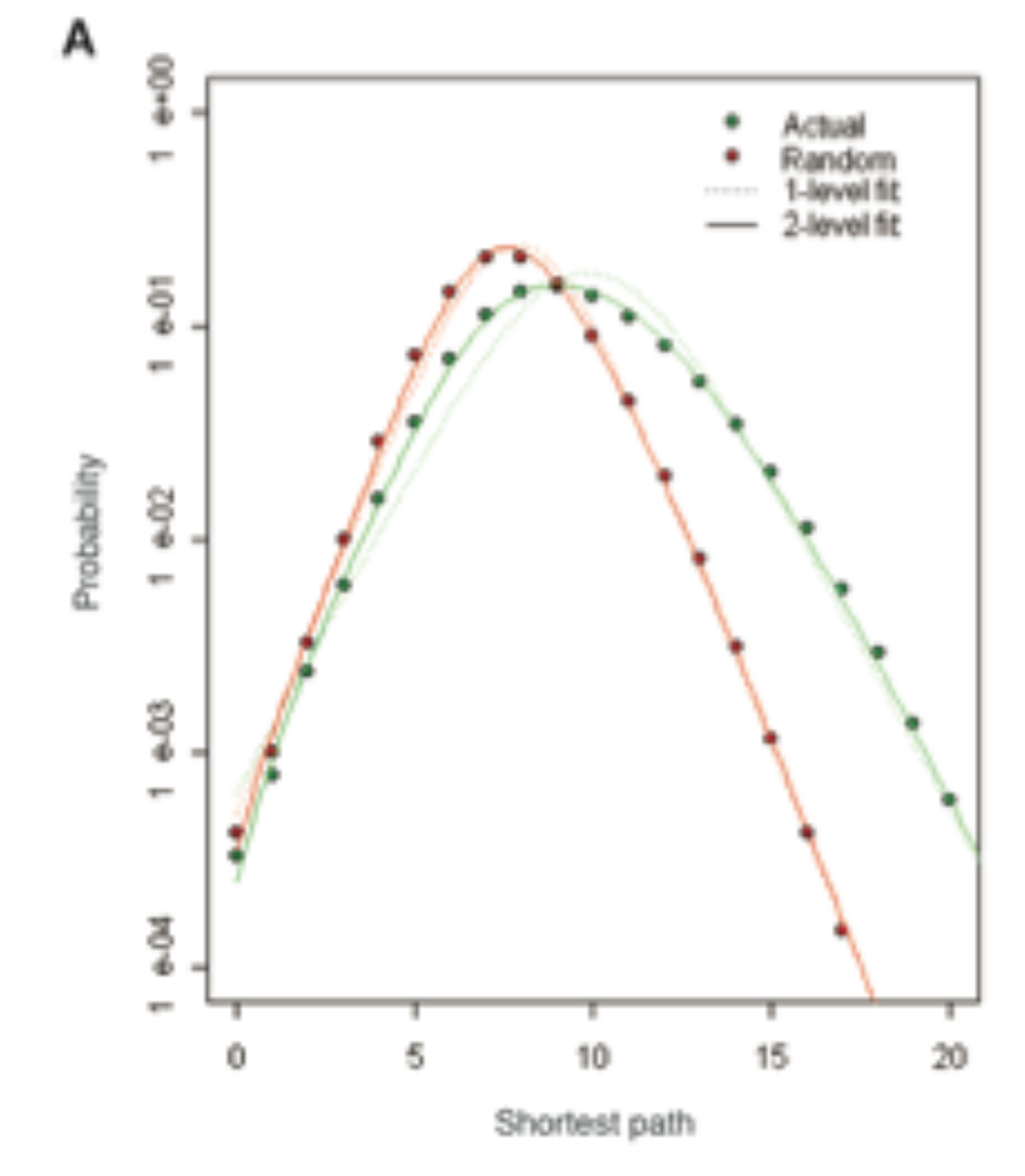
Zhu et al, 2007.



**Characteristic**: Huge # of low-degree nodes, with a few very high-degree nodes.

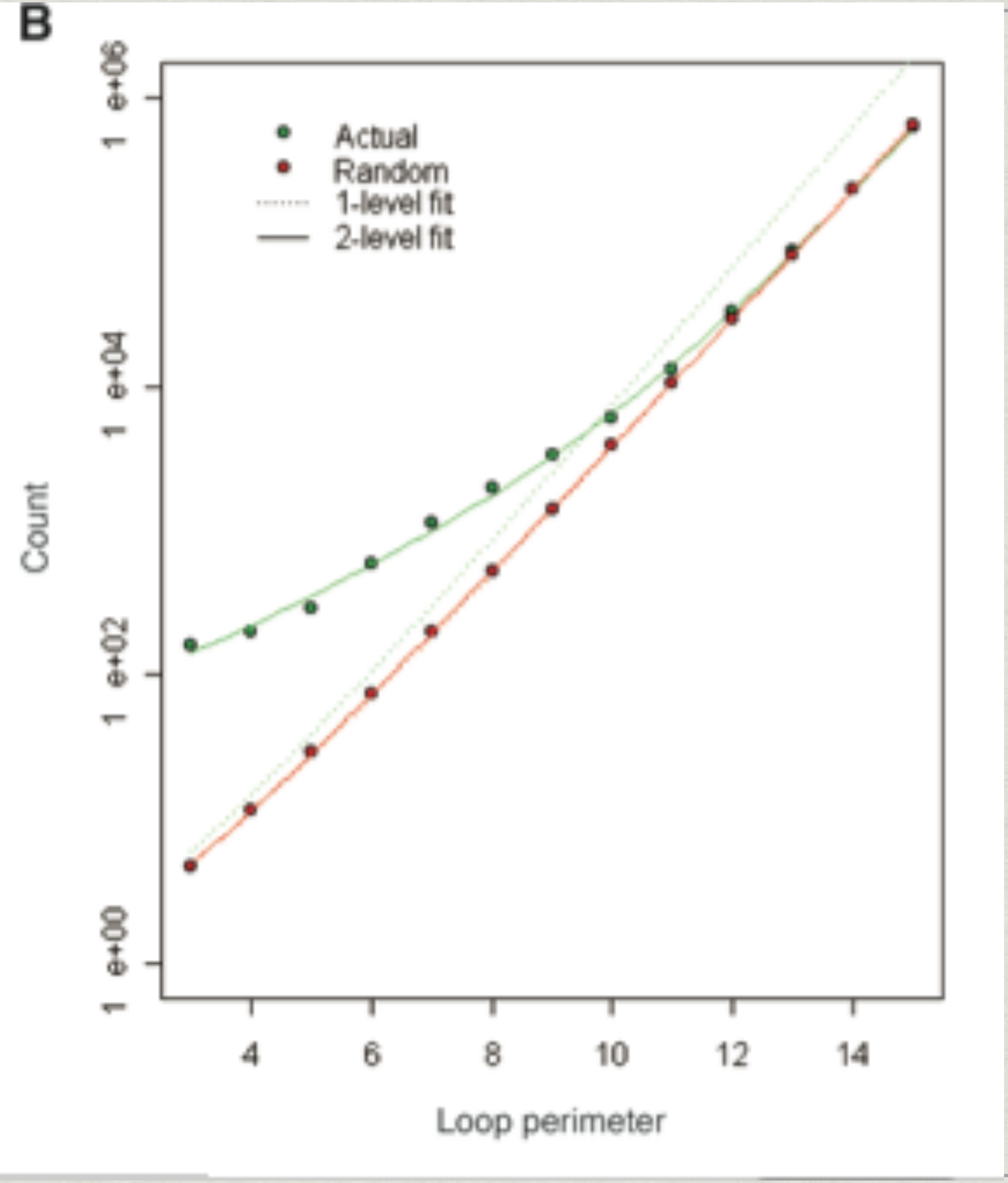
Giant connected component
= 3659 edges
= 3039 nodes

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

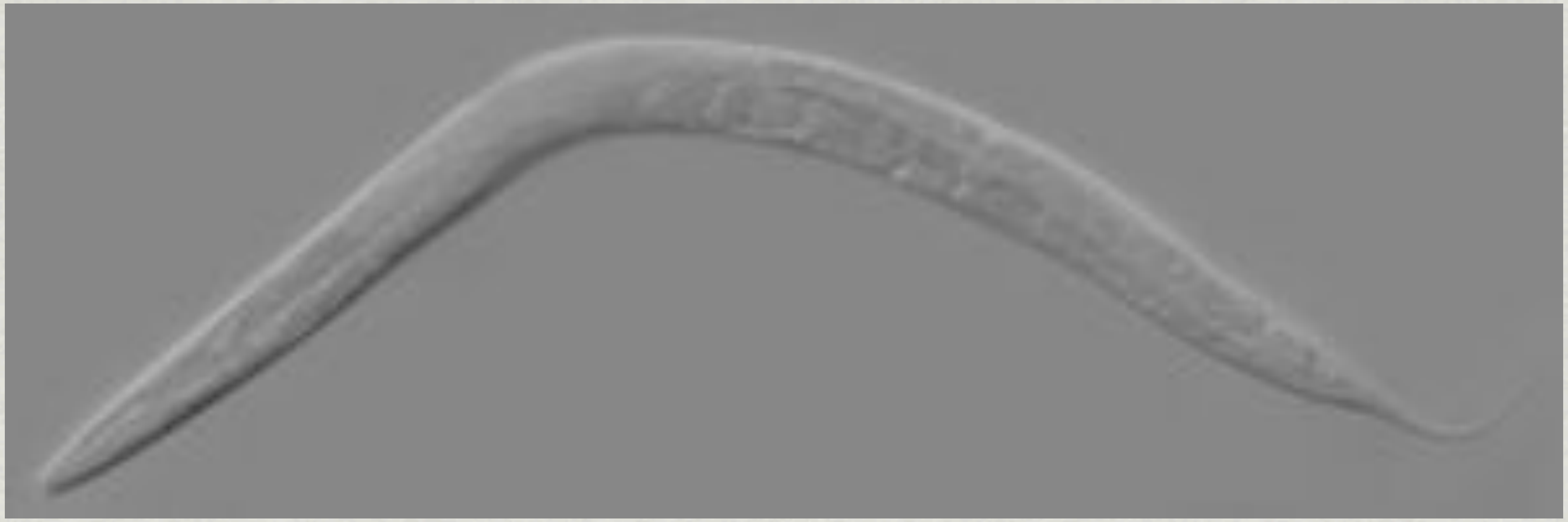


Node disjoint loops

Real network contains more triangles than random

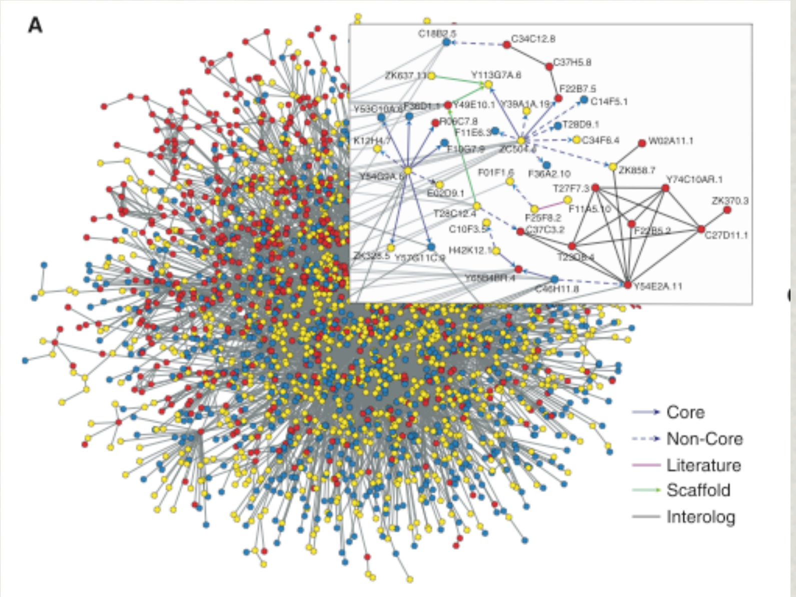


# *C. elegans*



# Li et al, 2004 Results:

| Interaction Set | # interactions | # proteins |
|---|---|---|
| Core | 2157 | 502 baits<br>1039 preys |
| Non-core | 1892 | 531 baits<br>1395 preys |



- “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

Sprinzak et al, 2003

fraction co-localized pairs in true edges

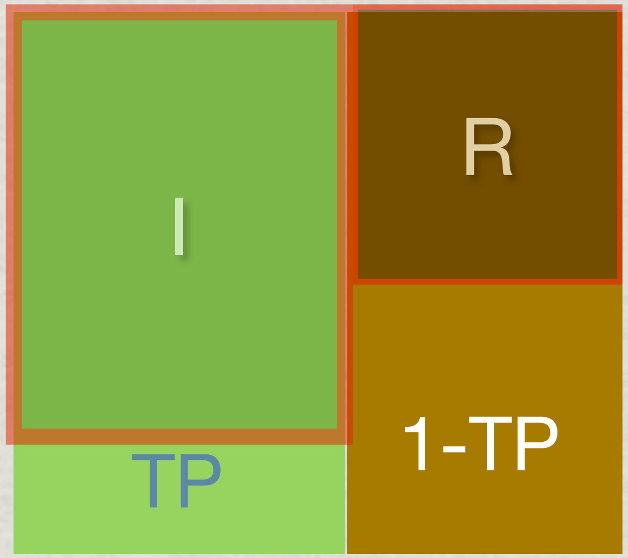
fraction of true edges

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

fraction of co-localized pairs in random edges

fraction of **co-localized** pairs in predicted edges

fraction of false positives



Assuming FP are 'random-like'

$$TP = (D - R) / (I - R)$$

# Estimating TP

$$TP = (D - R) / (I - R)$$

- D = fraction co-localized predicted edges
- R ≈ fraction of **all** pairs that are co-localized (~0.36)
- I ≈ 1 or 0.95 [assumed to be very high]

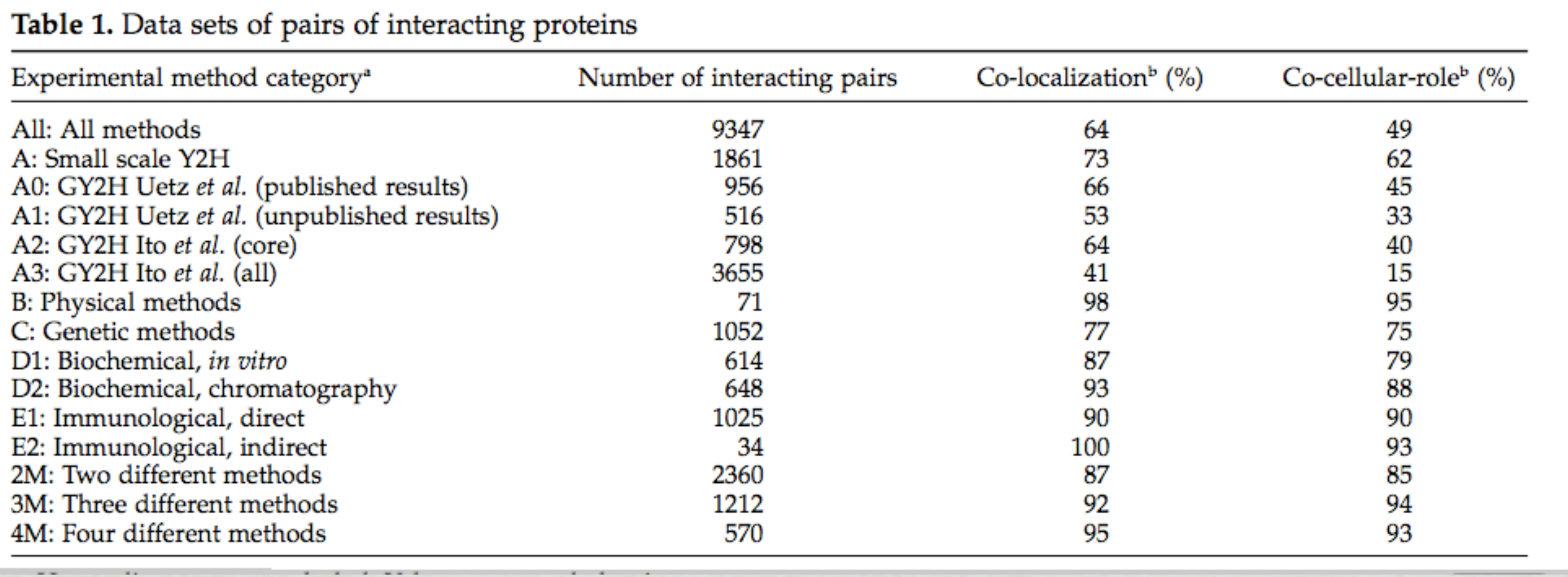
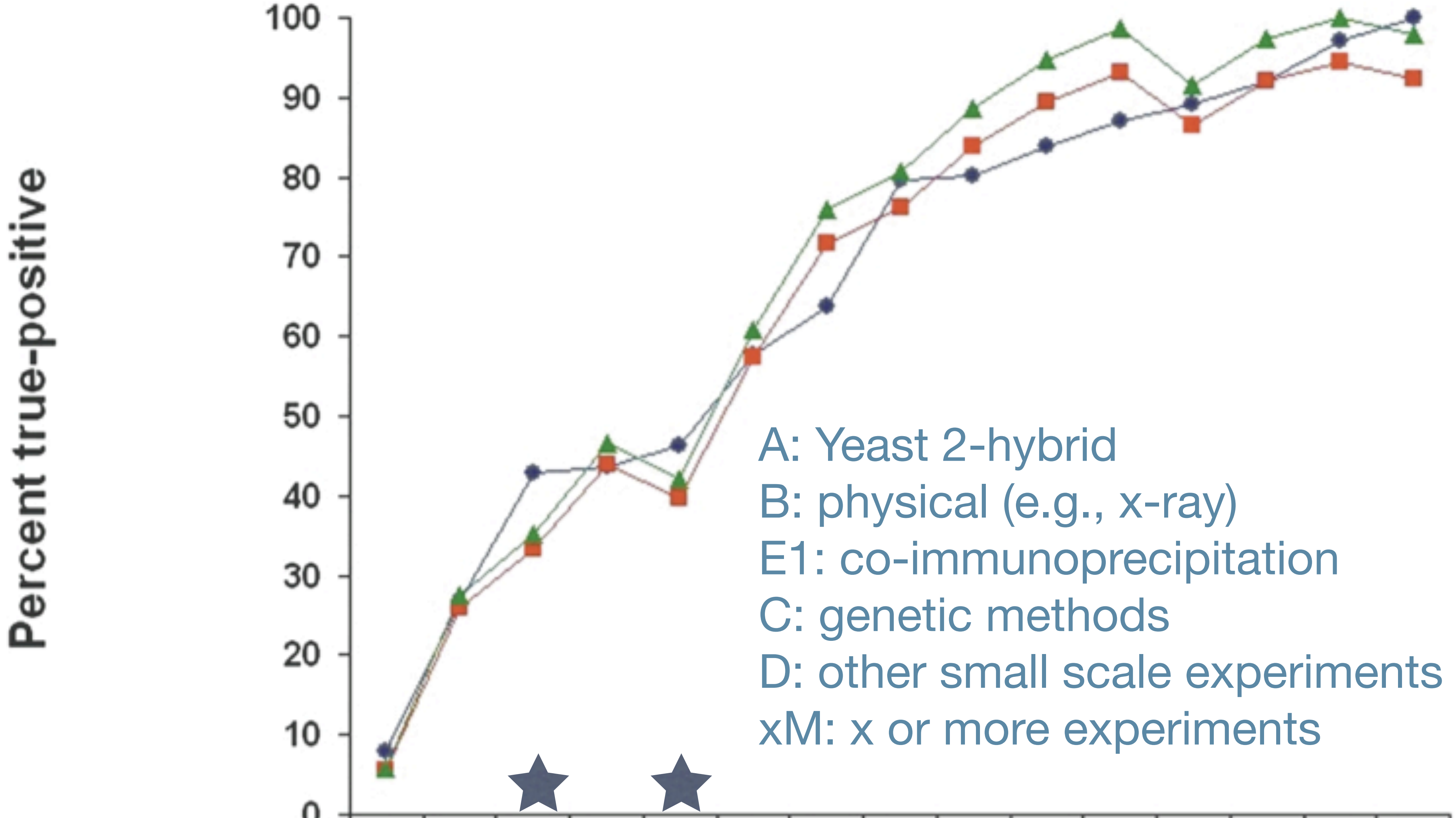


**Table 1.** Data sets of pairs of interacting proteins

| Experimental method category[a] | Number of interacting pairs | Co-localization[b] (%) | Co-cellular-role[b] (%) |
|---|---|---|---|
| All: All methods | 9347 | 64 | 49 |
| A: Small scale Y2H | 1861 | 73 | 62 |
| A0: GY2H Uetz *et al.* (published results) | 956 | 66 | 45 |
| A1: GY2H Uetz *et al.* (unpublished results) | 516 | 53 | 33 |
| A2: GY2H Ito *et al.* (core) | 798 | 64 | 40 |
| A3: GY2H Ito *et al.* (all) | 3655 | 41 | 15 |
| B: Physical methods | 71 | 98 | 95 |
| C: Genetic methods | 1052 | 77 | 75 |
| D1: Biochemical, *in vitro* | 614 | 87 | 79 |
| D2: Biochemical, chromatography | 648 | 93 | 88 |
| E1: Immunological, direct | 1025 | 90 | 90 |
| E2: Immunological, indirect | 34 | 100 | 93 |
| 2M: Two different methods | 2360 | 87 | 85 |
| 3M: Three different methods | 1212 | 92 | 94 |
| 4M: Four different methods | 570 | 95 | 93 |



A: Yeast 2-hybrid
B: physical (e.g., x-ray)
E1: co-immunoprecipitation
C: genetic methods
D: other small scale experiments
xM: x or more experiments

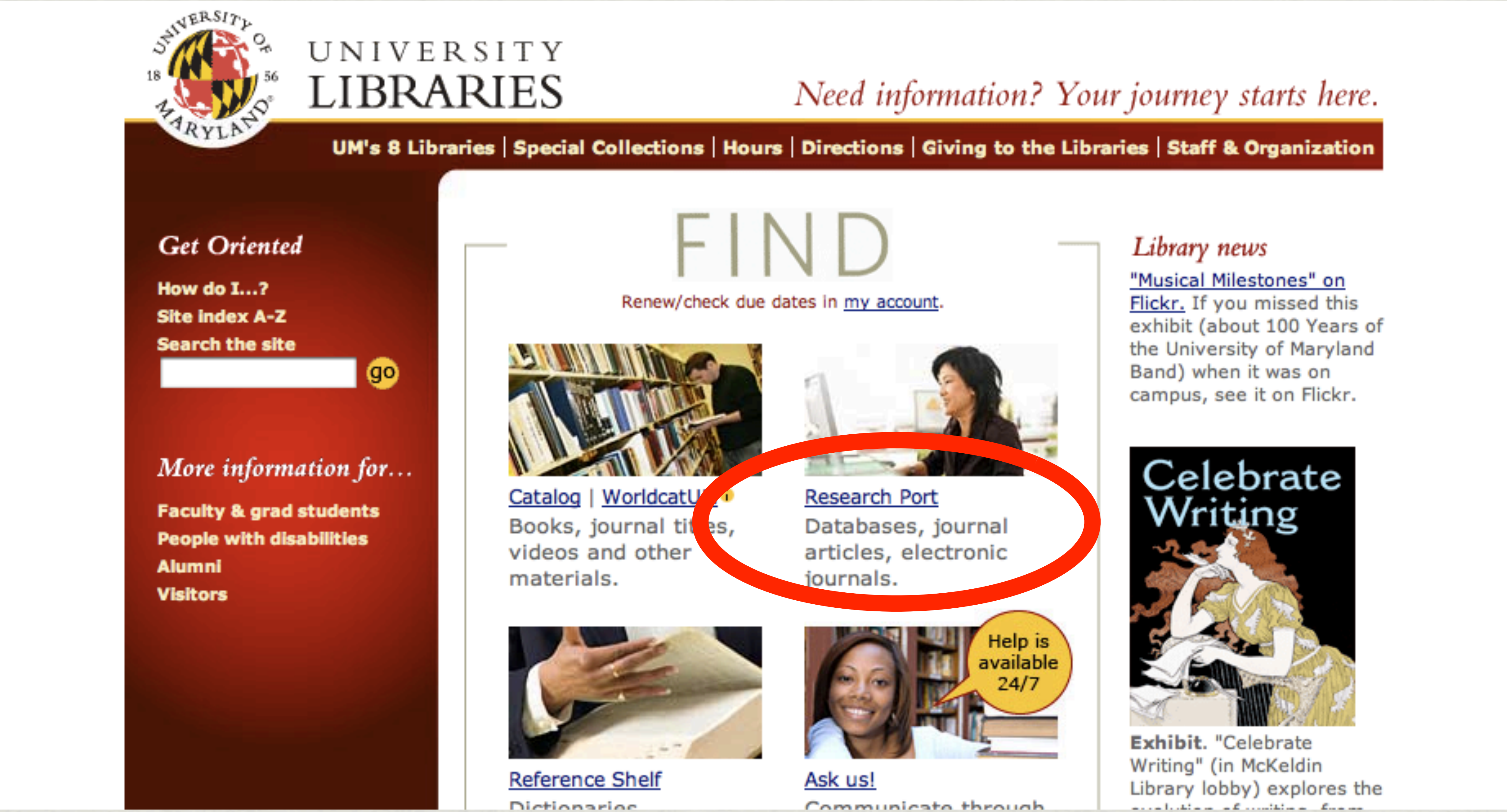
| | A3 | A1 | A2 | All | A0 | A | C | D1 | 2M | E1 | 3M | D2 | 4M | B | E2 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| —●— % TP by Localization, I=1 | 7.8 | 26.7 | 43.0 | 43.7 | 46.2 | 57.7 | 63.9 | 79.5 | 80.2 | 84.0 | 87.1 | 89.0 | 92.1 | 97.2 | 100. |
| —■— % TP by Cellular Role, I=1 | 5.5 | 25.9 | 33.3 | 43.9 | 39.7 | 57.4 | 71.8 | 76.3 | 83.8 | 89.4 | 93.2 | 86.5 | 92.1 | 94.5 | 92.4 |
| —▲— % TP by Cellular Role, I=0.95 | 5.9 | 27.4 | 35.3 | 46.5 | 42.0 | 60.8 | 76.0 | 80.8 | 88.7 | 94.7 | 98.6 | 91.6 | 97.5 | 100. | 97.8 |

Experimental method category

**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 ($I = 1$); red, green, percentage of TP based on shared cellular-role for $I = 1$ (red) and $I = 0.95$ (green). $I$ 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.



UNIVERSITY LIBRARIES

*Need information? Your journey starts here.*

UM's 8 Libraries | Special Collections | Hours | Directions | Giving to the Libraries | Staff & Organization

*Get Oriented*

How do I...?
Site Index A-Z
Search the site

go

*More information for...*

Faculty & grad students
People with disabilities
Alumni
Visitors

FIND

Renew/check due dates in my account.

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

Research Port
Databases, journal articles, electronic journals.

Reference Shelf
Dictionaries

Ask us!
Help is available 24/7
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.

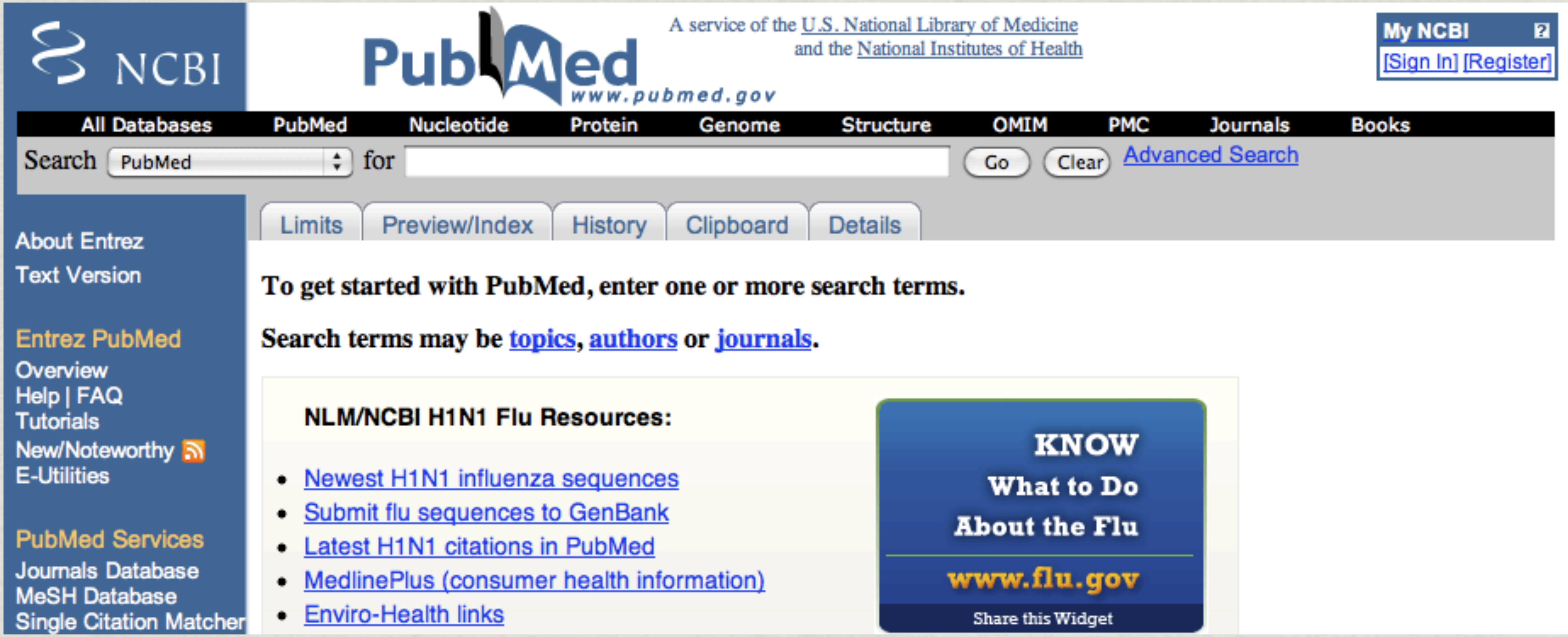
Celebrate Writing

**Exhibit.** "Celebrate Writing" (in McKeldin Library lobby) explores the

# PubMed: Where nearly all the relevant papers can be found.

Relevant Journals: Science, Nature, PLoS Biology, PLoS Comp. Biology, Bioinformatics, BMC Bioinformatics, Journal Computational Biology, Genome Biology

http://www.ncbi.nlm.nih.gov/PubMed/



A service of the U.S. National Library of Medicine and the National Institutes of Health

To get started with PubMed, enter one or more search terms.

Search terms may be topics, authors or journals.

NLM/NCBI H1N1 Flu Resources:

- Newest H1N1 influenza sequences
- Submit flu sequences to GenBank
- Latest H1N1 citations in PubMed
- MedlinePlus (consumer health information)
- Enviro-Health links

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

Asimov, The Human Brain, 1965