



---

# 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

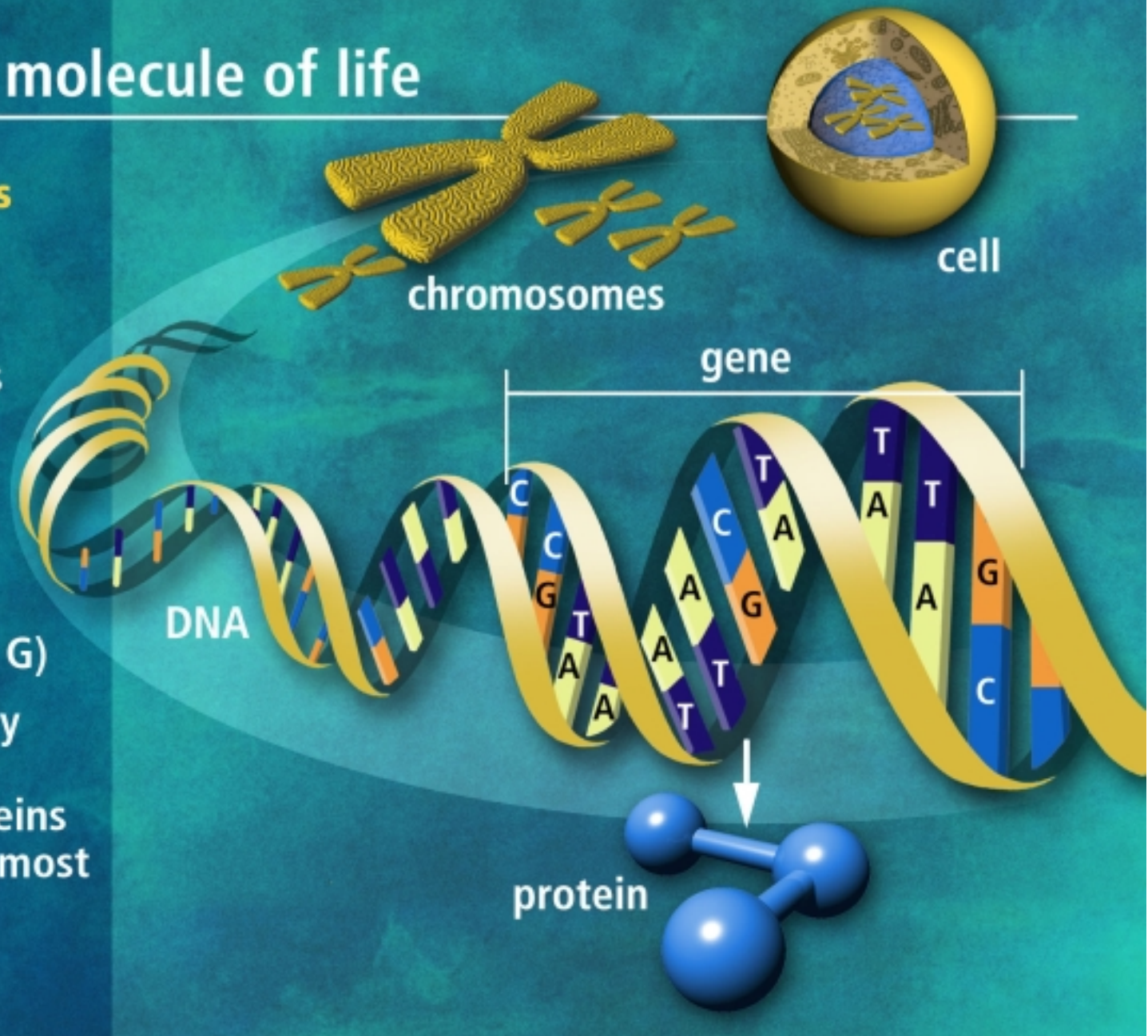
3/23/2004

# DNA the molecule of life

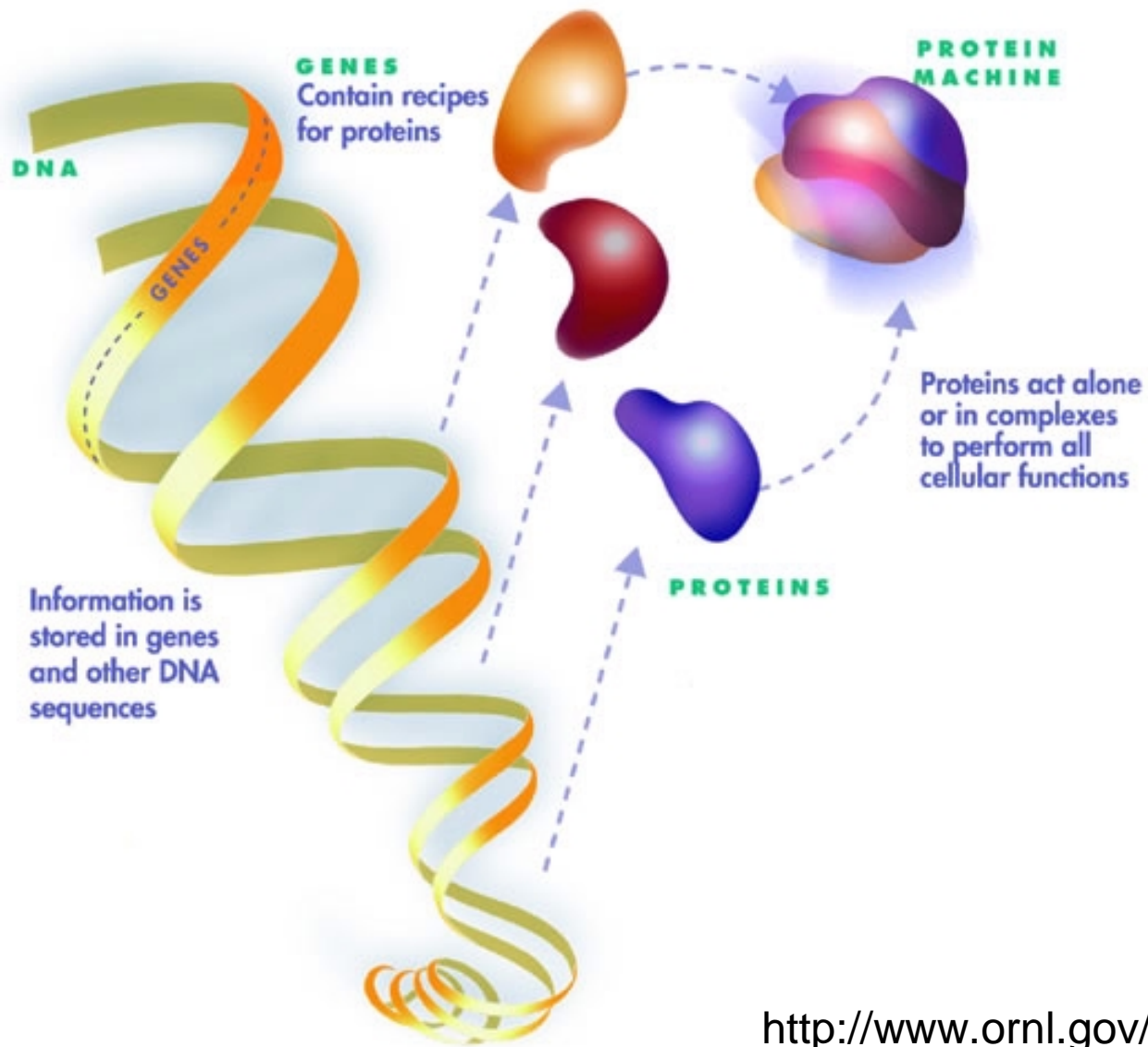
## Trillions of cells

Each cell:

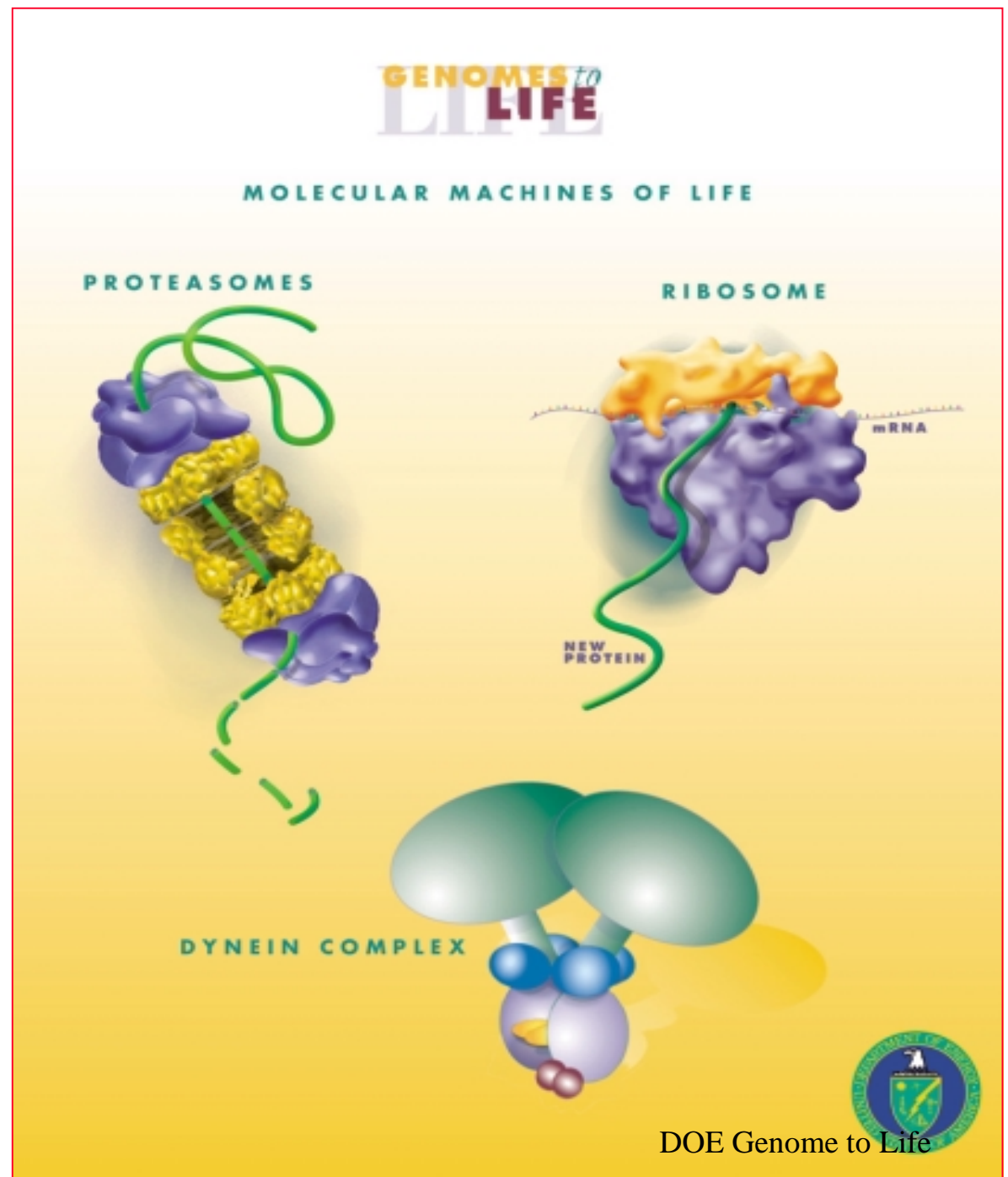
- 46 human chromosomes
- 2 meters of DNA
- 3 billion DNA subunits (the bases: A, T, C, G)
- Approximately 30,000 genes code for proteins that perform most life functions



## GENES, PROTEINS, AND MOLECULAR MACHINES



# 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



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

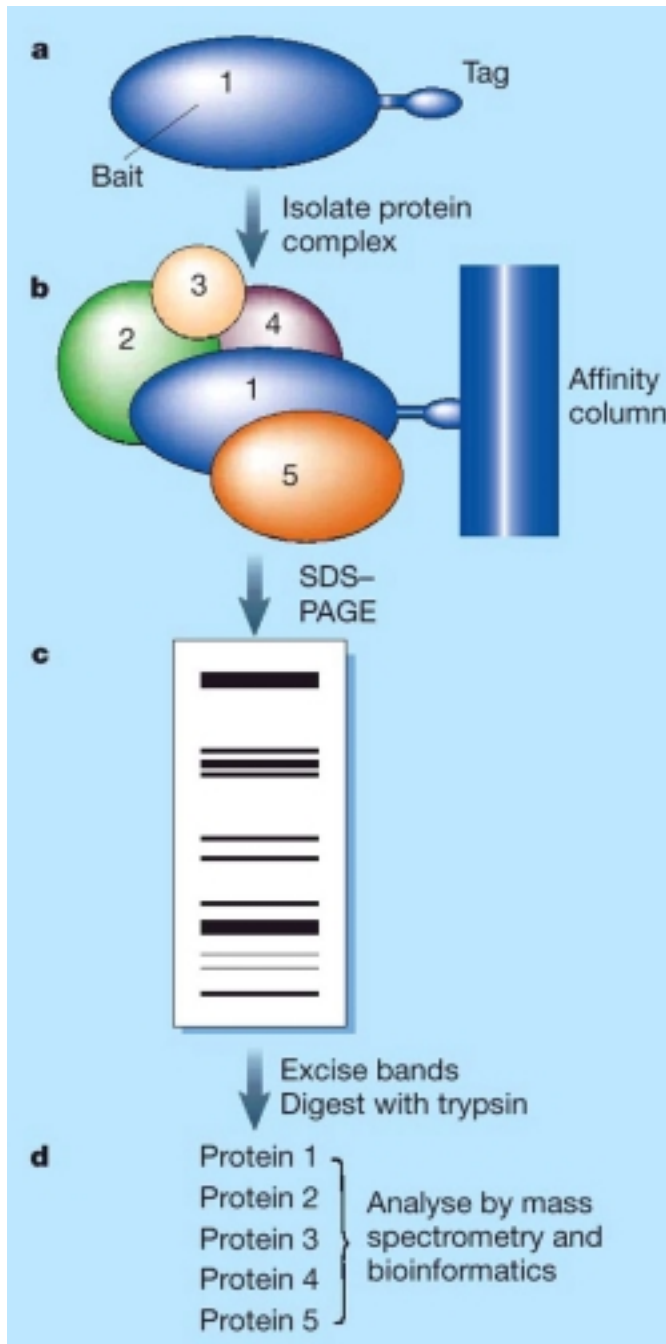
- **Two-hybrid** dataset by Uetz *et al* 2000 (the first comprehensive study in yeast)
- **Two-hybrid** dataset by Ito *et al* 2001 (broad coverage in yeast)
- **HMS-PCI** dataset by Ho *et al* 2002
- **TAP-MS** dataset by Gavin *et al* 2002

**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

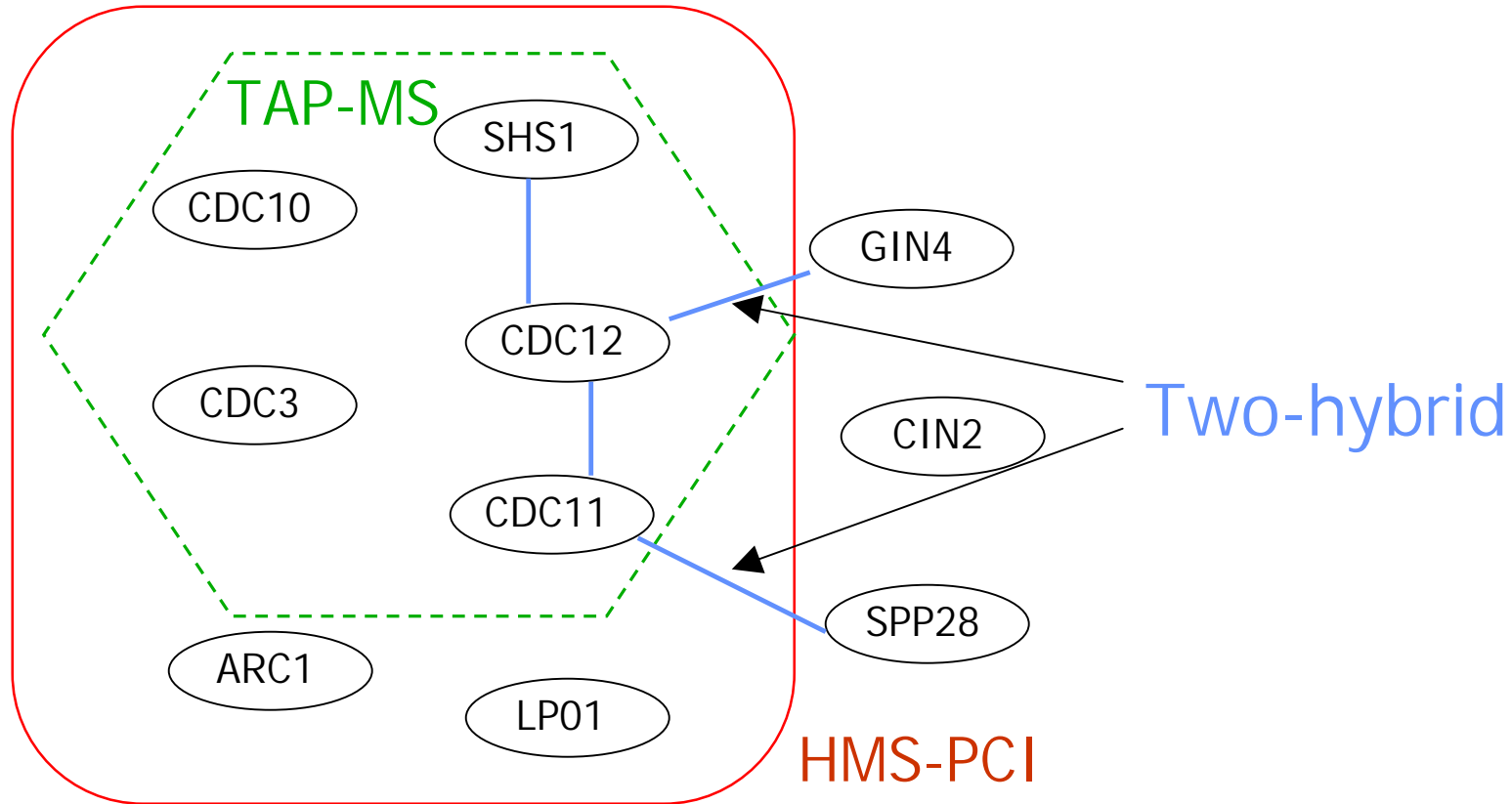


- Tandem-Affinity Purification coupled with Mass-Spectrometry (TAP-MS) determines the constituents of multi-protein complexes.

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.

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

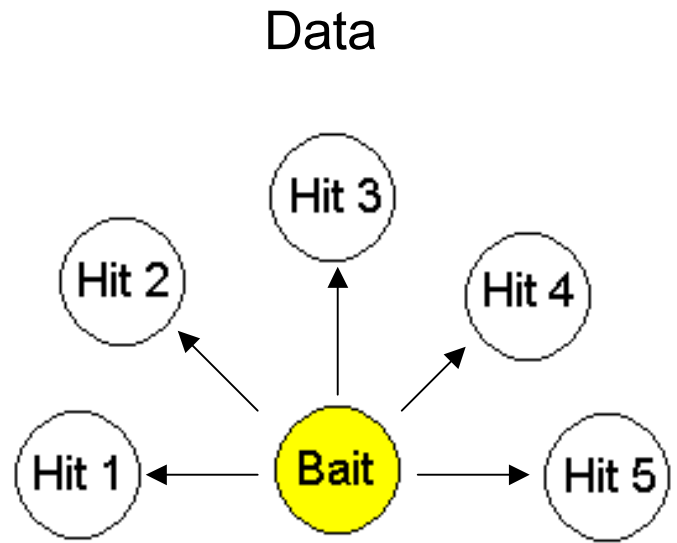
Copied from Salwinski and Eisenberg, 2003

# Protein interaction databases

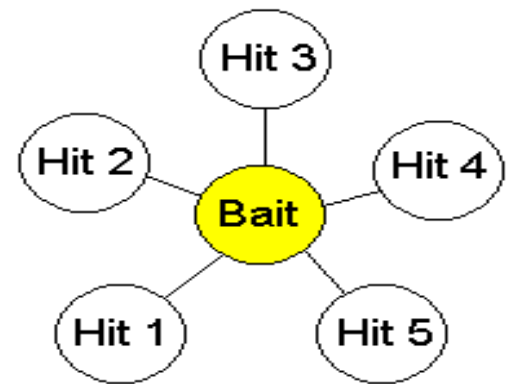


Database	URL
DIP	<a href="http://dip.doe-mbi.ucla.edu">dip.doe-mbi.ucla.edu</a>
MIPS	<a href="http://mips.gsf.de">mips.gsf.de</a>
BIND	<a href="http://www.bind.ca">www.bind.ca</a>
YPD	<a href="http://www.proteome.com/YPDhome.html">www.proteome.com/YPDhome.html</a>
The GRID	<a href="http://biodata.mshri.on.ca/grid/servlet/index">biodata.mshri.on.ca/grid/servlet/i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

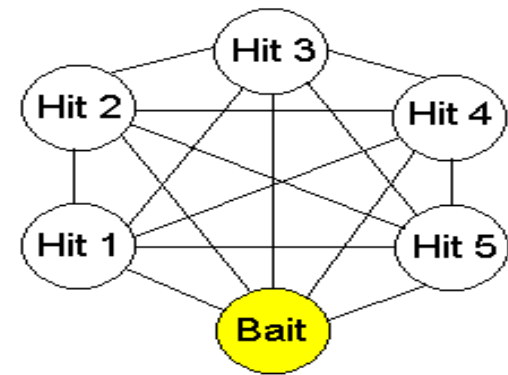
# Previous attempts to analyze TAP-MS data used simplified models



Spoke Model



Matrix Model



- find  $k$ -cores (Bader and Hogue, 2002)
- find cliques (Spirin and Mirny, 2003)
- Hypergraph –  $k$ -core (Pothén, 2003)





## 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

**Un-weighted** interaction strength  
- oversimplified

Focus only on **protein – protein interactions**

**K-core, clique**

## Our Model

**Weighted** interaction strength  
-more realistic

### **Unified representation**

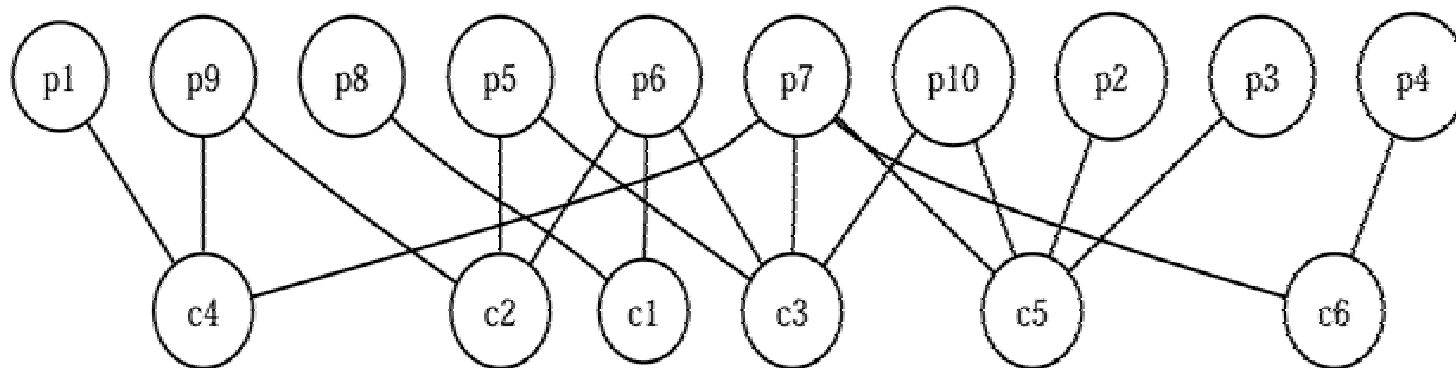
from protein complex data  
to derive

**protein – protein interactions**

**complex – complex network**

**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:

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

Interaction strength of protein complex – protein complex network:

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



## 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  $\Rightarrow$  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  $\Rightarrow$  highly connected

# Spectral clustering p-p and c-c network



**MinMaxCut** spectral clustering method:

Minimize similarity between clusters,  
Maximize similarity within cluster

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

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

Minimizing  $J_{\text{MMC}}$  leads to

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

and the solution is given by

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

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

# Biological usefulness of Clusters

---



Protein Cluster from protein-protein interaction network:

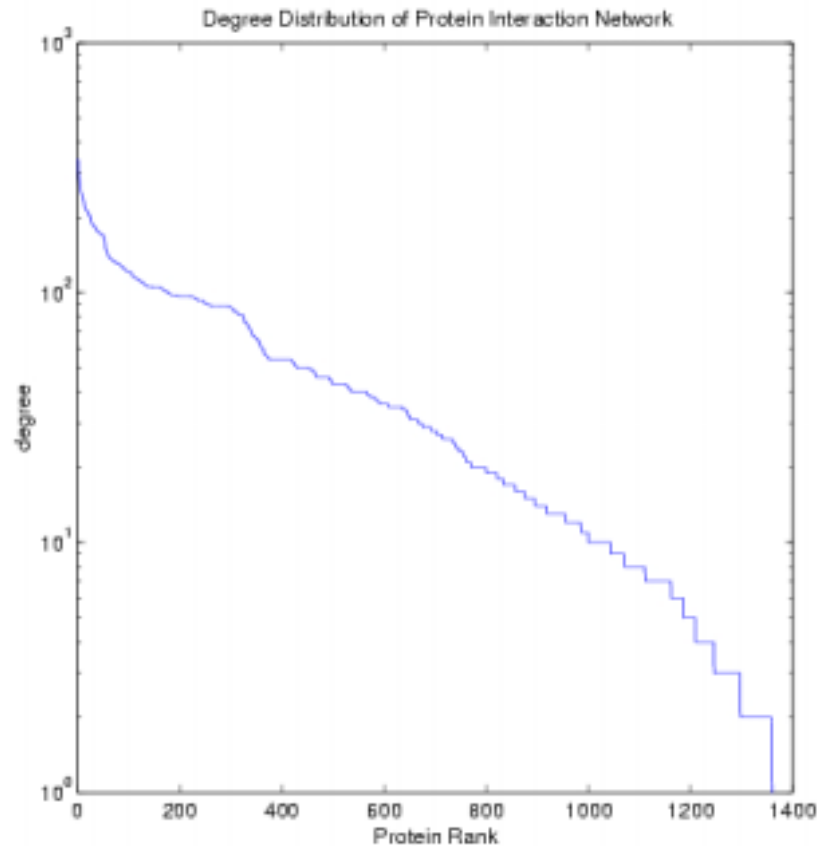
1. Assign annotations (functions) to uncharacterized proteins.
2. Predict possible functions for their orthologs in other species.
3. Predict biologically relevant modules carrying out cellular functions

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 -----

Cluster 28	Complex 128	Complex 129	Complex 166	Complex 168	Complex 169	Complex 161
YIL229C	YIL229C	YML117W				YML028C
YML117W						YLF424W
YML028C		YML214C	YLF424W	YLF424W	YLF424W	
YLF424W	YLF424W					
YML214C			Yju2			
Yju2						
YJF084W				YJF084W	YJF084W	YJF084W
YHP159C					YHP159C	YHP159C
Yhc1				Yhc1	Yhc1	Yhc1
YLF0276W			YLF0276W			YLF0276W
YOL128C	YOL128C		YOL128C			YOL128C
YOL209C			YOL209C	YOL209C	YOL209C	YOL209C
YOL179C		YOL179C				
YCF043W			YCF043W			YCF043W
Tss4						Tss4
TF48.52		TF48.52			TF48.52	
TF48.51		TF48.51			TF48.51	
Byp1			Byp1			Byp1
Bba1		Bba1	Bba1	Bba1	Bba1	Bba1
Bba2	Bba2			Bba2	Bba2	Bba2
Bba3						Bba3
Bba4						Bba4
Bba5						Bba5
Bba6						Bba6
Bba7						Bba7
Bba8						Bba8
Bba9	Bba9	Bba9		Bba9	Bba9	Bba9
Bba10						Bba10
Bba11						Bba11
Bba12						Bba12
Bba13						Bba13
Bba14						Bba14
Bba15						Bba15
Bba16						Bba16
Bba17						Bba17
Bba18						Bba18
Bba19						Bba19
Bba20						Bba20
Bba21						Bba21
Bba22						Bba22
Bba23						Bba23
Bba24						Bba24
Bba25						Bba25
Bba26						Bba26
Bba27						Bba27
Bba28						Bba28
Bba29						Bba29
Bba30						Bba30
Bba31						Bba31
Bba32						Bba32
Bba33						Bba33
Bba34						Bba34
Bba35						Bba35
Bba36						Bba36
Bba37						Bba37
Bba38						Bba38
Bba39						Bba39
Bba40						Bba40
Bba41						Bba41
Bba42						Bba42
Bba43						Bba43
Bba44						Bba44
Bba45						Bba45
Bba46						Bba46
Bba47						Bba47
Bba48						Bba48
Bba49						Bba49
Bba50						Bba50
Bba51						Bba51
Bba52						Bba52
Bba53						Bba53
Bba54						Bba54
Bba55						Bba55
Bba56						Bba56
Bba57						Bba57
Bba58						Bba58
Bba59						Bba59
Bba60						Bba60
Bba61						Bba61
Bba62						Bba62
Bba63						Bba63
Bba64						Bba64
Bba65						Bba65
Bba66						Bba66
Bba67						Bba67
Bba68						Bba68
Bba69						Bba69
Bba70						Bba70
Bba71						Bba71
Bba72						Bba72
Bba73						Bba73
Bba74						Bba74
Bba75						Bba75
Bba76						Bba76
Bba77						Bba77
Bba78						Bba78
Bba79						Bba79
Bba80						Bba80
Bba81						Bba81
Bba82						Bba82
Bba83						Bba83
Bba84						Bba84
Bba85						Bba85
Bba86						Bba86
Bba87						Bba87
Bba88						Bba88
Bba89						Bba89
Bba90						Bba90
Bba91						Bba91
Bba92						Bba92
Bba93						Bba93
Bba94						Bba94
Bba95						Bba95
Bba96						Bba96
Bba97						Bba97
Bba98						Bba98
Bba99						Bba99
Bba100						Bba100
Bba101						Bba101
Bba102						Bba102
Bba103						Bba103
Bba104						Bba104
Bba105						Bba105
Bba106						Bba106
Bba107						Bba107
Bba108						Bba108
Bba109						Bba109
Bba110						Bba110
Bba111						Bba111
Bba112						Bba112
Bba113						Bba113
Bba114						Bba114
Bba115						Bba115
Bba116						Bba116
Bba117						Bba117
Bba118						Bba118
Bba119						Bba119
Bba120						Bba120
Bba121						Bba121
Bba122						Bba122
Bba123						Bba123
Bba124						Bba124
Bba125						Bba125
Bba126						Bba126
Bba127						Bba127
Bba128						Bba128
Bba129						Bba129
Bba130						Bba130
Bba131						Bba131
Bba132						Bba132
Bba133						Bba133
Bba134						Bba134
Bba135						Bba135
Bba136						Bba136
Bba137						Bba137
Bba138						Bba138
Bba139						Bba139
Bba140						Bba140
Bba141						Bba141
Bba142						Bba142
Bba143						Bba143
Bba144						Bba144
Bba145						Bba145
Bba146						Bba146
Bba147						Bba147
Bba148						Bba148
Bba149						Bba149
Bba150						Bba150
Bba151						Bba151
Bba152						Bba152
Bba153						Bba153
Bba154						Bba154
Bba155						Bba155
Bba156						Bba156
Bba157						Bba157
Bba158						Bba158
Bba159						Bba159
Bba160						Bba160
Bba161						Bba161
Bba162						Bba162
Bba163						Bba163
Bba164						Bba164
Bba165						Bba165
Bba166						Bba166
Bba167						Bba167
Bba168						Bba168
Bba169						Bba169
Bba170						Bba170
Bba171						Bba171
Bba172						Bba172
Bba173						Bba173
Bba174						Bba174
Bba175						Bba175
Bba176						Bba176
Bba177						Bba177
Bba178						Bba178
Bba179						Bba179
Bba180						Bba180
Bba181						Bba181
Bba182						Bba182
Bba183						Bba183
Bba184						Bba184
Bba185						Bba185
Bba186						Bba186
Bba187						Bba187
Bba188						Bba188
Bba189						Bba189
Bba190						Bba190
Bba191						Bba191
Bba192						Bba192
Bba193						Bba193
Bba194						Bba194
Bba195						Bba195
Bba196						Bba196
Bba197						Bba197
Bba198						Bba198
Bba199						Bba199
Bba200						Bba200
Bba201						Bba201
Bba202						Bba202
Bba203						Bba203
Bba204						Bba204
Bba205						Bba205
Bba206						Bba206
Bba207						Bba207
Bba208						Bba208
Bba209						Bba209
Bba210						Bba210
Bba211						Bba211
Bba212						Bba212
Bba213						Bba213
Bba214						Bba214
Bba215						Bba215
Bba216						Bba216
Bba217						

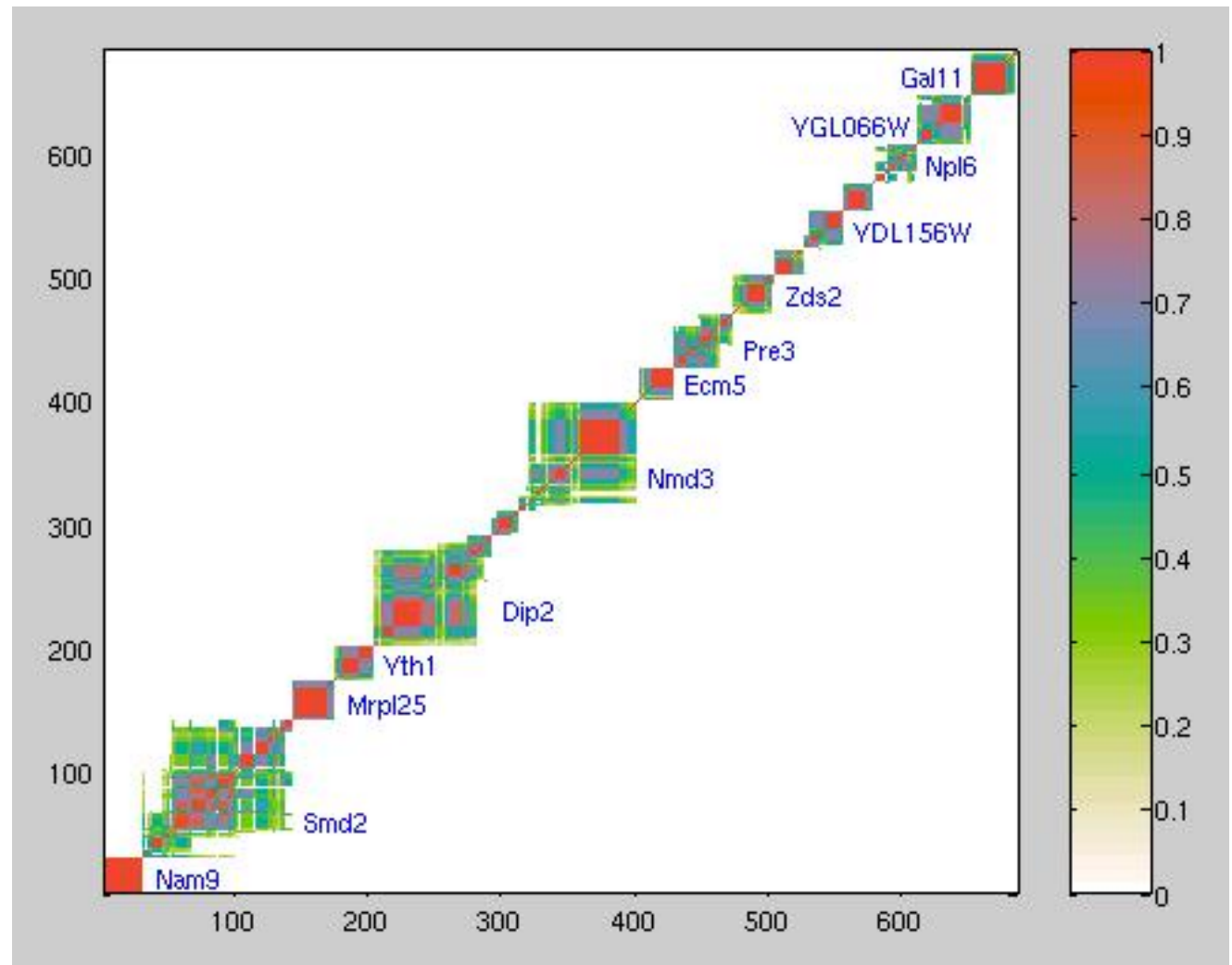
# Predicted clusters of protein-protein network



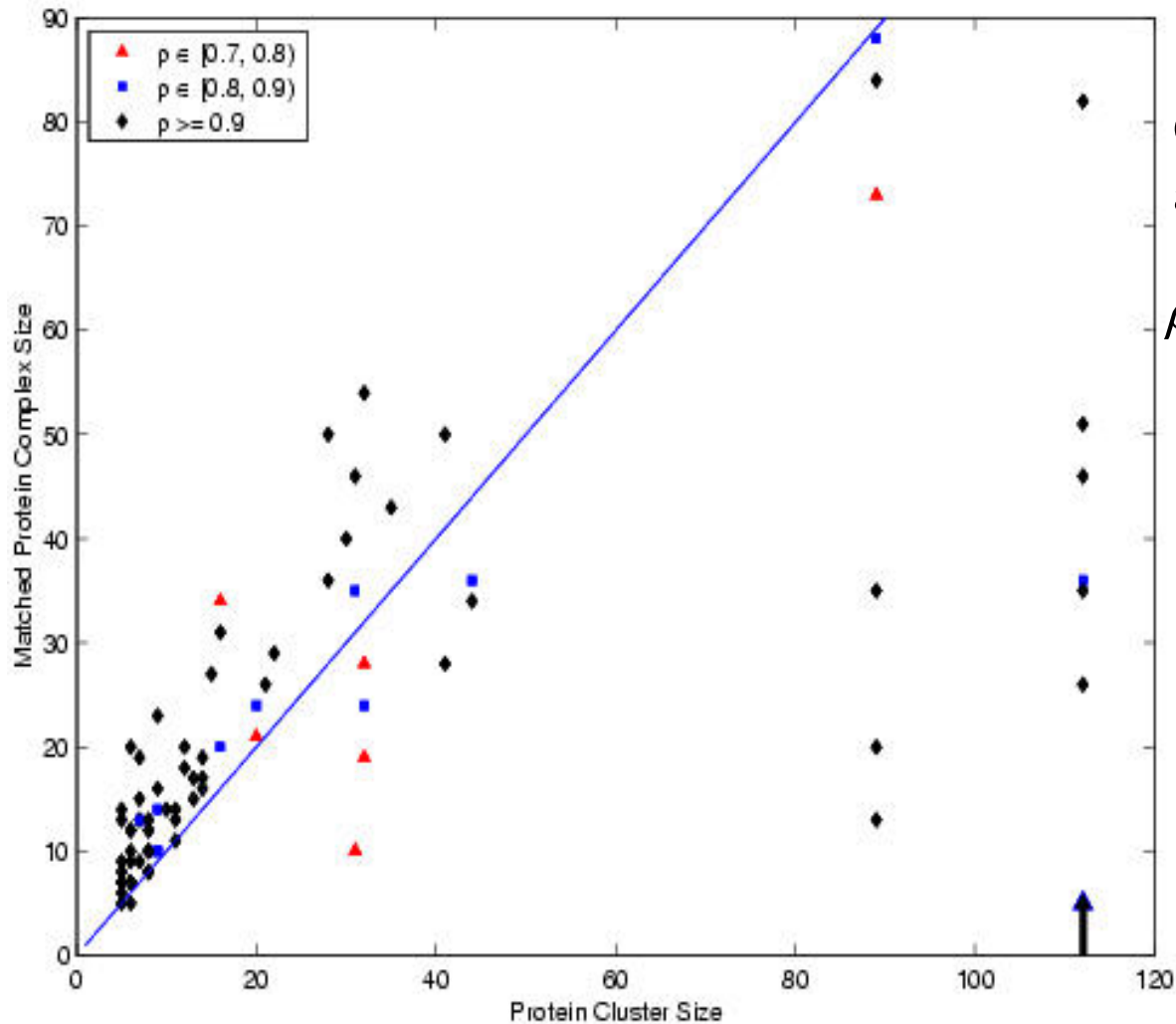
Interaction strength between gene varies, more realistic

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

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

## 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

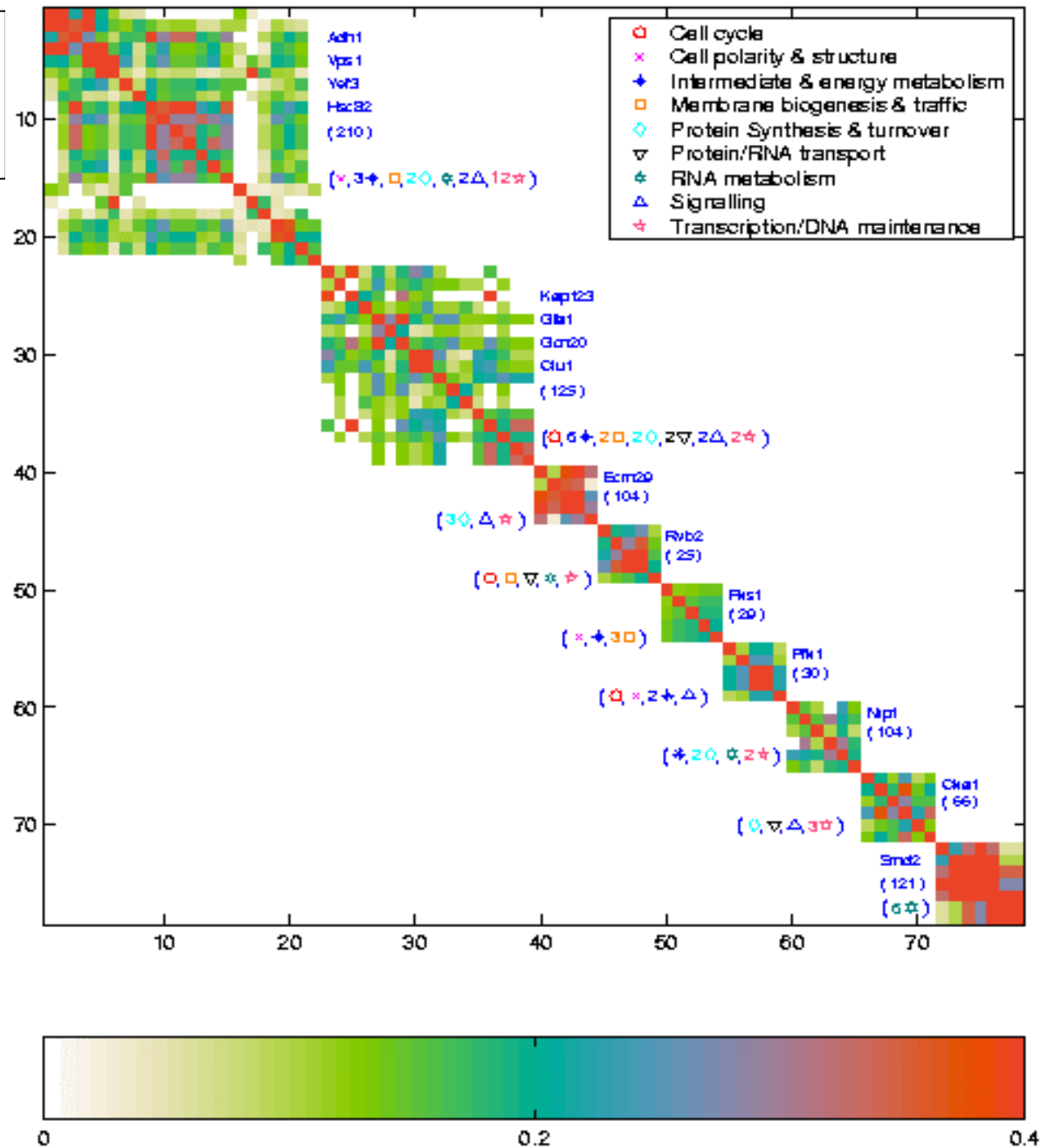
# Protein complex - Protein complex

Higher order organization

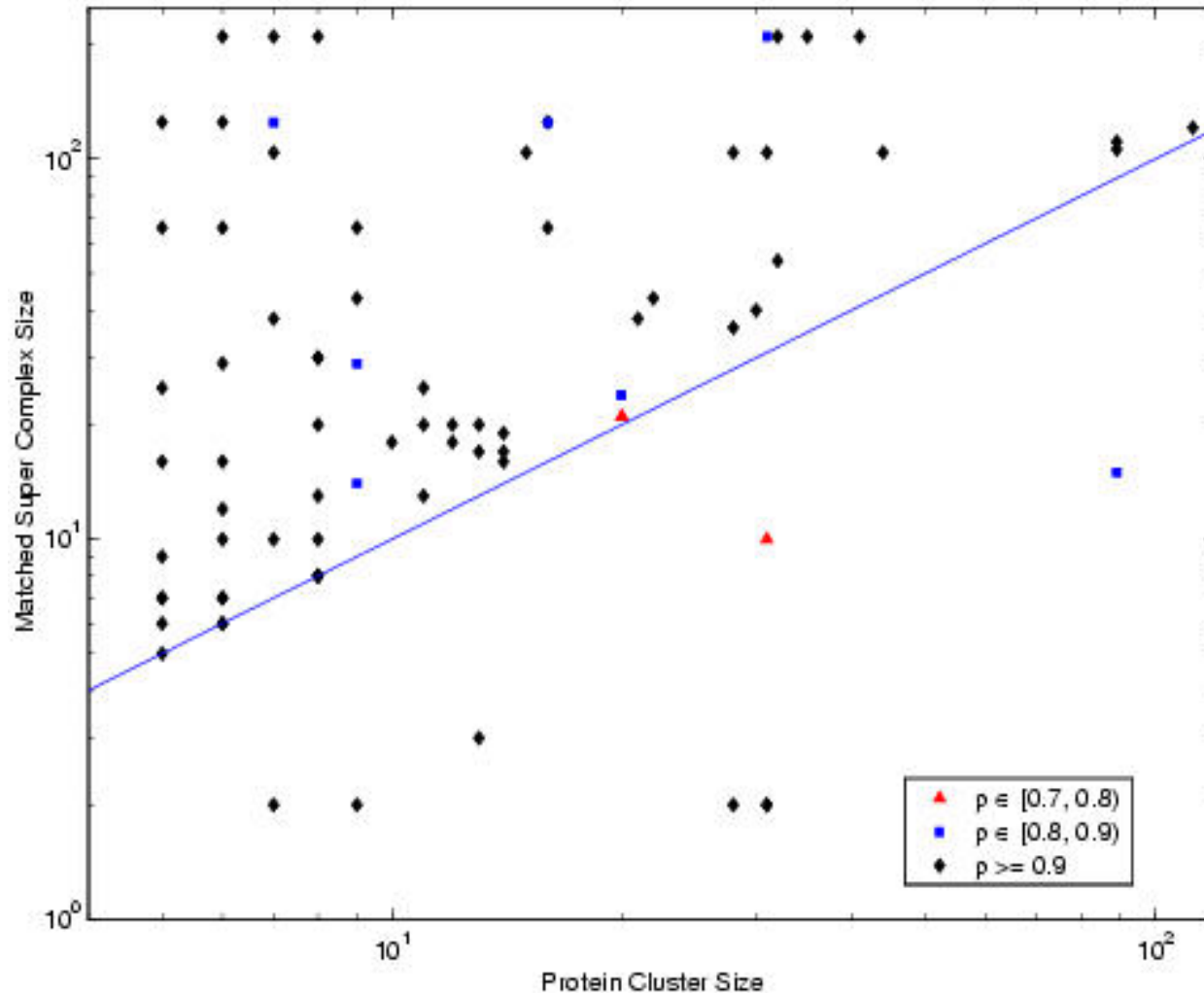
Supercomplex:

Cellular Process information more apparent:

Chromatin dynamics, transcription regulation, cell cycle control, biogenesis



# Protein cluster vs. supercomplex

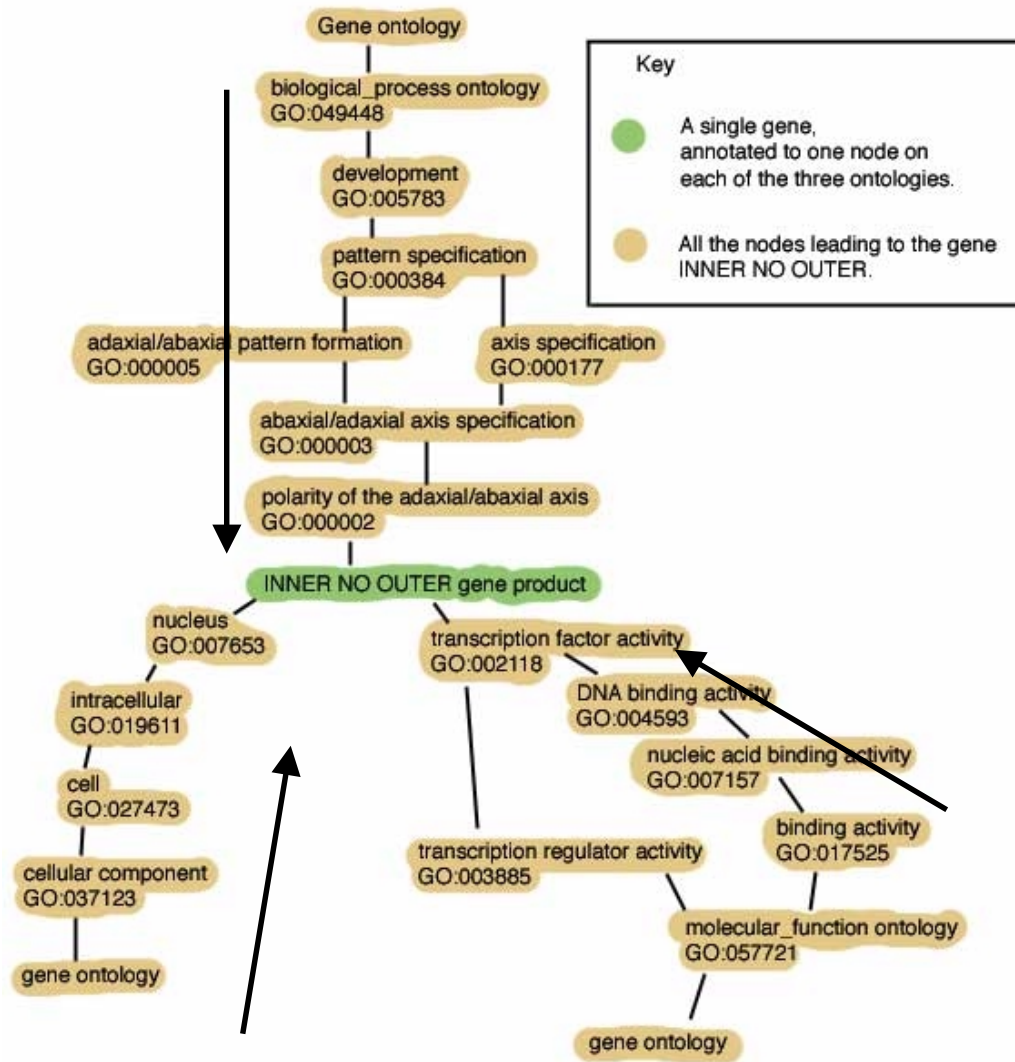


Most supercomplex  
overlap with more than  
1 protein cluster.  
 $\Rightarrow$  higher order  
organization of  
biological process

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)

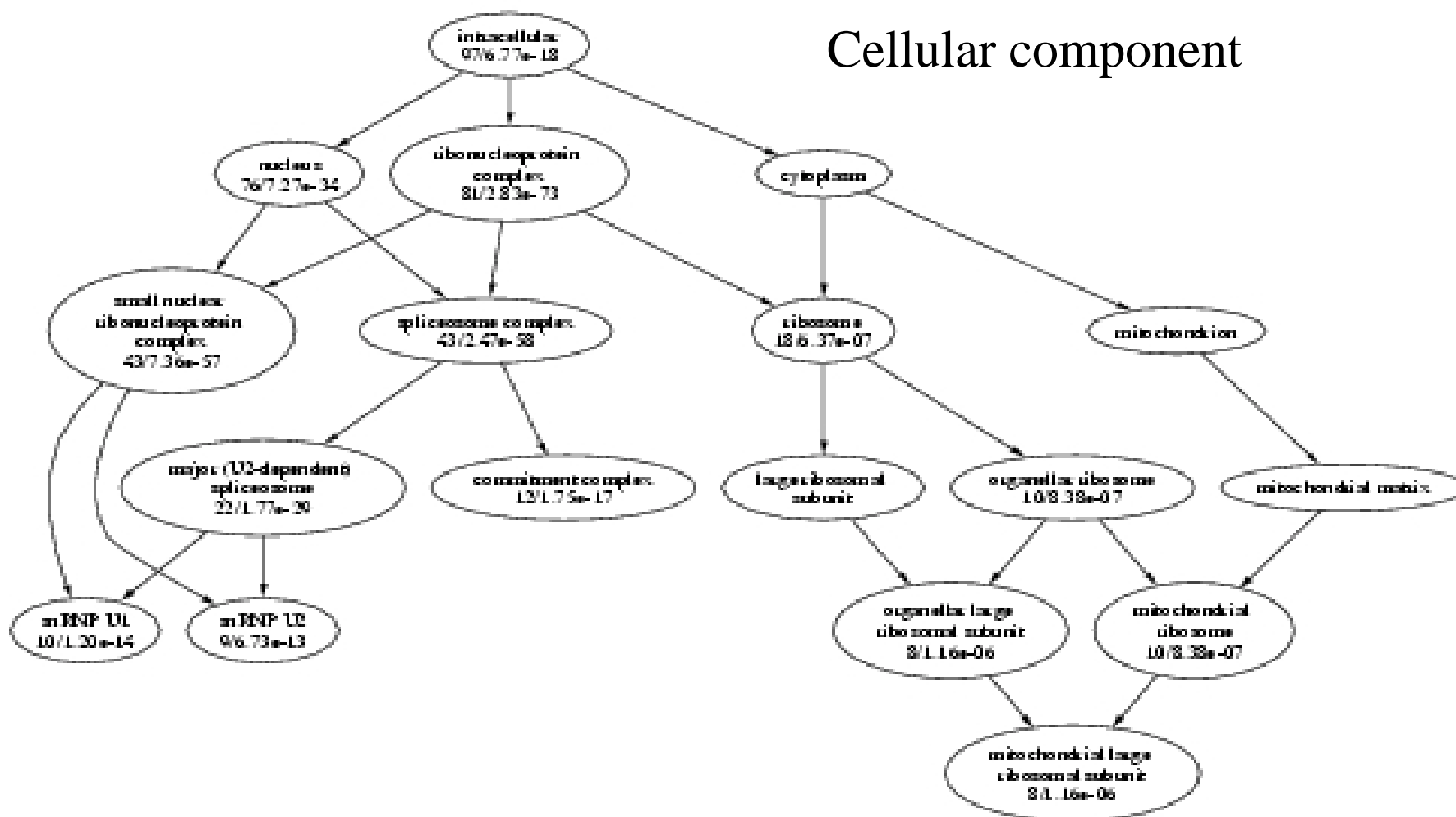
**Biological process** is accomplished by ordered assemblies of molecular functions (e.g. 'signal transduction' or 'nuclear export').

**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<sub>28</sub>



## Cellular component

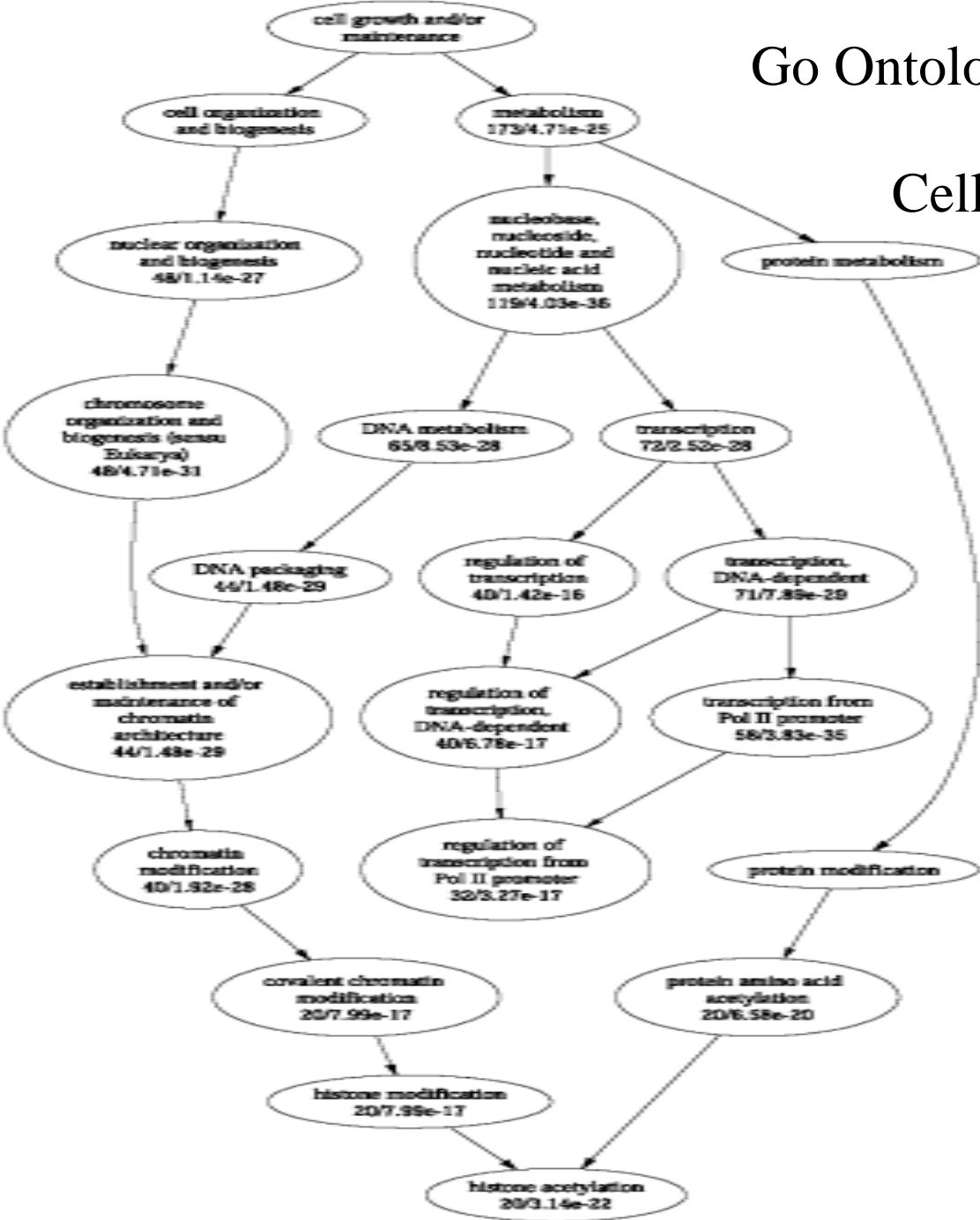


# Biological Significance of Supercomplex C<sub>47</sub>

MIPS Annotation Category	# ORFs in C <sub>47</sub>	# 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