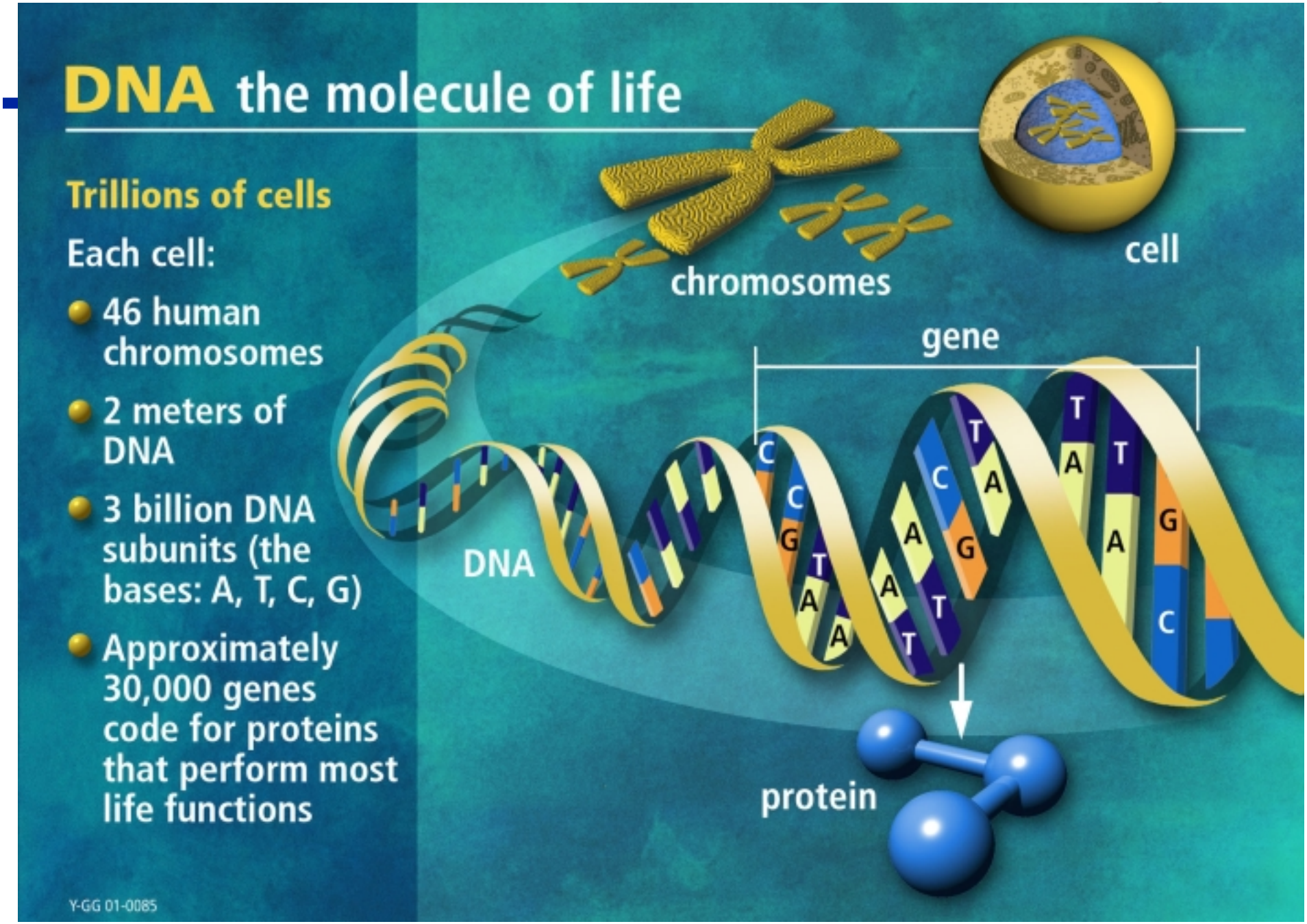
Multi-protein Complex Data Mining for Predicting Protein Interactions and Functional Organizations

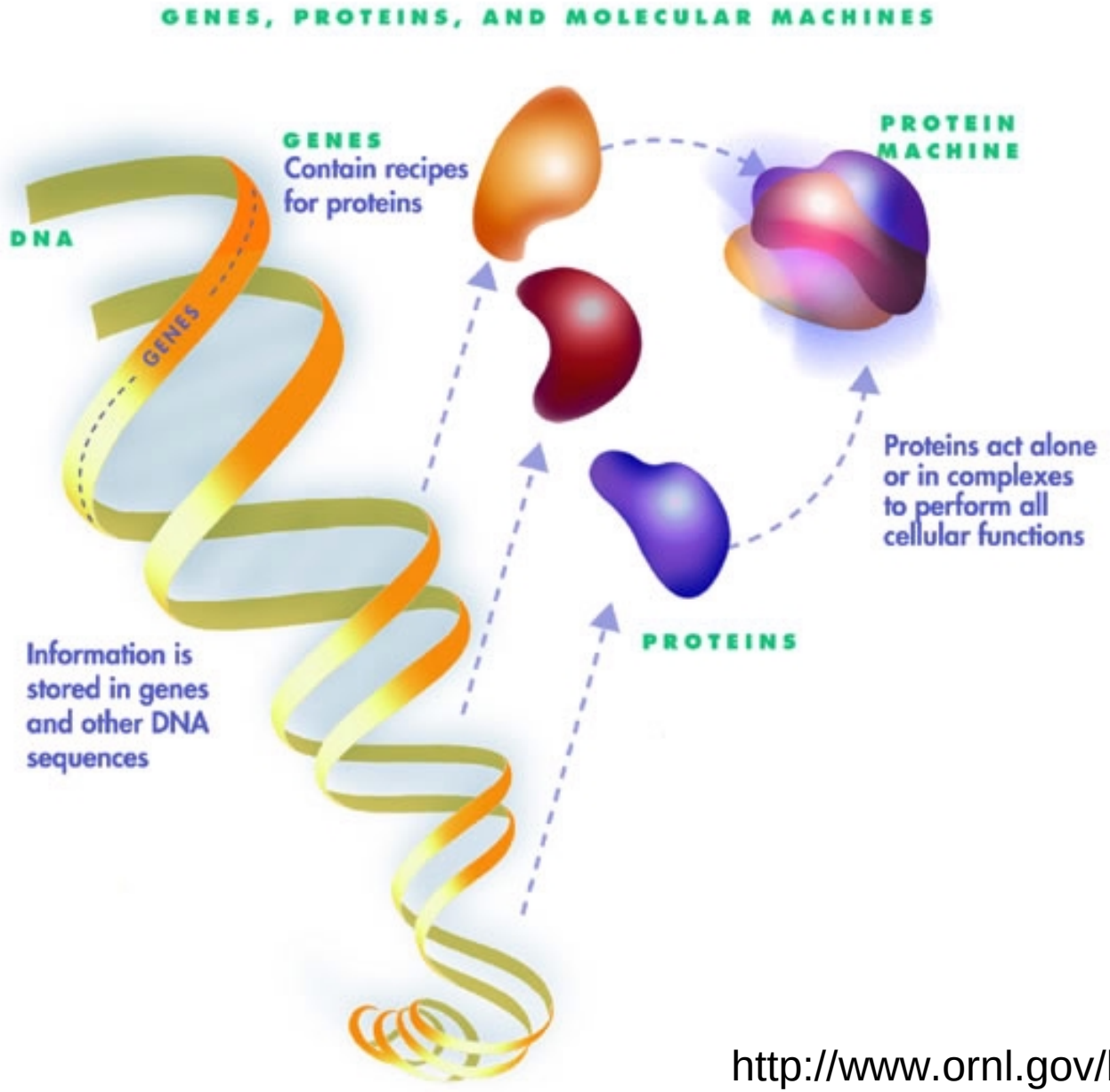
Xiaofeng He
Computational Research Division
Lawrence Berkeley National Laboratory

Joint work with
Chris Ding, Computational Research Division, LBL
Richard Meraz, Physical Biosciences Division, LBL
Steve Holbrook, Physical Biosciences Division, LBL

3/23/2004

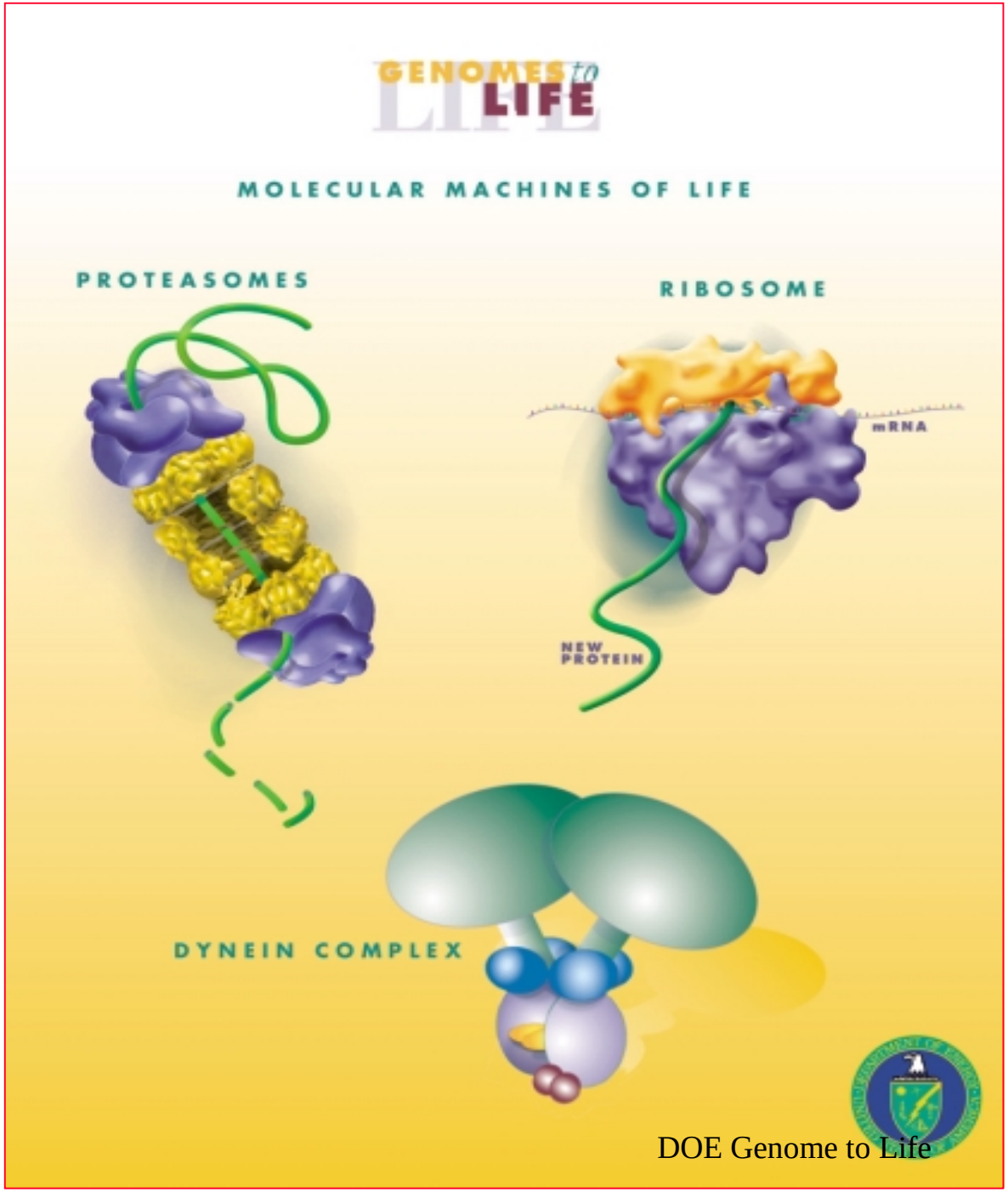


DOE Genome to Life



http://www.ornl.gov/hgmis

Protein Complex



DOE Genome to Life

- Identify and characterize protein complexes
- Identify gene regulatory networks
- Microbial genome
- Systems level modeling

Protein – Protein interactions

Proteins carry out tasks together with other proteins => Protein – protein interactions

- Proteins bind each other
- Binary interactions
- Multi-protein complexes (assemblies)

Multi-Protein complex

1. Multi-protein complex: module of functionally related proteins.
2. Cellular process carried out by multi-protein complex.
3. Higher order functional units.

Challenge for Post-Genomic biology: protein interaction

Protein interactions traditionally studied individually by genetic, biochemical and biophysical techniques.

Current progress:

1. Completion of dozens of genome sequencing projects
2. New high-throughput experimental methods to determine functions of newly discovered genes

Systematically analyze interactions / coordinations of proteins on genomic scale

Outline:

- Protein–protein interaction and protein complex
- Protein interaction experiments and data
- Unified representation of protein complex data
    1. Protein – protein complex network (Bipartite graph)
    2. protein – protein network
    3. protein complex – protein complex network
- MinMaxCut spectral clustering
- Main computational results: protein cluster & supercomplex
- Biological significance of discovered cluster & supercomplex

High-throughput methods for detecting protein interactions

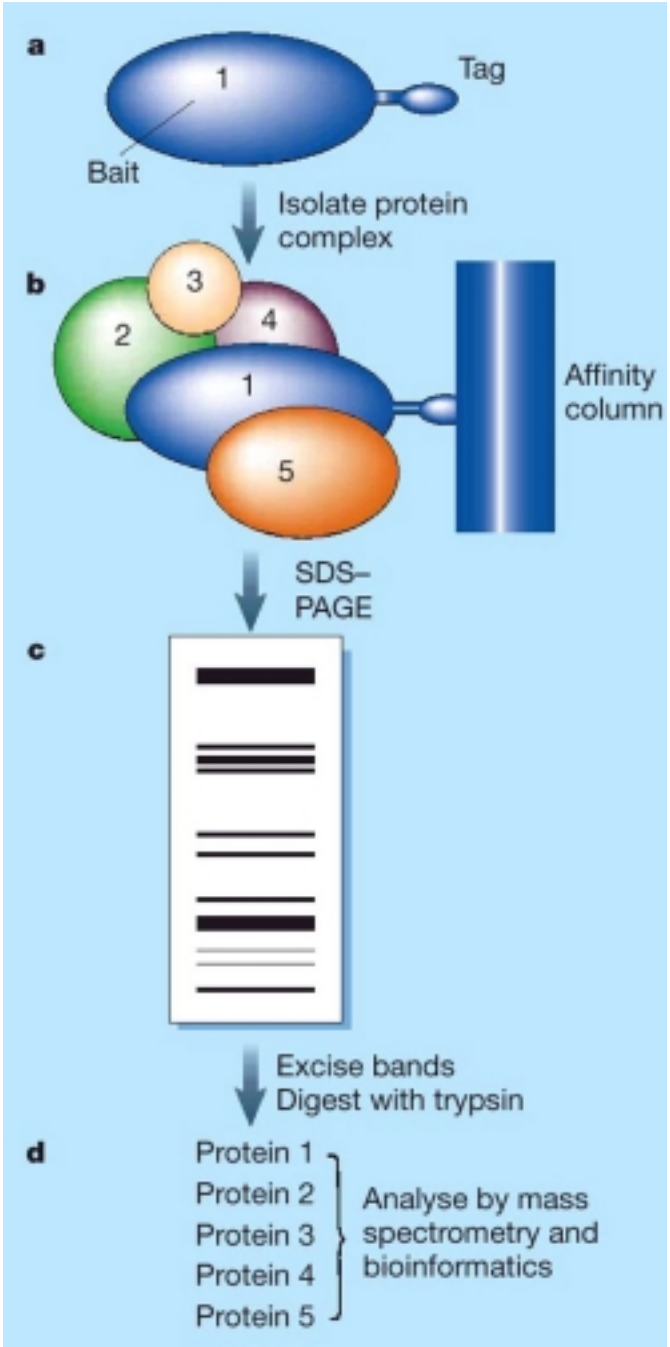
Recent high-throughput analyses of protein interaction datasets in S. cerevisiae:

- Two-hybrid dataset by Uetz *et al* 2000 (the first comprehensive study in yeast)
- Two-hybrid dataset by Ito *et al* 2001 (broad coverage in yeast)
- HMS-PCI dataset by Ho *et al* 2002
- TAP-MS dataset by Gavin *et al* 2002

TAP-MS dataset is the most reliable one (Deng, *et al*)

Protein interaction experiments

- **Two-hybrid Assay (fuse proteins)**
  - —Binary interactions
  - —Capture transient and unstable interactions
- **Mass Spectrometry**
  - —TAP-MS: Tandem affinity purification
  - —HMS-PCI: high throughput protein interaction id.
  - —Use bait proteins
  - —Capture multi-protein complexes
- **Problems:**
  - —**Results do not agree. Lots of noise**

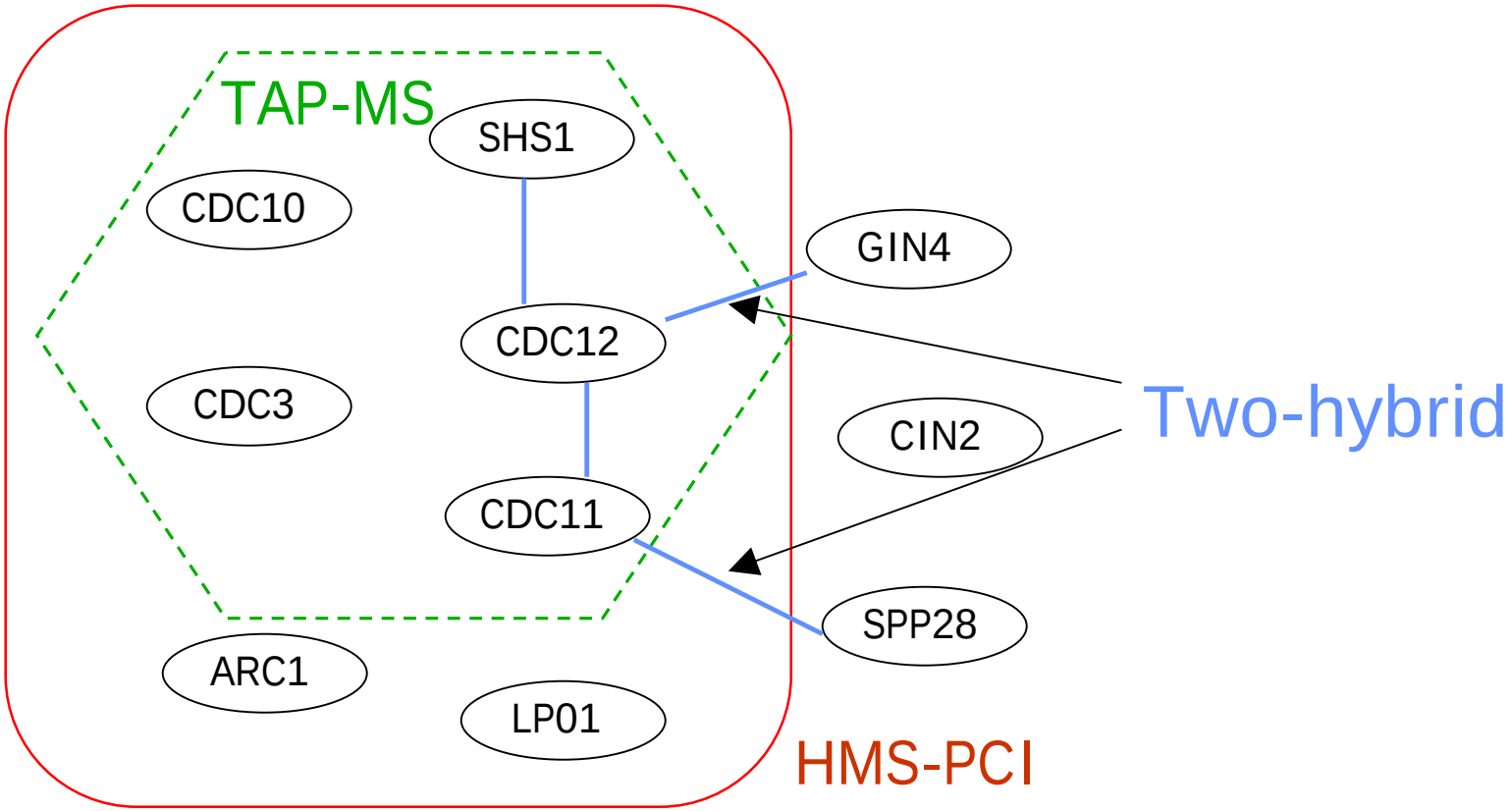


- Tandem-Affinity Purification coupled with Mass-Spectrometry (TAP-MS) determines the constituents of multi-protein complexes.

Proved to be the most reliable dataset (Deng, *et al)*

Gavin AC, *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 2002;415(6868):141-147.

Counting Protein interactions



Protein interaction database

Small overlaps among different experiments.

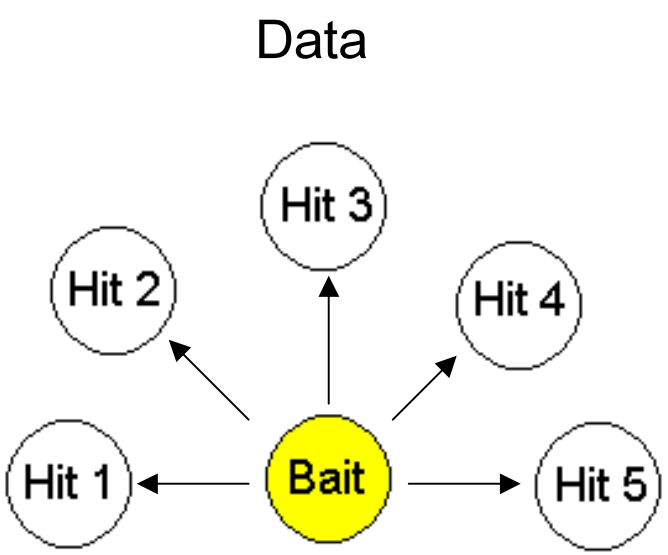
	ITO *et al*	Uetz *et al*	Gavin *et al*	Ho *et al*
Ito *et al*	**4363**	186	54	63
Uetz *et al*	186	**1403**	54	56
Gavin *et al*	54	54	**3222**	198
Ho *et al*	63	56	198	**3596**
Small-scale experiments in DIP	442	415	528	391

Copied from Salwinski and Eisenberg, 2003

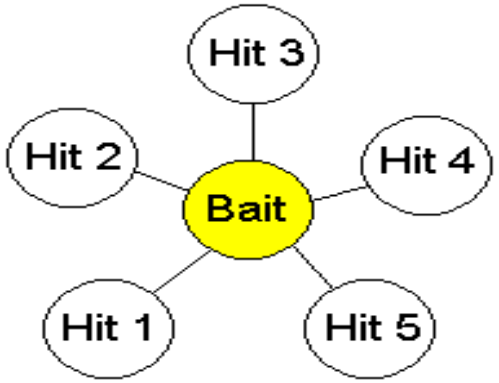
Protein interaction databases

Database	URL
DIP	dip.doe-mbi.ucla.edu
MIPS	mips.gsf.de
BIND	www.bind.ca
YPD	www.proteome.com/YPDhome.html
The GRID	biodata.mshri.on.ca/grid/servlet/index
LivDIP	dip.doc-mbi.ucla.edu/ldip.html
PREDICTOME	predictome.bu.edu
STRING	www.bork.embl-heidelberg.de/STRING
interDOM	InterDom.lit.org.sg
PreBIND	Bind.ca

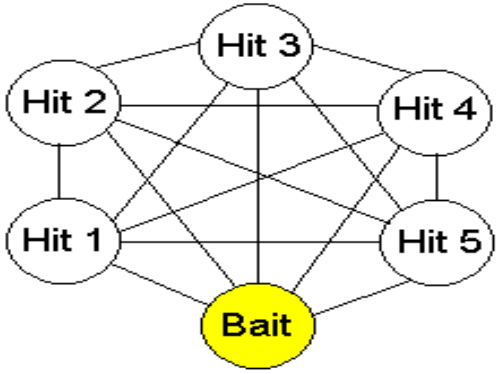
Previous attempts to analyze TAP-MS data used simplified models



Spoke Model



Matrix Model



-find *k*-cores (Bader and Hogue, 2002)

-find cliques (Spirin and Mirn y, 2003)

-Hypergraph – *k*-core (Pothen, 2003)

Previous attempts to analyze TAP-MS data used simplified models

binary interactions with unit weights

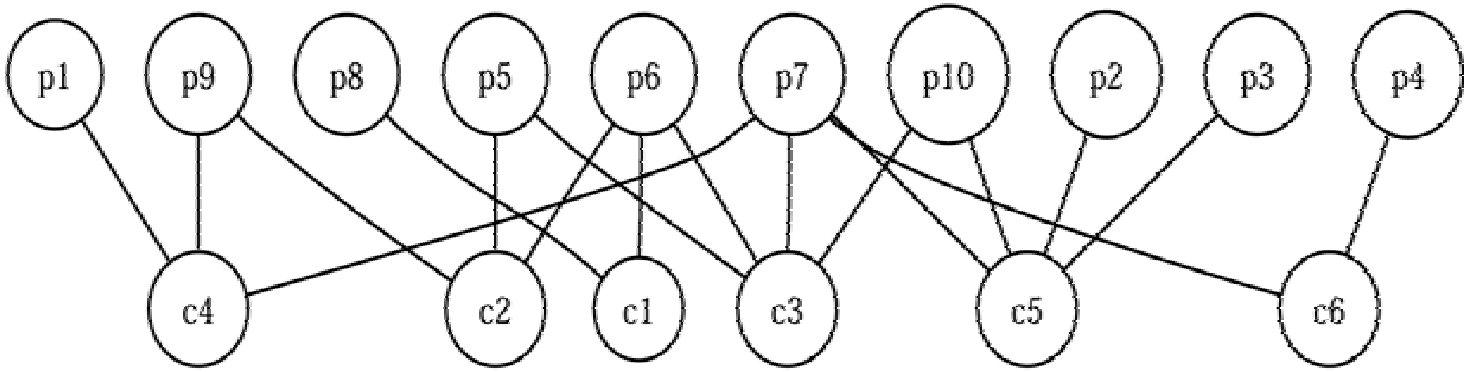
Limitations:

- Oversimplify realistic physical interactions between protein;
- Unable to represent diversity of interconnected cellular processes.

Previous models vs. our models

Previous Models	Our Model
Un-weighted interaction strength - oversimplified	Weighted interaction strength -more realistic
Focus only on protein – protein interactions	**Unified representation** from protein complex data to derive protein – protein interactions complex – complex network
K-core, clique	Vigorous clustering

Bipartite graph model of protein complex data



P-nodes represent proteins and *c*-nodes represent protein Complexes

Proteins and multi-protein complexes form the bipartite graph (p-c interaction network)

Unified representation of protein complex data

Dual relationship between protein and protein complex is specified by adjacency matrix ***B***.

Interaction strength of protein – protein network:

$$(BB^T)_{ij} = \begin{pmatrix} \#\text{ of protein complexes} \\ \text{containing both proteins } p_i, p_j \end{pmatrix}$$

Interaction strength of protein complexe – protein complex network:

$$(B^T B)_{ij} = \begin{pmatrix} \#\text{ of proteins shared by} \\ \text{protein complexes } c_i, c_j \end{pmatrix}$$

Bipartite graph model of protein complex data

Unified representation

1. Protein-protein (p-p) interaction network arises naturally
   Strength of interaction: number of protein complexes containing the pair of proteins

2. Protein complex – protein complex (c-c) interaction network also arises.
   Strength of interaction: number of common proteins contained

3. System-level understanding of cellular processes

Cluster interaction networks

Previous: *k*-core, clique => densely connected subgraphs

Our work: clustering --- a more consistent and flexible way to find clusters in a mathematically rigorous way

Cluster Cohesion to assess cluster connectedness:

Cut a cluster *G* into subsets: *A,B*

Cohesion = between-subset connections

weighed by within-subset connections

Large cohesion => highly connected

Spectral clustering p-p and c-c network

MinMaxCut spectral clustering method:

Minimize similarity between clusters,

Maximize similarity within cluster

$$\Rightarrow \quad J_{MMC}(A,B) = \frac{s(A,B)}{s(A,A)} + \frac{s(A,B)}{s(B,B)}$$ **= cohesion**

where $s(A,B) = \sum_{i\in A}\sum_{j\in B} W_{ij}$

Minimizing J_{MMC} leads to

$$\min_q J(q) = \min_q \frac{q^T(D-W)q}{q^T Dq}$$

and the solution is given by

$$(D-W)q = \lambda Dq$$

Ding, He, Zha, Gu, Simon (2001)

Biological usefulness of Clusters

Protein Cluster from protein-protein interaction network:

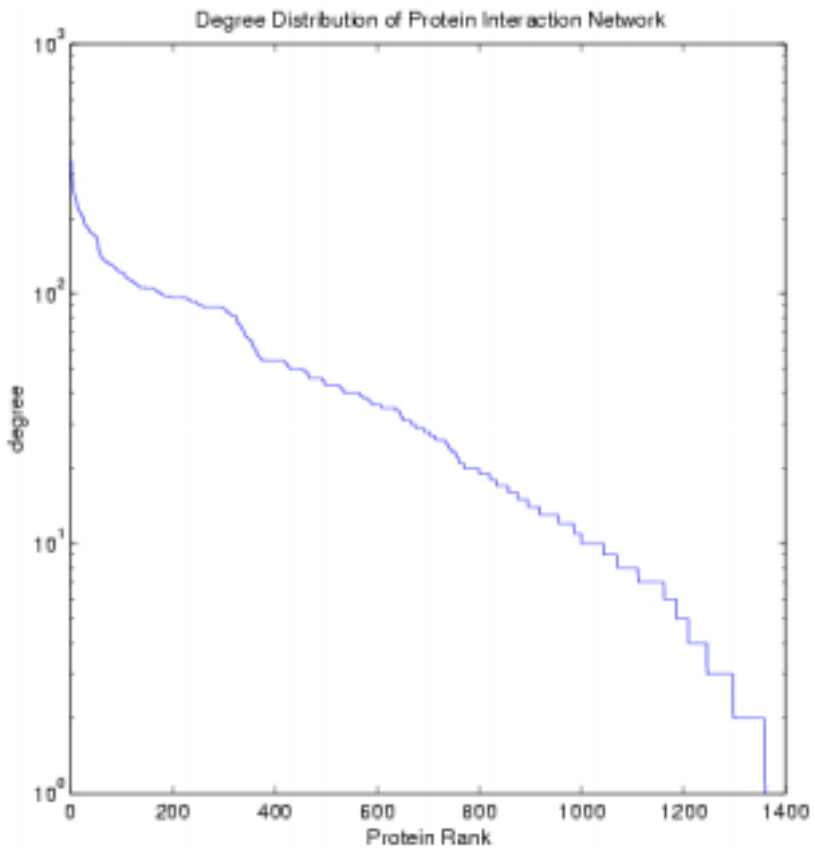
1. Assign annotations (functions) to uncharacterized proteins.
2. Predict possible functions for their orthologs in other species.
3. Predict biologically relevant modules carrying out cellular functions

Supercomplex from protein complex – protein complex network:

Detect higher order organization of the proteome.

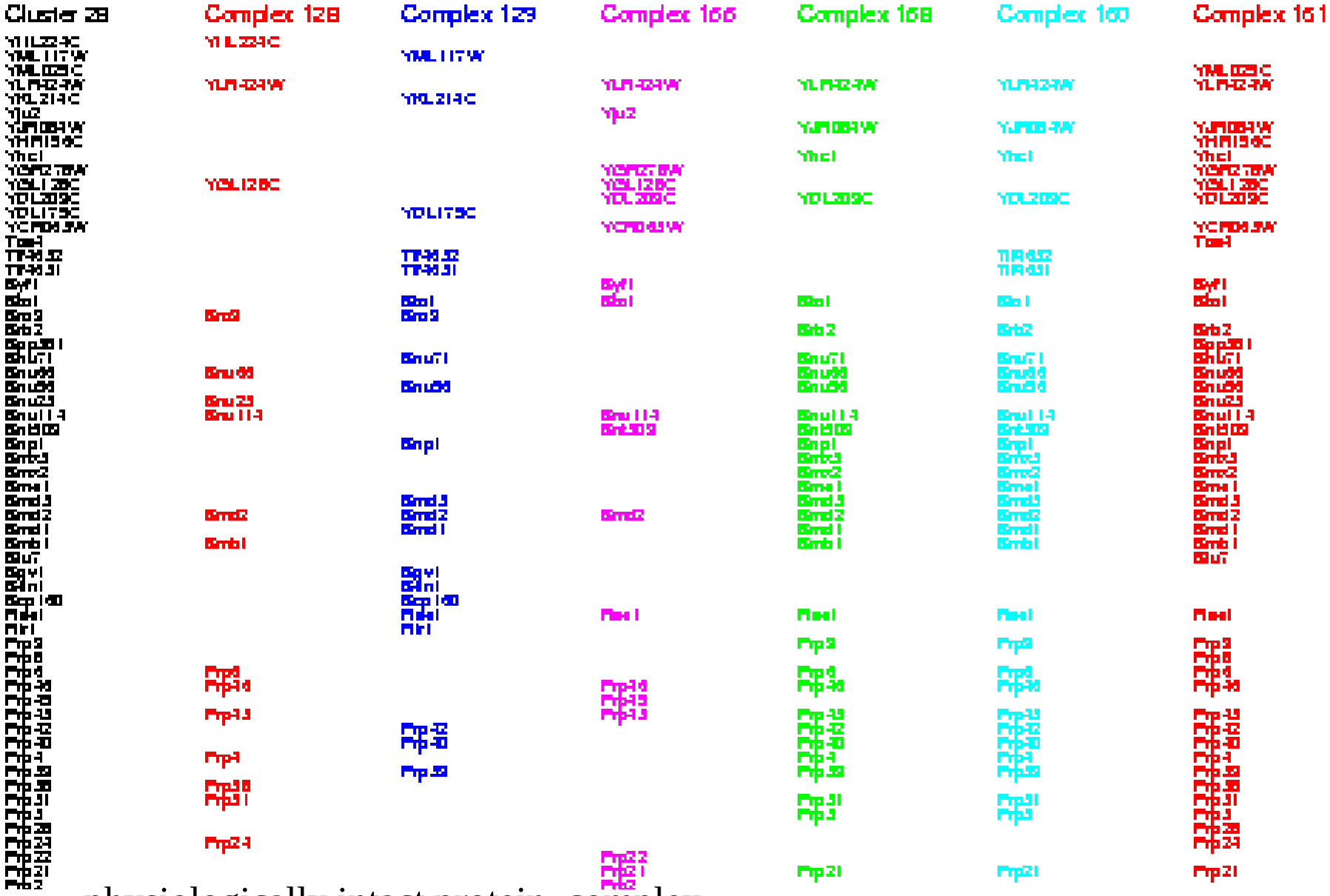
Provide a more system-level picture of protein interactions.

Main results of protein complex data analysis



Distribution of degrees in protein-protein interaction network, a scale-free network.

Predicted ---------- Experimental Protein Complex ---------



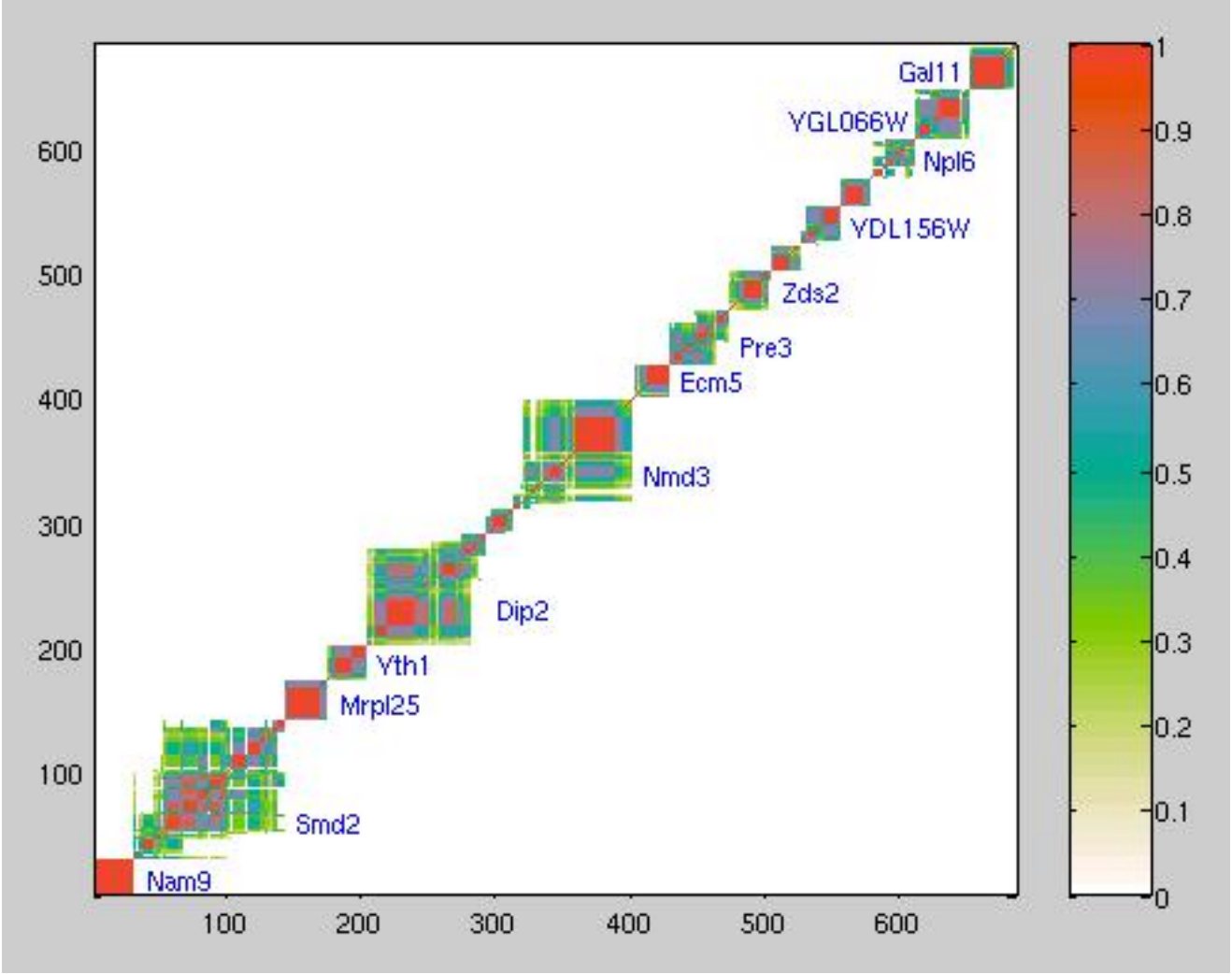
physiologically intact protein complex

Predicted clusters of protein-protein network

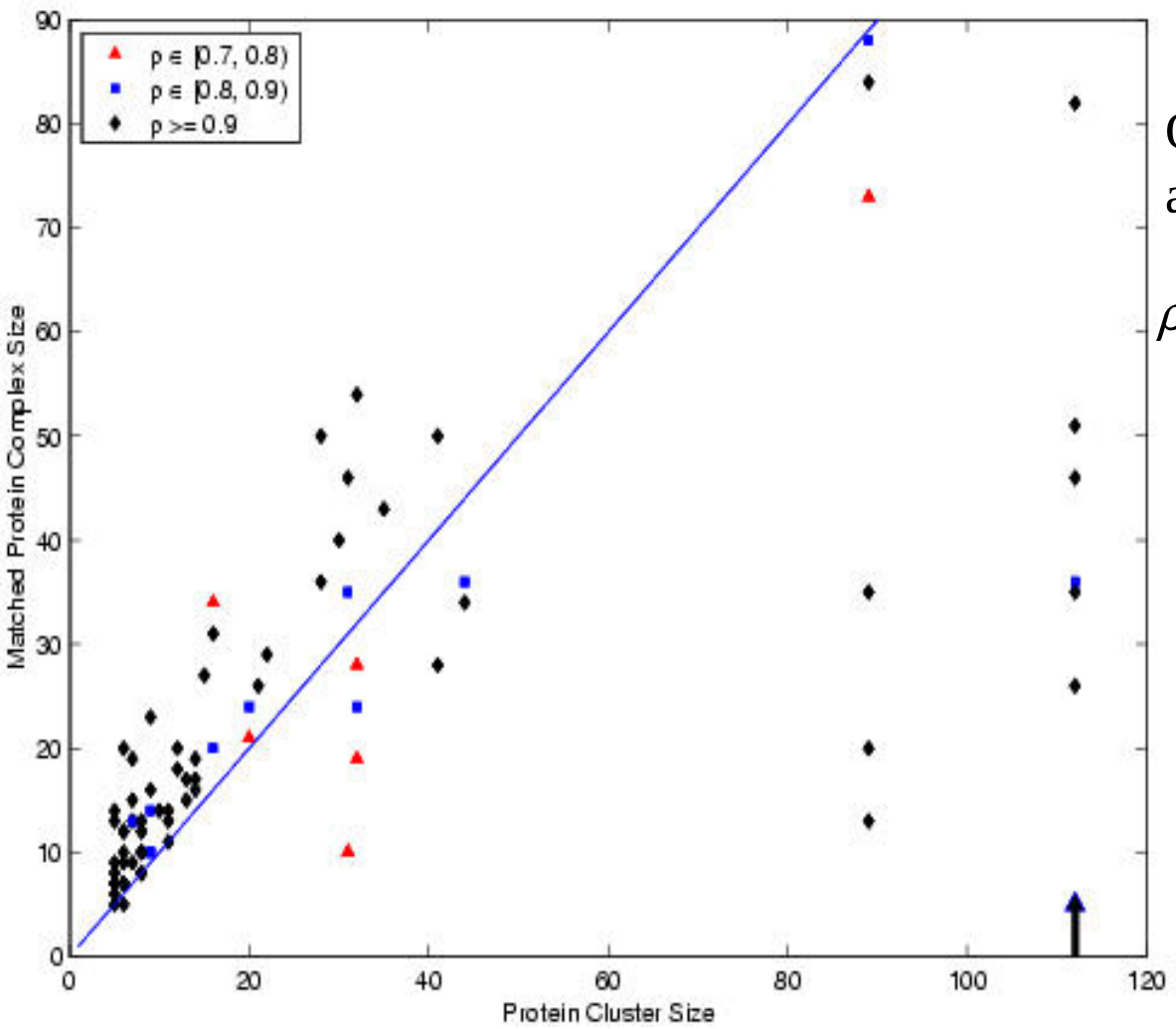
Interaction strength between gene varies, more realistic

Protein clusters obtained via clustering

Gives a comprehensive description of protein complex



Discovered Protein clusters vs. experimental protein complexes



Overlap between protein clusters and protein complexes defined as

$$\rho = n(P_k, c_j) / \min(|P_k|, |c_j|)$$

- Discovered protein clusters highly overlap with experiment complexes
- Uncharacterized proteins in discovered clusters might infer novel functions

Implications of discovered protein clusters on protein interactions: *F*-statistics

F - statistics of amino acids and physical properties across all protein clusters measure statistical significance

Lys	100	Asn	56	Val	30	Ile	24
Asp	89	Gln	50	Tyr	29	Ser	23
Arg	73	Cys	39	Met	29	Leu	22
Pro	70	His	33	Trp	28	Gly	21
Glu	66	Ala	31	Thr	28	Phe	21
pI	169	Basic	149	Acidic	97	MW	60
Aromatic	30	Helix	37	Beta-Sheet	33	Coil	27

$$F = \frac{1}{K-1}\sum_{k=1}^{K} n_k(\bar{f}_k - \bar{f}) \Big/ \frac{1}{n-K}\sum_{k=1}^{K}(n_k - 1)\sigma_k^2$$

Implications of discovered protein clusters on protein interactions: *F*-statistics

Lys	100	Asn	56	Val	30	Ile	24
Asp	89	Gln	50	Tyr	29	Ser	23
Arg	73	Cys	39	Met	29	Leu	22
Pro	70	His	33	Trp	28	Gly	21
Glu	66	Ala	31	Thr	28	Phe	21
pI	169	Basic	149	Acidic	97	MW	60
Aromatic	30	Helix	37	Beta-Sheet	33	Coil	27

Lys, Gln, Arg, Asn, Asp are most significant: => electrostatic forces are dominant surface factors influencing protein interactions

Arg is significant: => hydrogen bonding is important

Pro is significant: => hydrophobic interactions has strong stabilizing

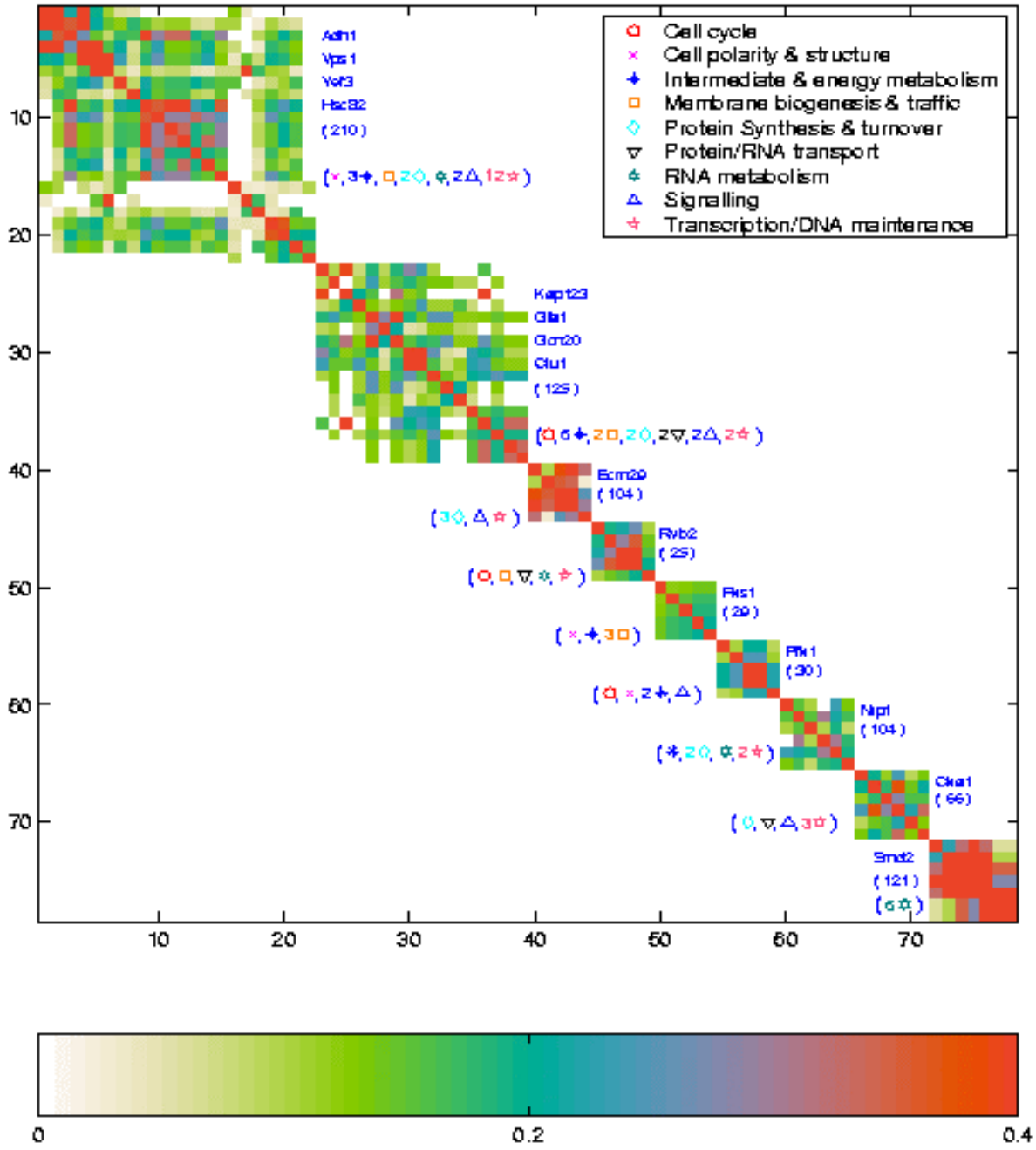
Protein complex - Protein complex

Higher order organization

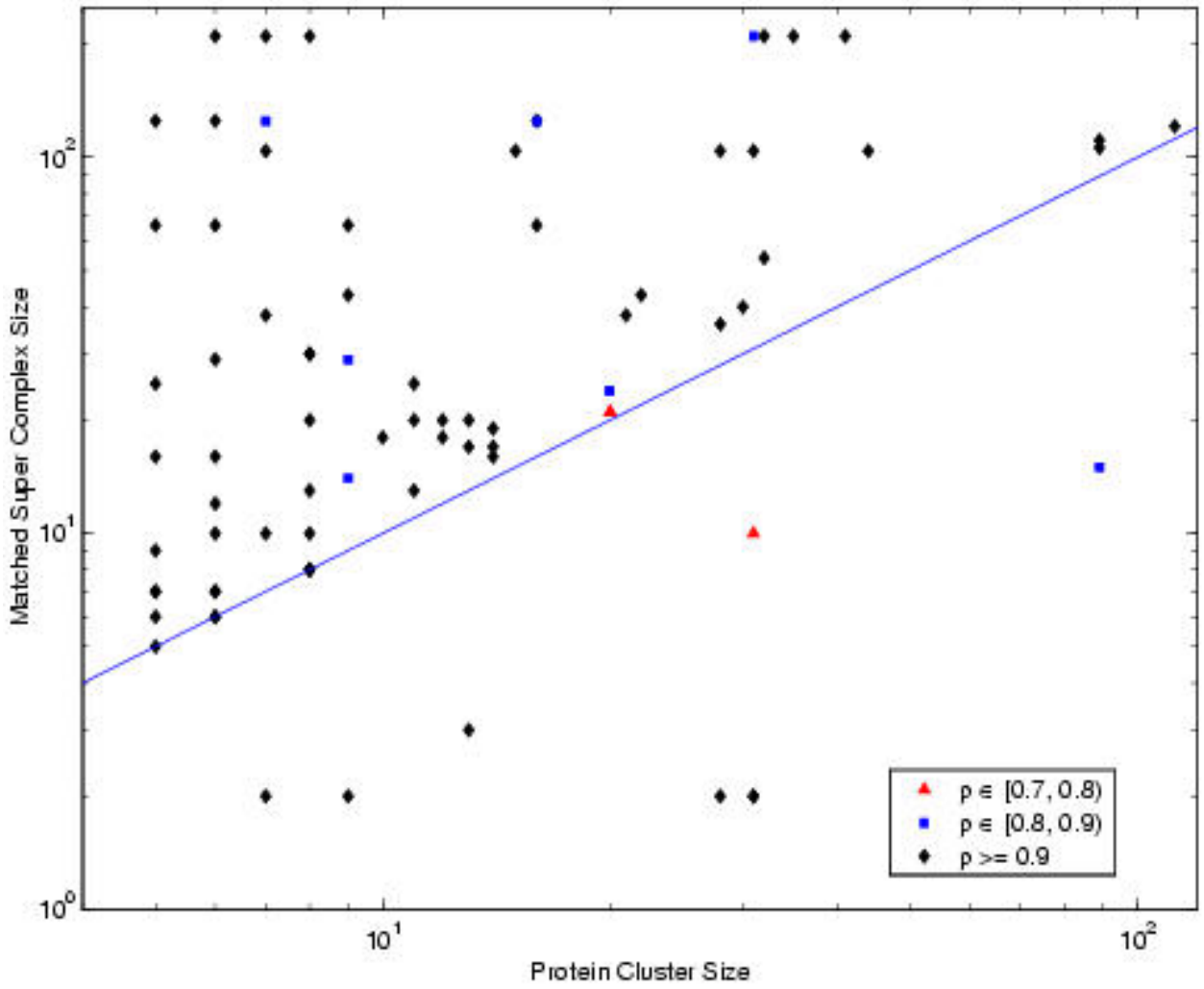
Supercomplex:

Cellular Process information more apparent:

Chromatin dynamics, transcription regulation, cell cycle control, biogenesis



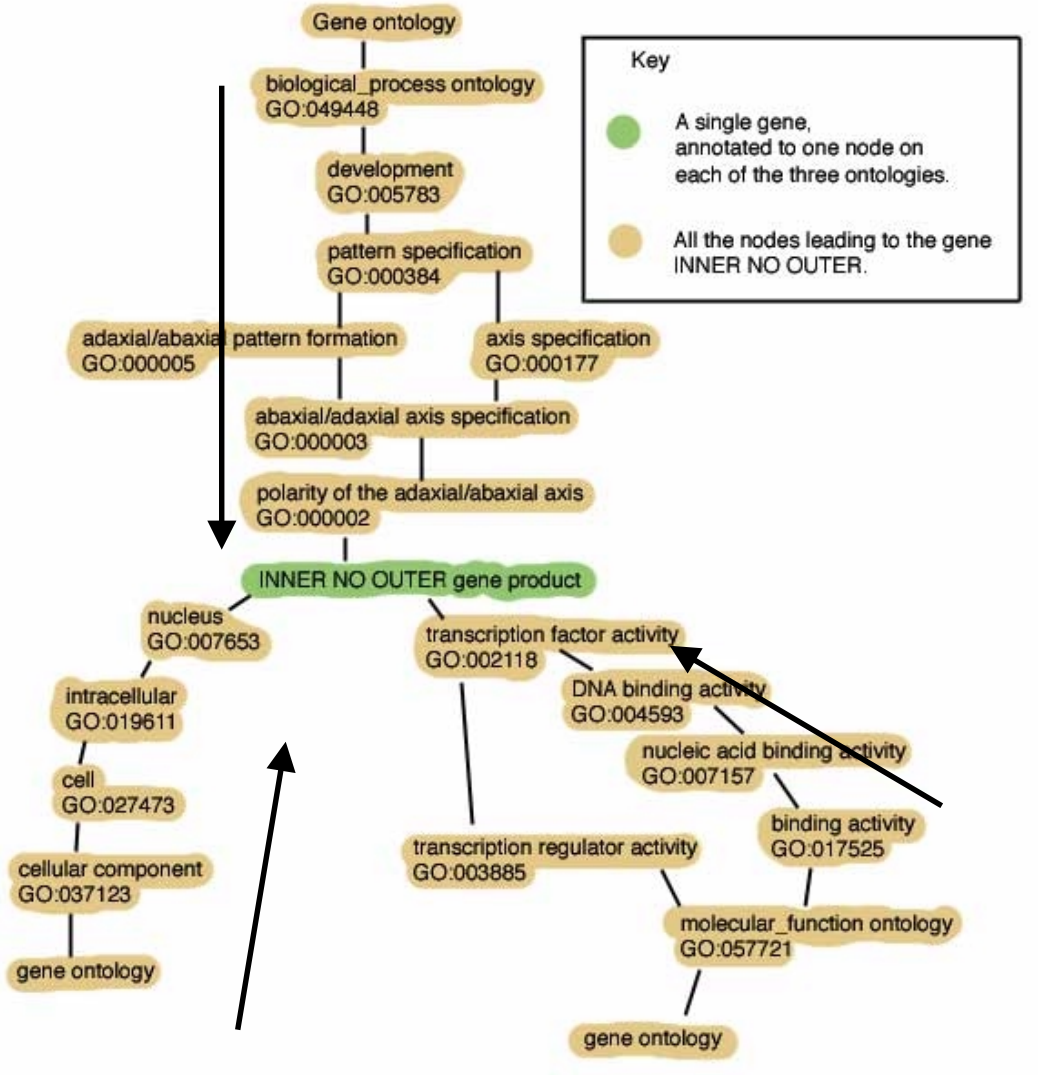
Protein cluster vs. supercomplex



Most supercomplex overlap with more than 1 protein cluster.
⇒ higher order organization of biological process

Overlap between protein clusters and supercomplex

Gene Ontology (GO)



Three separate ontologies: Biological Process, Molecular Function, Cellular Component.

Organized as a DAG describing gene products (proteins and functional RNA).

Makes the represented biological relationships computable.

Collaborative effort between major genome databases.

http://www.geneontology.org

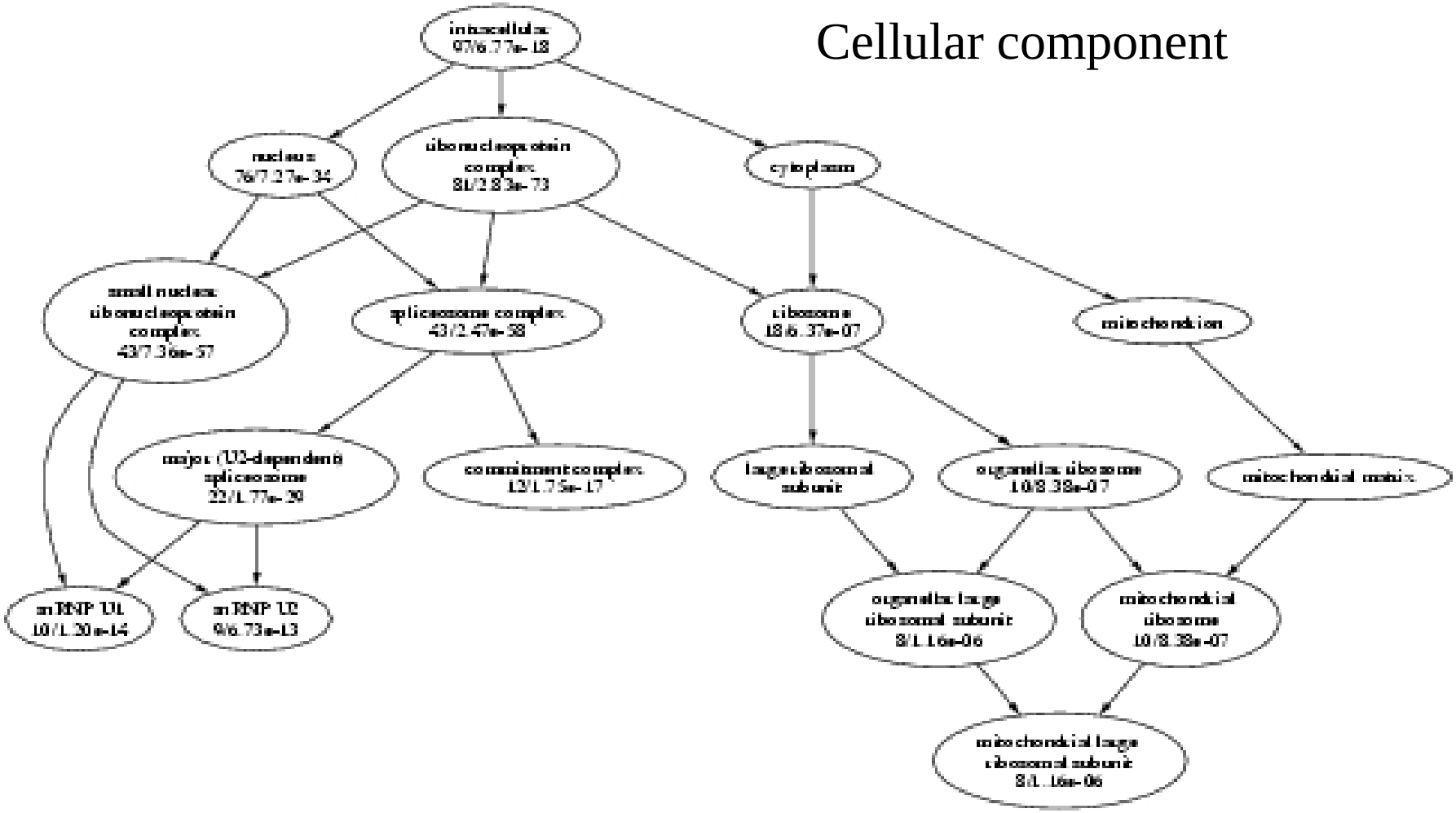
GO Category

Molecular function describes activities, such as catalytic or binding activities, at the molecular level (e.g. nucleic acid binding or exonuclease)

Biological process is accomplished by ordered assemblies of molecular functions (e.g. 'signal transduction' or 'nuclear export').

Cellular component is a component of cell that is part of a larger object, which may be an anatomical structure (e.g. nucleus) or a gene product group (e.g. spliceosome).

Annotation of protein cluster P_{28}

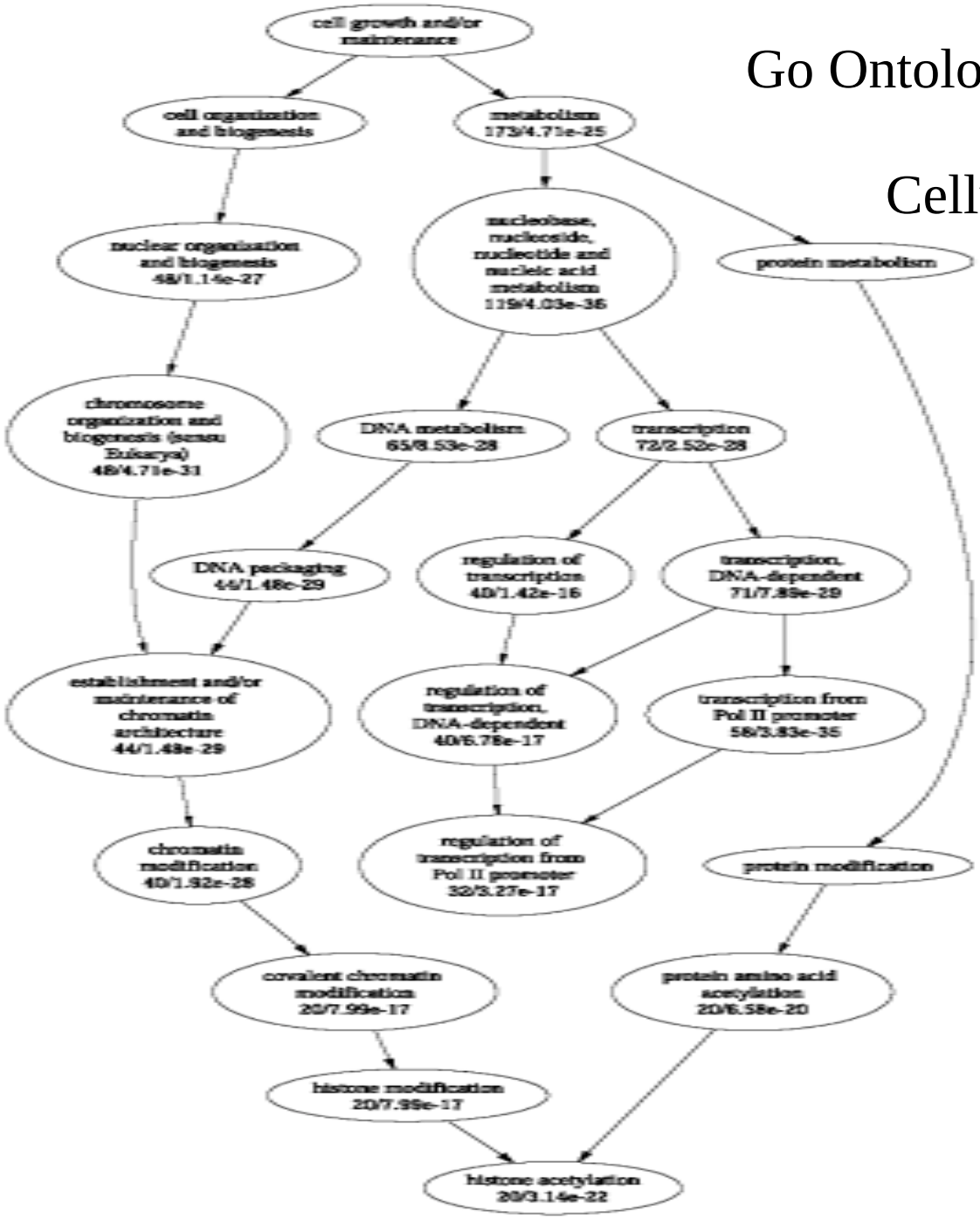


Biological Significance of Supercomplex C_{47}

MIPS Annotation Category	# ORFs in C_{47}	# ORFs matched
RNA Pol II holoenzyme	35	23
Kornberg's mediator	21	21
Other transcription	73	17
HAT A	15	14
TFIID	13	13
SAGA	14	13
Ada-Spt	14	13
TAFIIs	12	12
DNA repair	33	9
RSC	10	6
ADA	6	6
Replication fork	30	6
DNA mismatch repair	5	5
Cytoplasmic translation initiation	27	4
SAGA-like	5	4
Nucleotide excision repairosome	16	3
RNA Polymerase III	13	3
Replication factor A	3	3
Actin-associated motorproteins	7	3
MSH2/MSH3	3	3
Srb10p	4	3
NEF4	2	2
eIF4A	2	2
NuA4	2	2
Nuclear pore	24	2
Sir	2	2

Go Ontology for C_{47}

Cellular Process



Summary

- Study of protein interactions is important part of DOE Genome to Life program
- Genomic scale data from high-throughput experiments
- A new unified representation captures dual relationship between protein and protein-complex => naturally lead to protein – protein and complex – complex interactions
- MinMaxCut spectral clustering provides protein clusters and supercomplexes
- Protein cluster represents physiologically intact protein complex
- Important implications derived from clusters & supercomplexes
- Gene ontology (component) validates discovered protein clusters
- Gene ontology (process) validates supercomplex