Making Sense of High throughput Protein-Protein Interaction Data

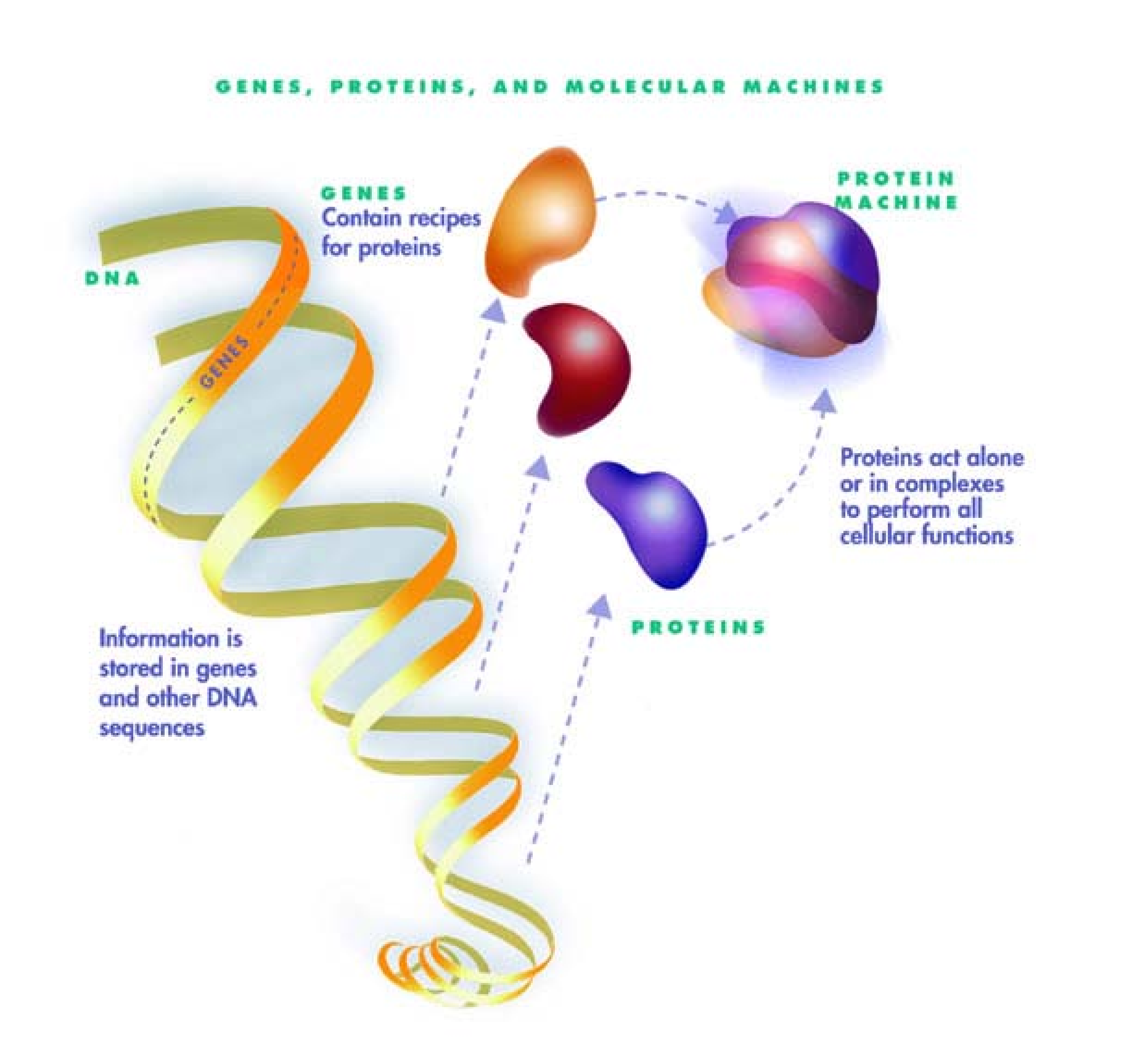
A Graph Theoretic Algorithm for AP-MS Data

Denise Scholtens

Robert Gentleman

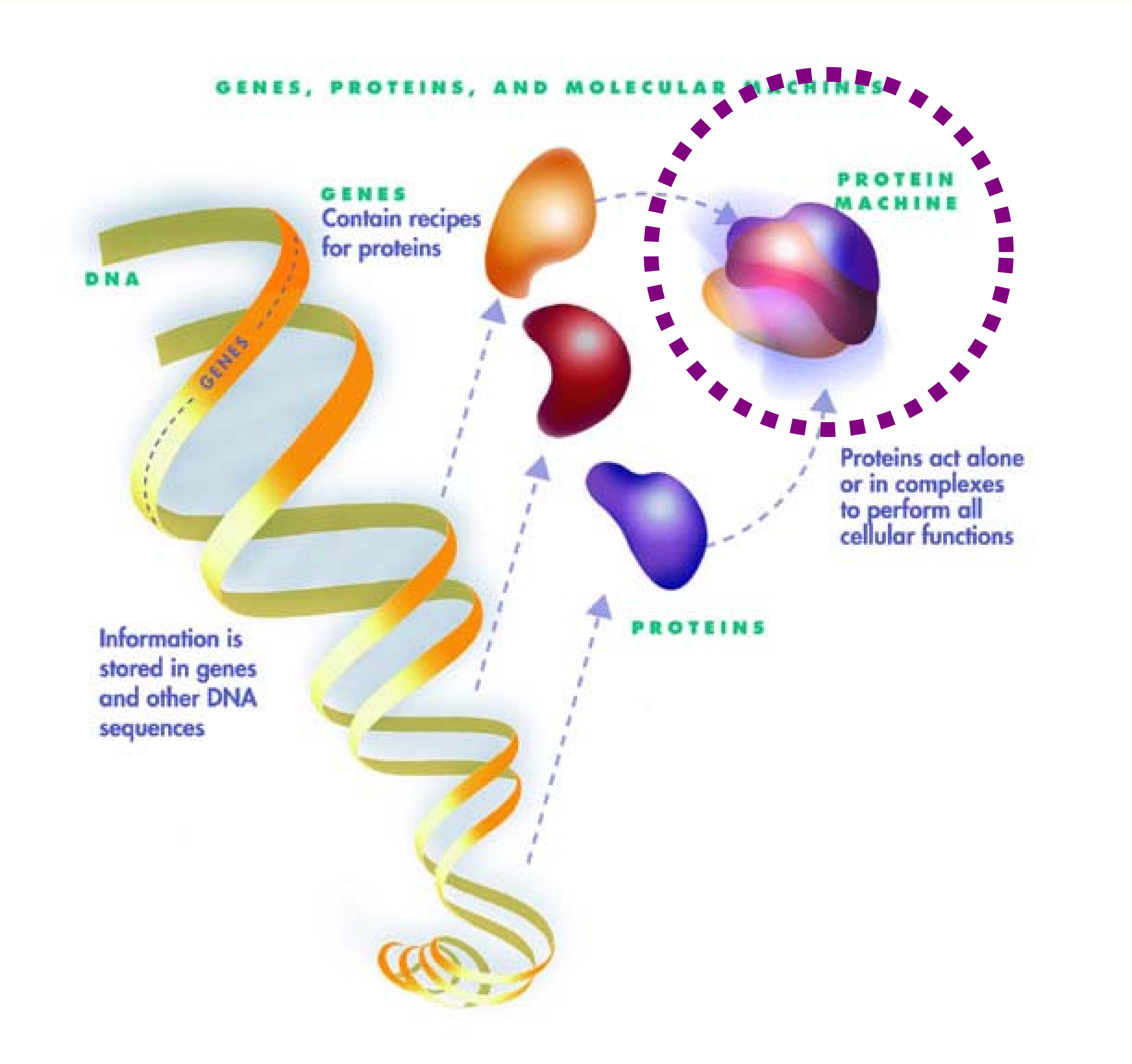
Auckland

Dec. 2003


GENES, PROTEINS, AND MOLECULAR MACHINES
DNA
GENES
GENES
Contain recipes for proteins
PROTEIN MACHINE
Proteins act alone or in complexes to perform all cellular functions
PROTEINS
Information is stored in genes and other DNA sequences

Which proteins are these?

Graphic courtesy of:
U.S. Department of Energy Human Genome Program
http://www.ornl.gov/hgmis


GENES, PROTEINS, AND MOLECULAR MACHINES
DNA
GENES
GENES
Contain recipes for proteins
PROTEIN MACHINE
Proteins act alone or in complexes to perform all cellular functions
PROTEINS
Information is stored in genes and other DNA sequences

Which proteins are these?

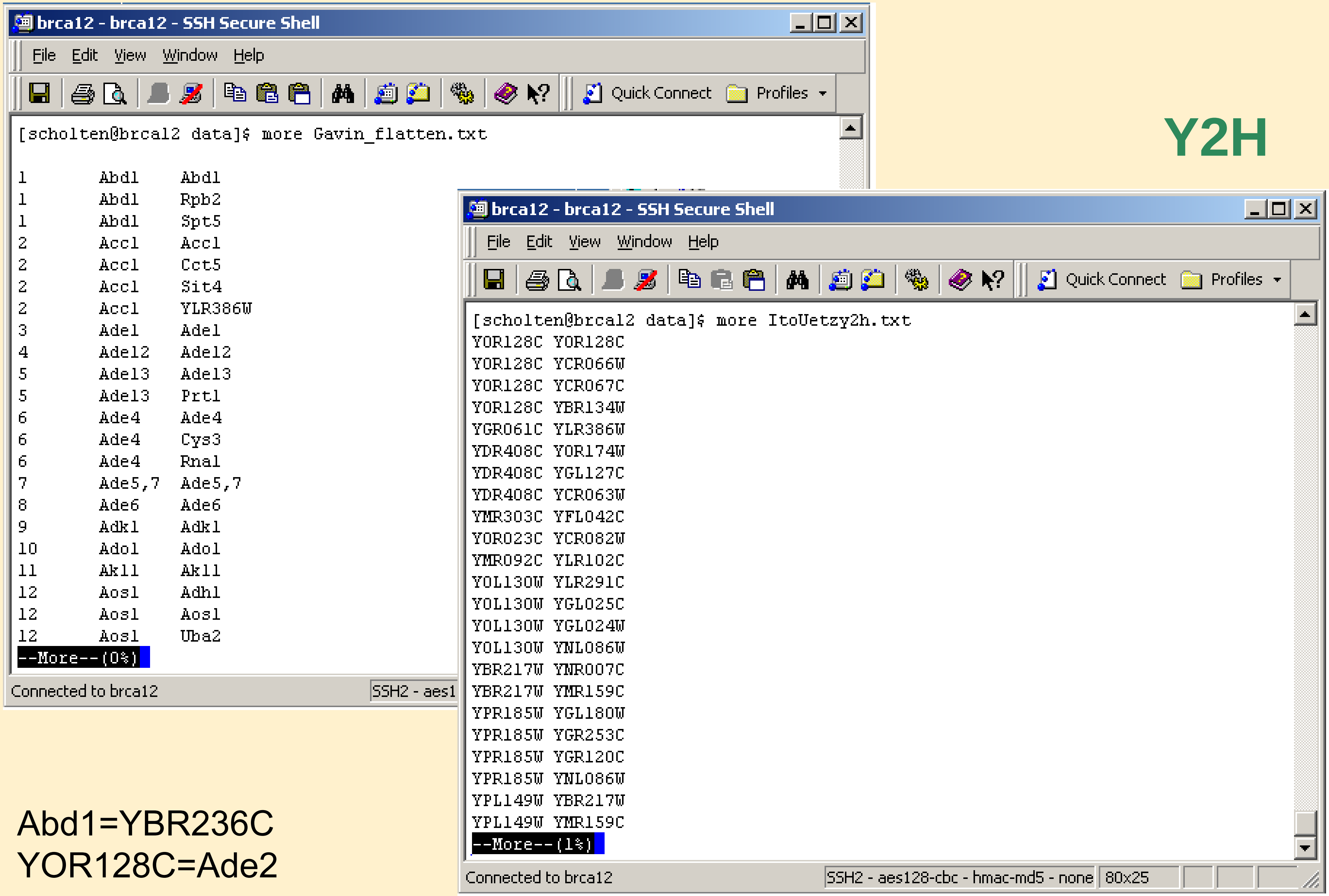
Graphic courtesy of:
U.S. Department of Energy Human Genome Program
http://www.ornl.gov/hgmis

Two Types of Data: Pairwise Protein Relationships

- **AP-MS (Affinity Purification - Mass Spectrometry)**
  - Measures ***Complex Comembership***
    - Gavin, et al. (Nature, 2002)
      - **TAP** : Tandem Affinity Purification
    - Ho, et al. (Nature, 2002)
      - **HMS-PCI**: High-throughput Mass Spectromic Protein Complex Identification
- **Y2H (Yeast Two Hybrid)**
  - Measures ***Physical Interactions***
    - Ito, et al. (PNAS, 1998)
    - Uetz, et al. (Nature, 2000)

AP-MS

Y2H

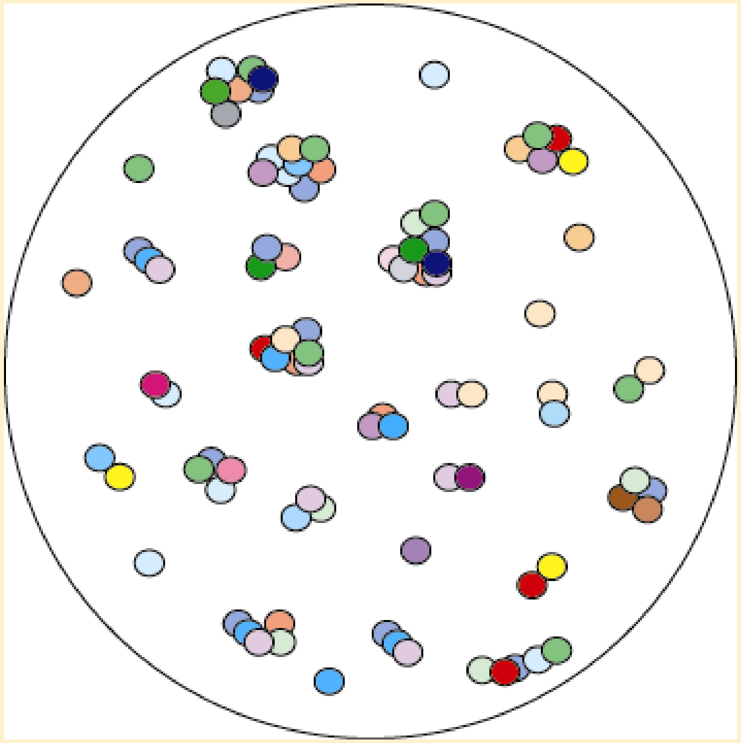

brca12 - brca12 - SSH Secure Shell
File Edit View Window Help
Quick Connect
Profiles
Connected to brca12
SSH2 - aes1

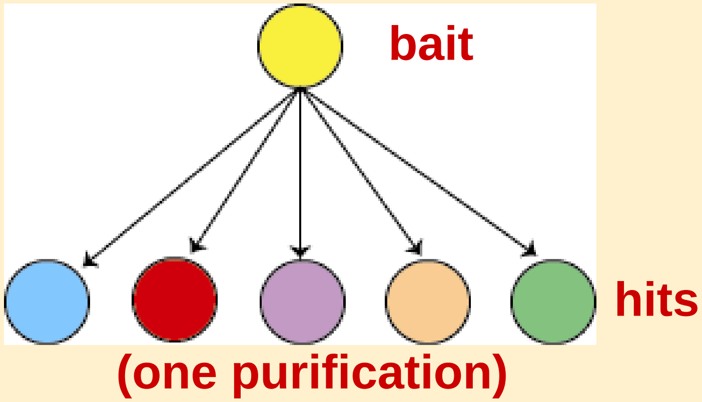
```
[scholten@brca12 data]$ more Gavin_flatten.txt

1       Abd1    Abd1
1       Abd1    Rpb2
1       Abd1    Spt5
2       Acc1    Acc1
2       Acc1    Cct5
2       Acc1    Sit4
2       Acc1    YLR386W
3       Ade1    Ade1
4       Ade12   Ade12
5       Ade13   Ade13
5       Ade13   Prt1
6       Ade4    Ade4
6       Ade4    Cys3
6       Ade4    Rna1
7       Ade5,7  Ade5,7
8       Ade6    Ade6
9       Adk1    Adk1
10      Ado1    Ado1
11      Akl1    Akl1
12      Aos1    Adh1
12      Aos1    Aos1
12      Aos1    Uba2
--More--(0%)
```

brca12 - brca12 - SSH Secure Shell
File Edit View Window Help
Quick Connect
Profiles
Connected to brca12
SSH2 - aes128-cbc - hmac-md5 - none
80x25

```
[scholten@brca12 data]$ more ItoUetzy2h.txt
YOR128C YOR128C
YOR128C YCR066W
YOR128C YCR067C
YOR128C YBR134W
YGR061C YLR386W
YDR408C YOR174W
YDR408C YGL127C
YDR408C YCR063W
YMR303C YFL042C
YOR023C YCR082W
YMR092C YLR102C
YOL130W YLR291C
YOL130W YGL025C
YOL130W YGL024W
YOL130W YNL086W
YBR217W YNR007C
YBR217W YMR159C
YPR185W YGL180W
YPR185W YGR253C
YPR185W YGR120C
YPR185W YNL086W
YPL149W YBR217W
YPL149W YMR159C
--More--(1%)
```

Abd1=YBR236C
YOR128C=Ade2

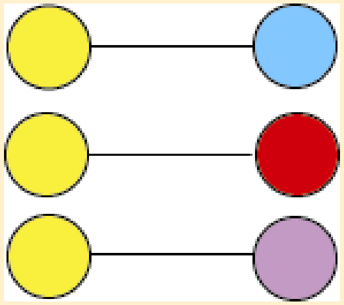




bait

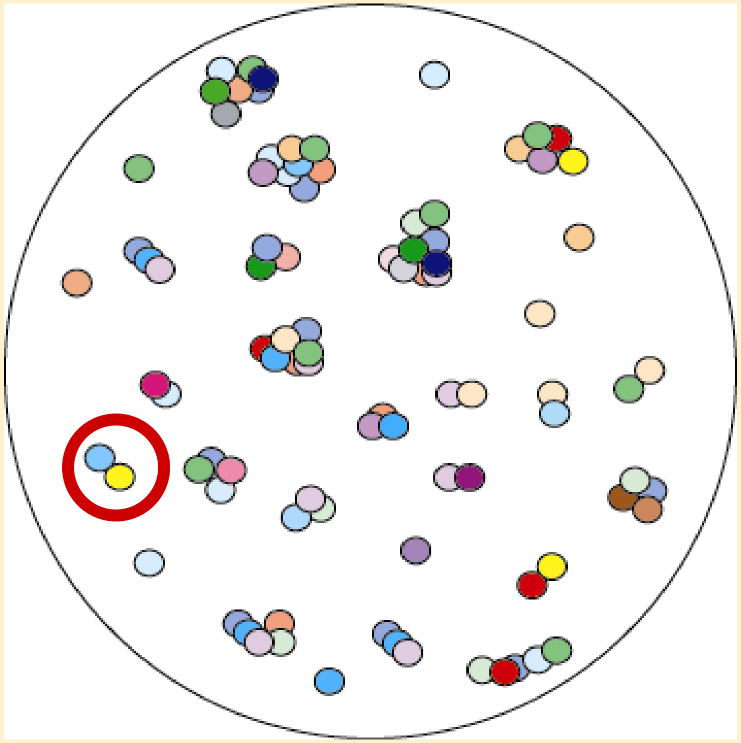
hits

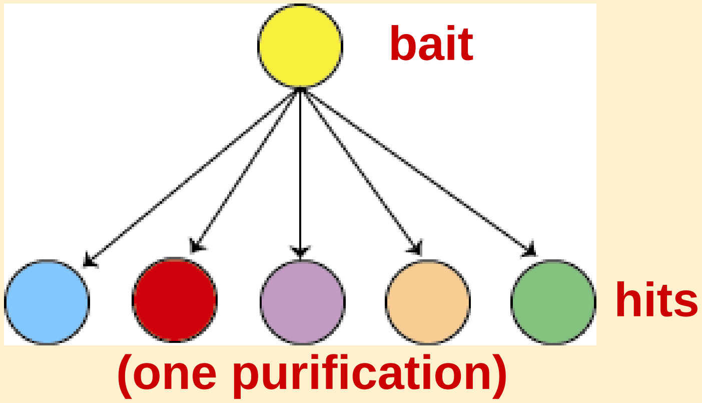
(one purification)

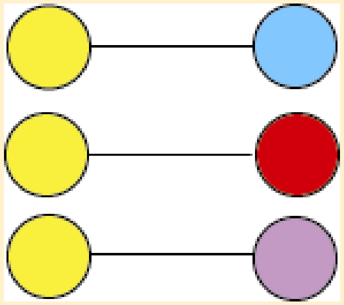


Using a ***bait*** protein, **AP-MS** technology finds ***hit*** proteins that are comembers of ***at least one*** complex with the bait.

Y2H technology finds pairs of ***physically interacting*** proteins.

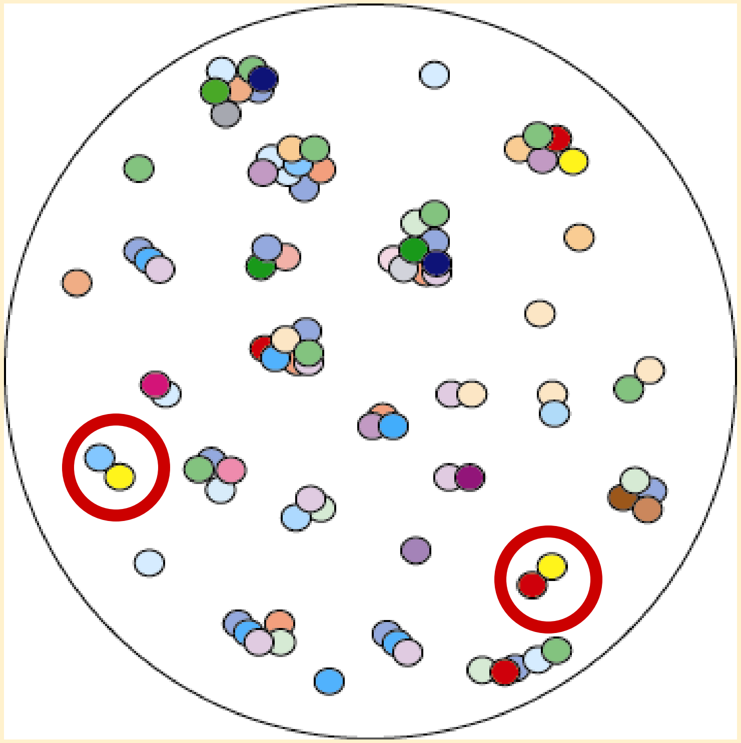


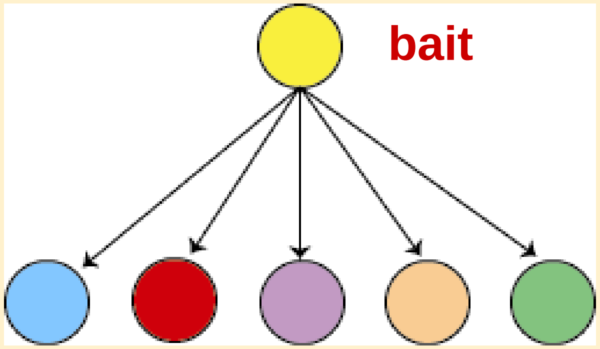




Using a ***bait*** protein, **AP-MS** technology finds ***hit*** proteins that are comembers of ***at least one*** complex with the bait.

Y2H technology finds pairs of ***physically interacting*** proteins.

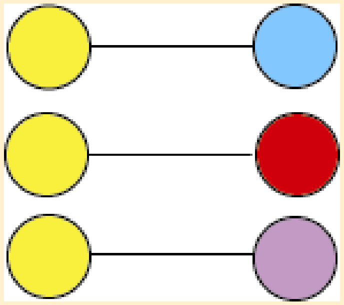




bait

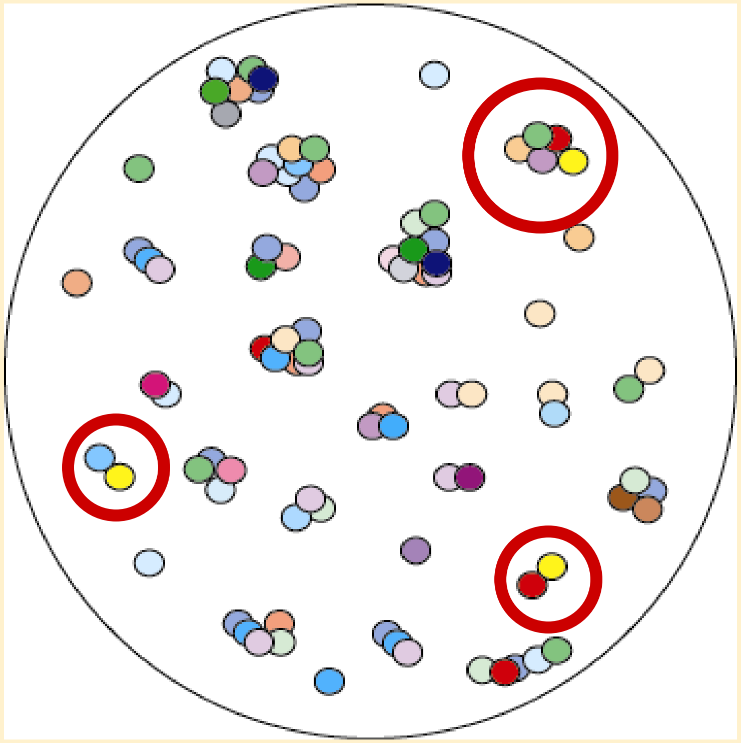
hits

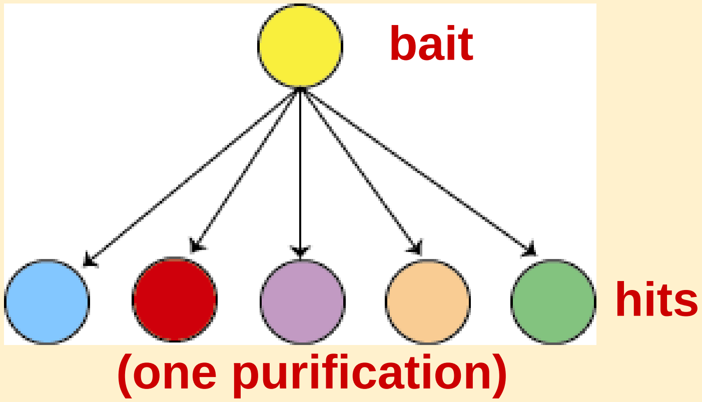
(one purification)

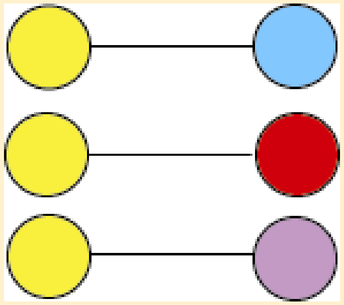


Using a ***bait*** protein, **AP-MS** technology finds ***hit*** proteins that are comembers of ***at least one*** complex with the bait.

Y2H technology finds pairs of ***physically interacting*** proteins.

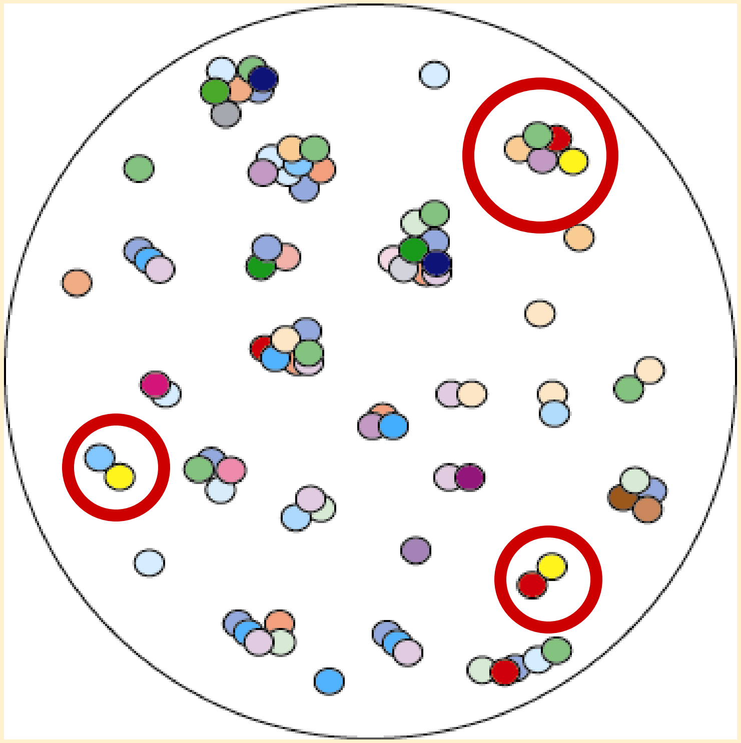




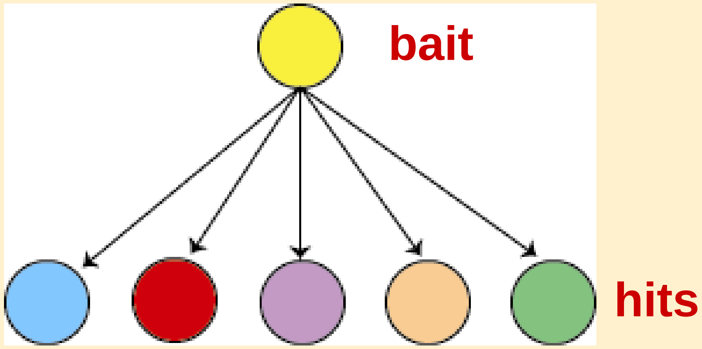


Using a ***bait*** protein, **AP-MS** technology finds ***hit*** proteins that are comembers of ***at least one*** complex with the bait.

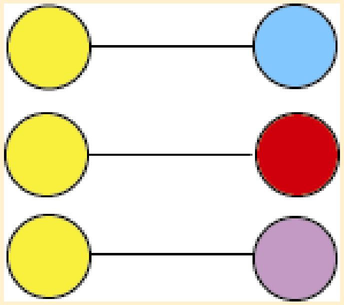
Y2H technology finds pairs of ***physically interacting*** proteins.



AP-MS data:

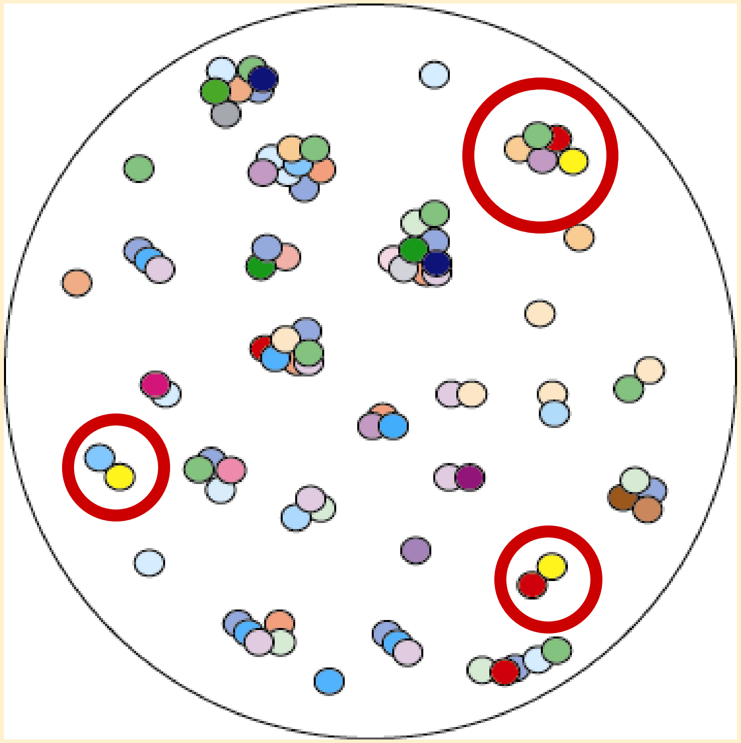


(one purification)

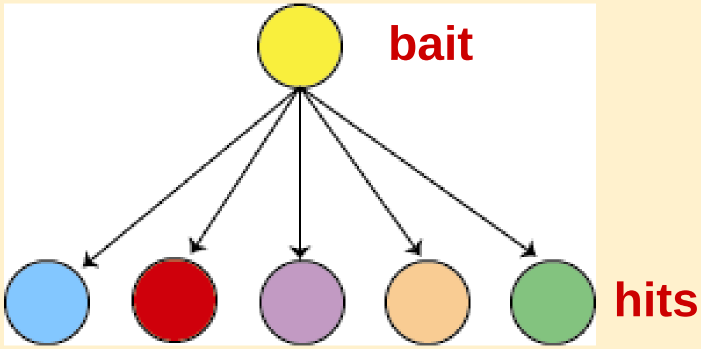


Using a ***bait*** protein, **AP-MS** technology finds ***hit*** proteins that are comembers of ***at least one*** complex with the bait.

Y2H technology finds pairs of ***physically interacting*** proteins.



AP-MS data:

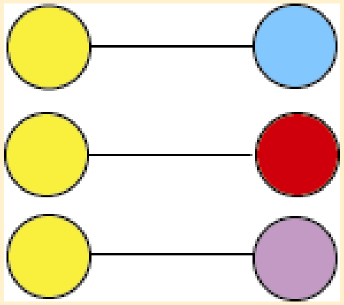


bait

hits

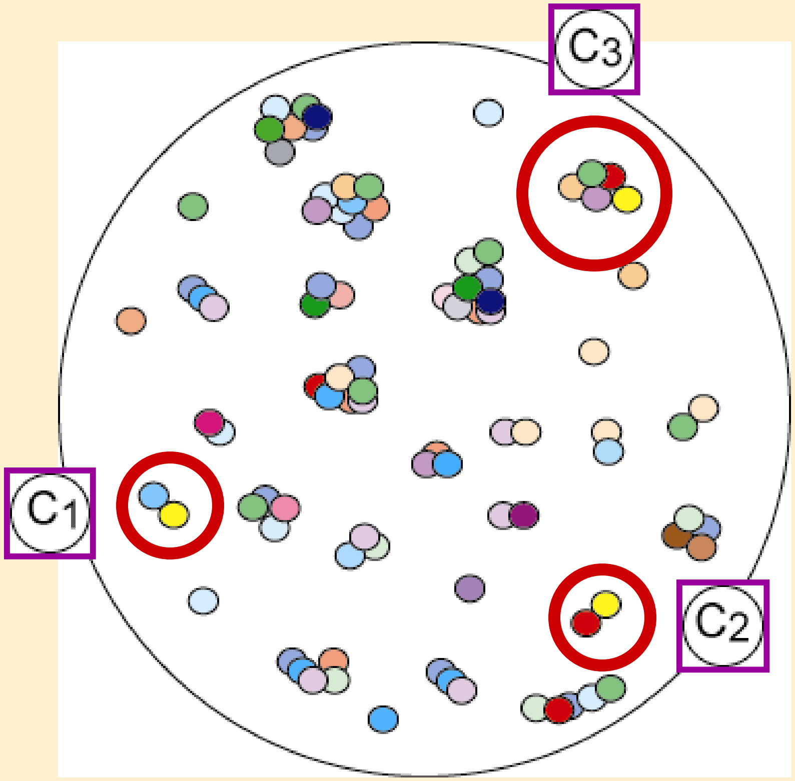
(one purification)

Y2H data:

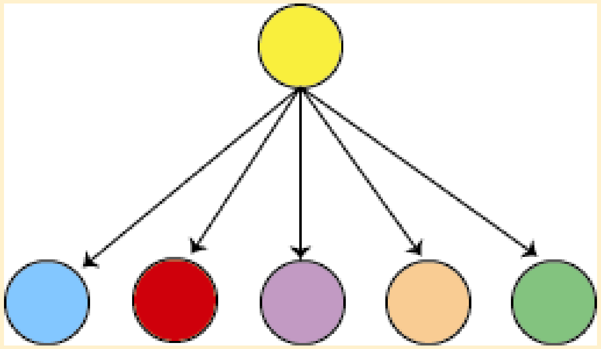


Using a ***bait*** protein, **AP-MS** technology finds ***hit*** proteins that are comembers of ***at least one*** complex with the bait.

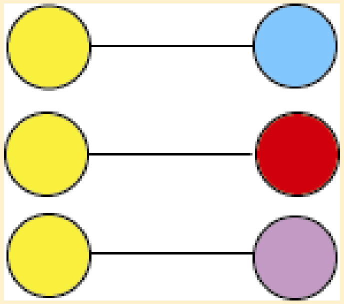
Y2H technology finds pairs of ***physically interacting*** proteins.

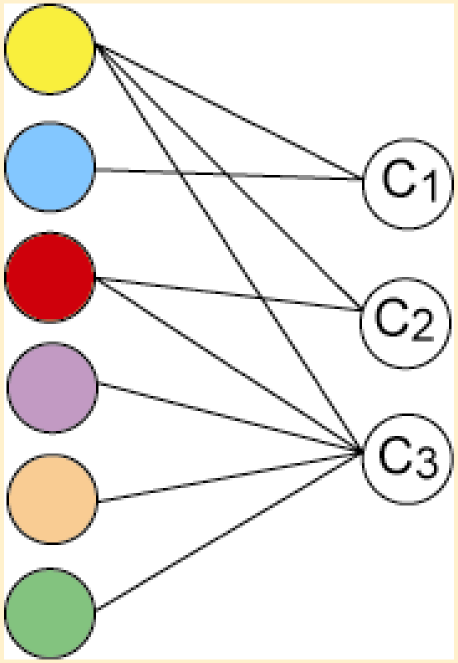


AP-MS data:

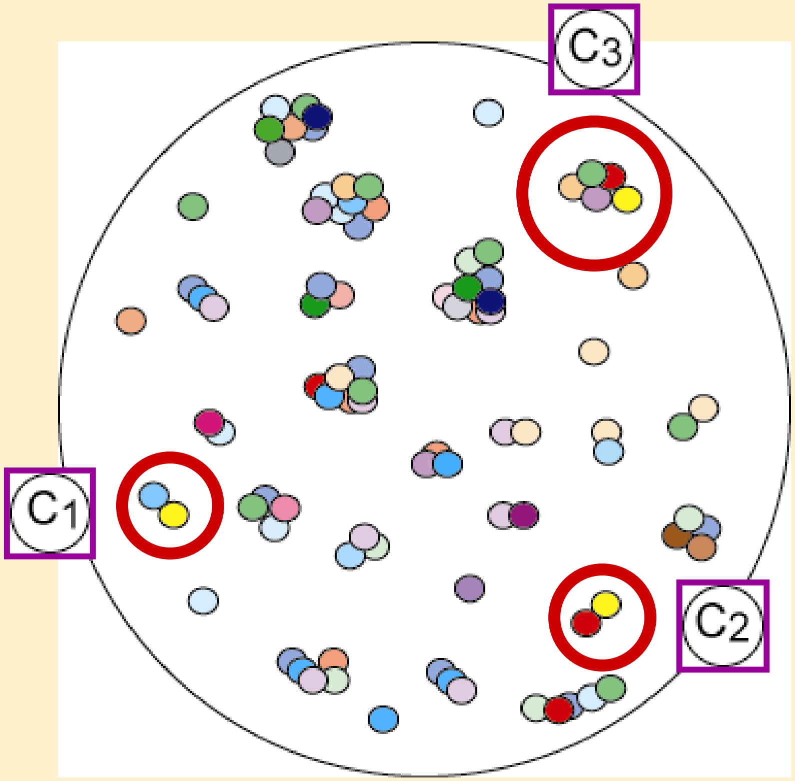


Y2H data:

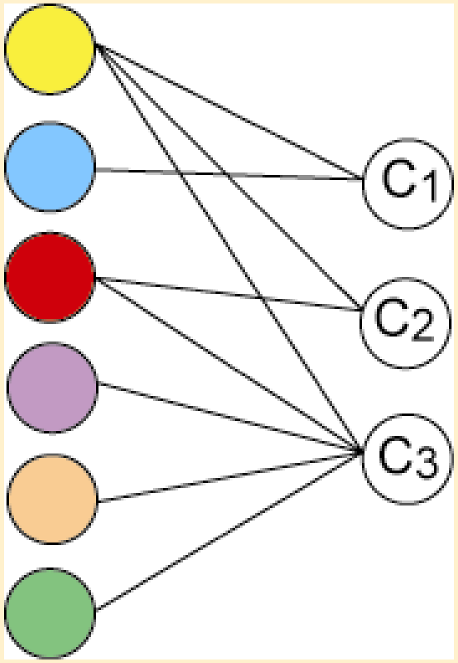




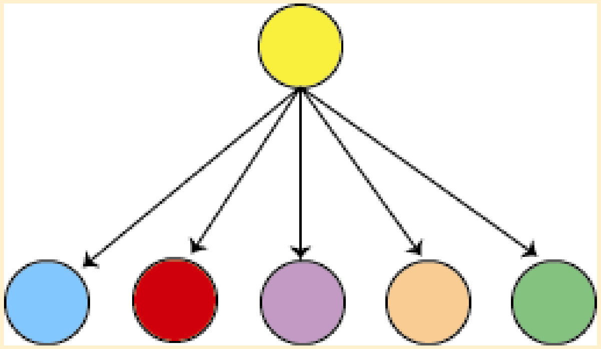
*Estimation of A requires estimation of K, the number of complexes.



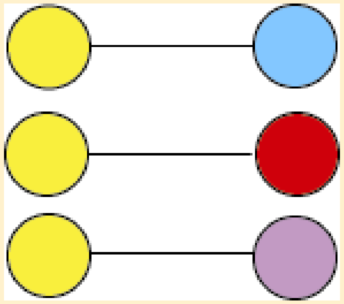
We want to estimate the bipartite protein complex membership graph, *A*:



AP-MS data:



Y2H data:



*Estimation of *A* requires estimation of K, the number of complexes.

Existing analyses of AP-MS data

- **Gavin, et al.**
  - *Functional organization of the yeast proteome by systematic analysis of protein complexes* (Nature 2002)
    - Purifications grouped together based on significant overlap (p.143)

- **Bader and Hogue**
  - *Analyzing Yeast Protein-Protein Interaction Data Obtained from Different Sources* (Nature Biotechnology, 2002)
  - *An Automated Method for Finding Molecular Complexes in Large Protein Interaction Networks* (Bioinformatics 2003)
    - Works within the realm of pairwise interactions without recognition of the bipartite graph structure for complex membership
    - “Spoke” and “Matrix” models
    - Treat AP-MS data as “hypothetical pairwise interactions”

- **Jansen, et al.**
  - *A Bayesian Networks Approach for Predicting Protein-Protein Interactions from Genomic Data (Science 2003)*
    - Deals with pairwise complex *comemberships*, not comprehensive complex *membership*

Four Unique Aspects to our Algorithm

1. Some proteins participate in more than one complex

2. In an AP-MS experiment, some proteins are used as baits and some proteins are only ever found as hits

3. Graph theoretic paradigm to allow for succinct expression of constructs involved
   - Bipartite graph for complex membership *(A)*
   - Relationship of complex *membership (A)* to complex *comembership (Y)* assayed in an AP-MS experiment *(Z)*
   - AP-MS and Y2H are different technologies that measure different relationships between proteins

4. Statistical paradigm to allow for false positive and false negative observations

1. Some proteins participate in more than one complex

PP2A

Heterotrimeric complex consisting of:

Tpd3

- regulatory A subunit

Rts1 *or* Cdc55

- regulatory B subunits

Pph21 *or* Pph22

- catalytic subunits

Jiang and Broach (1999). *EMBO.*

Gavin, et al. (2002)
Rgraphviz plot of
yTAP C151

Bader & Hogue (2002)
Portion of Figure 2:
Overlap of the spoke models
of TAP and HMS-PCI.

Jansen, et al. (2003)
PIT Bayesian Network, LR>600
central node=Tpd3
http://genecensus.org/intint
YGL109C=Cdc55, YDL134C=Pph21, YDL188C=Pph22
YCR002C=Cdc10, YJR076C=Cdc11, YMR109W=Myo5

1. Some proteins participate in more than one complex

PP2A

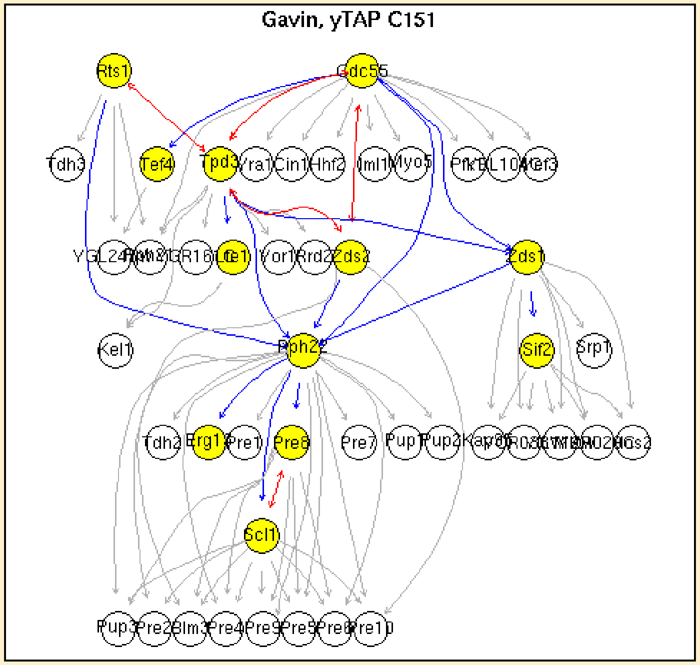
Heterotrimeric complex consisting of:

Tpd3
- regulatory A subunit

Rts1 *or* Cdc55
- regulatory B subunits

Pph21 *or* Pph22
- catalytic subunits

Jiang and Broach (1999). *EMBO.*



Gavin, et al. (2002)
Rgraphviz plot of yTAP C151

Bader & Hogue (2002)
Portion of Figure 2:
Overlap of the spoke models of TAP and HMS-PCI.

Jansen, et al. (2003)
PIT Bayesian Network, LR>600
central node=Tpd3
http://genecensus.org/intint
YGL109C=Cdc55, YDL134C=Pph21, YDL188C=Pph22
YCR002C=Cdc10, YJR076C=Cdc11, YMR109W=Myo5

1. Some proteins participate in more than one complex

PP2A

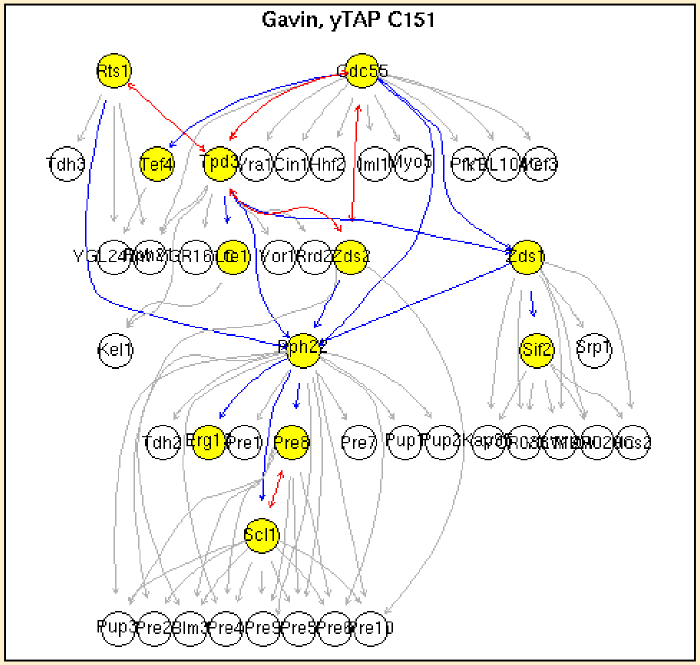
Heterotrimeric complex consisting of:

Tpd3
- regulatory A subunit

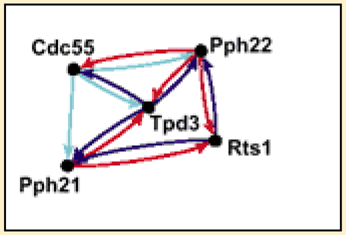
Rts1 *or* Cdc55
- regulatory B subunits

Pph21 *or* Pph22
- catalytic subunits

Jiang and Broach (1999). *EMBO.*



Gavin, et al. (2002)
Rgraphviz plot of yTAP C151



Bader & Hogue (2002)
Portion of Figure 2: Overlap of the spoke models of TAP and HMS-PCI.

Jansen, et al. (2003)
PIT Bayesian Network, LR>600
central node=Tpd3
http://genecensus.org/intint
YGL109C=Cdc55, YDL134C=Pph21, YDL188C=Pph22
YCR002C=Cdc10, YJR076C=Cdc11, YMR109W=Myo5

1. Some proteins participate in more than one complex

PP2A

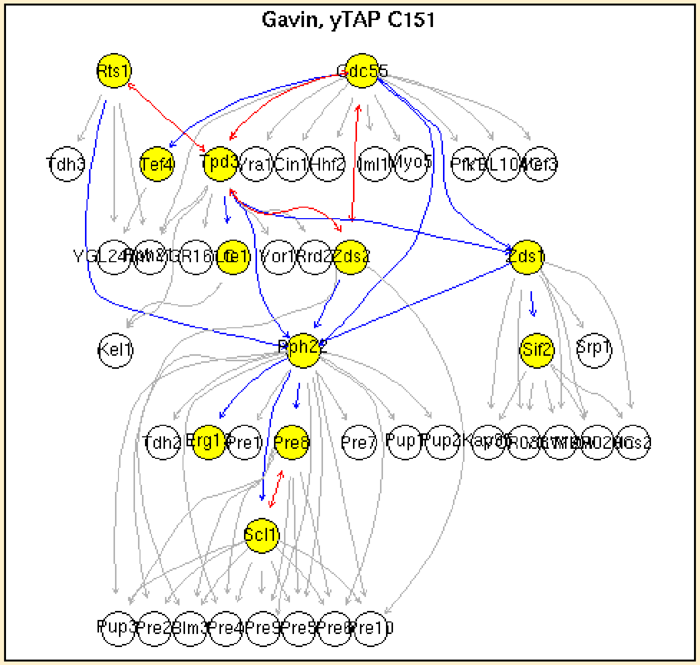
Heterotrimeric complex consisting of:

Tpd3
- regulatory A subunit

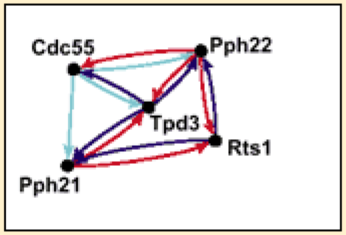
Rts1 *or* Cdc55
- regulatory B subunits

Pph21 *or* Pph22
- catalytic subunits

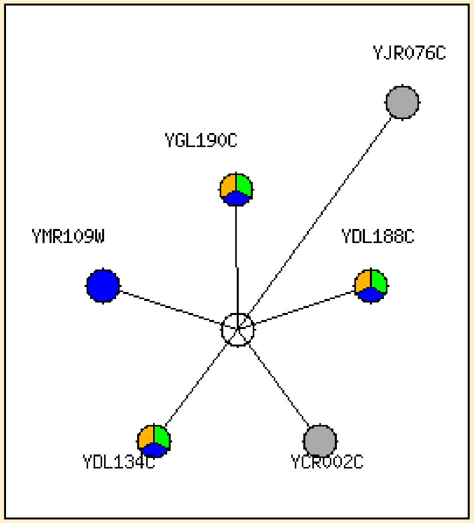
Jiang and Broach (1999). *EMBO.*



Gavin, et al. (2002)
Rgraphviz plot of yTAP C151



Bader & Hogue (2002)
Portion of Figure 2:
Overlap of the spoke models of TAP and HMS-PCI.



Jansen, et al. (2003)
PIT Bayesian Network, LR>600
central node=Tpd3
http://genecensus.org/intint
YGL109C=Cdc55, YDL134C=Pph21, YDL188C=Pph22
YCR002C=Cdc10, YJR076C=Cdc11, YMR109W=Myo5

1. Some proteins participate in more than one complex

Our algorithm detects:

PP2A

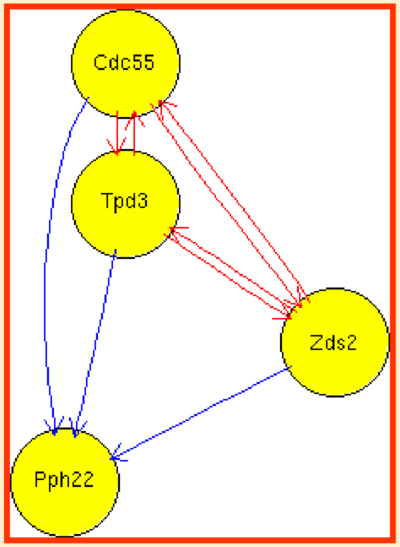
Heterotrimeric complex consisting of:

Tpd3
- regulatory A subunit

Rts1 *or* Cdc55
- regulatory B subunits

Pph21 *or* Pph22
- catalytic subunits

Jiang and Broach (1999). *EMBO.*



Zds1 and Zds2 (known cell-cycle regulators) only exist in complexes with the Cdc55-Pph22 trimer!

1. Some proteins participate in more than one complex

PP2A

Heterotrimeric complex consisting of:

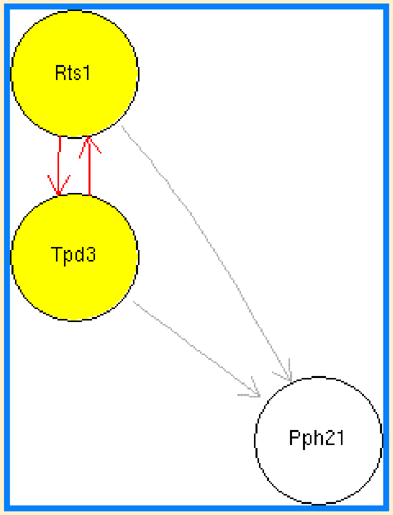
Tpd3
- regulatory A subunit

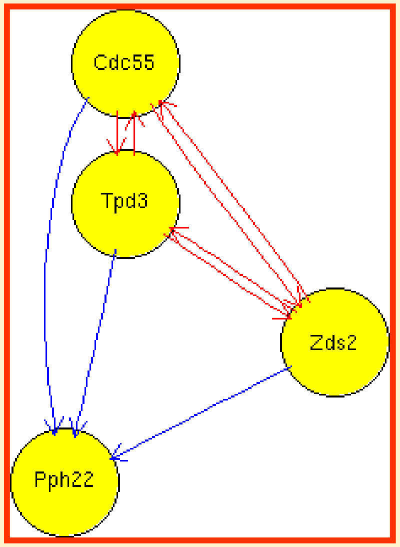
Rts1 ***or*** **Cdc55**
- regulatory B subunits

Pph21 ***or*** **Pph22**
- catalytic subunits

Jiang and Broach (1999). *EMBO.*

Our algorithm detects:





Zds1 and Zds2 (known cell-cycle regulators) only exist in complexes with the Cdc55-Pph22 trimer!

1. Some proteins participate in more than one complex

PP2A

Heterotrimeric complex consisting of:

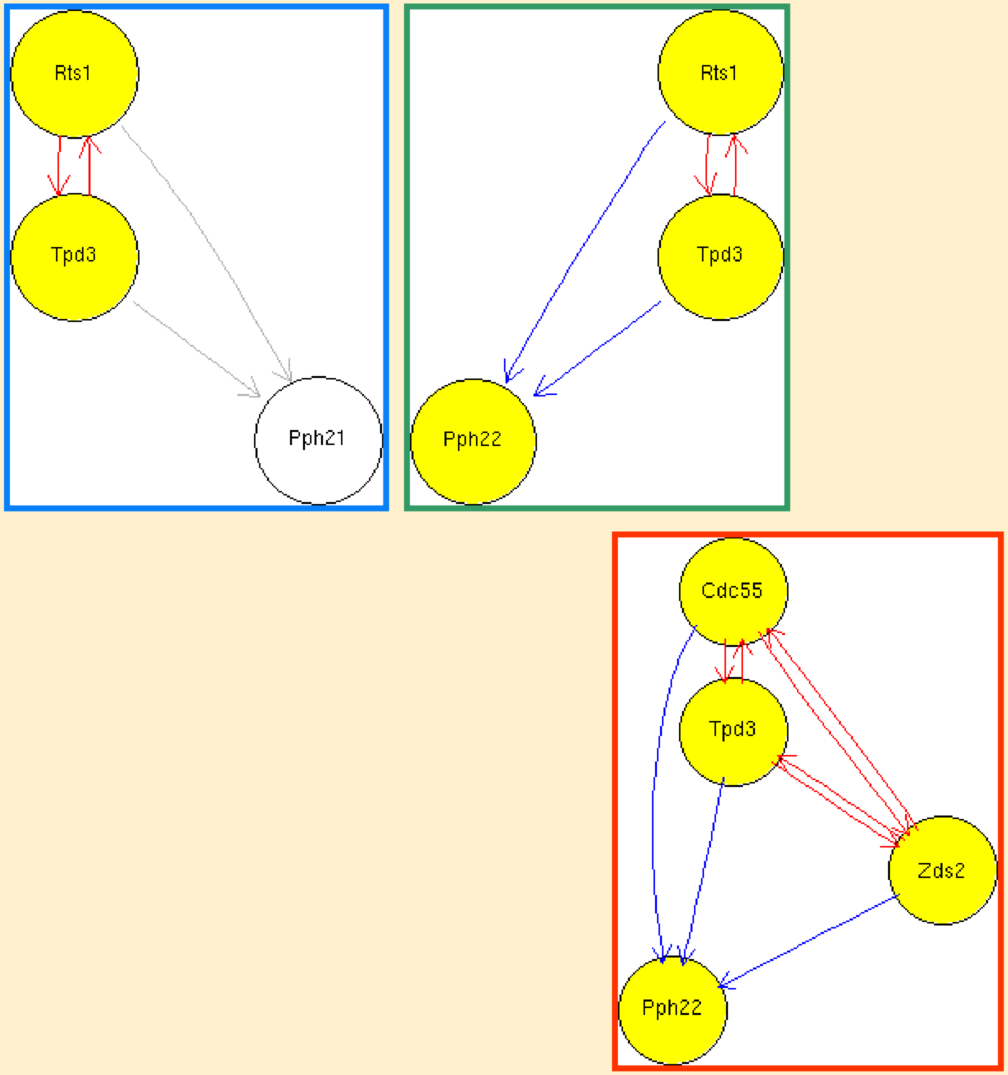
Tpd3
- regulatory A subunit

Rts1 *or* Cdc55
- regulatory B subunits

Pph21 *or* Pph22
- catalytic subunits

Jiang and Broach (1999). *EMBO.*

Our algorithm detects:



Zds1 and Zds2 (known cell-cycle regulators) only exist in complexes with the Cdc55-Pph22 trimer!

1. Some proteins participate in more than one complex

PP2A

Heterotrimeric complex consisting of:

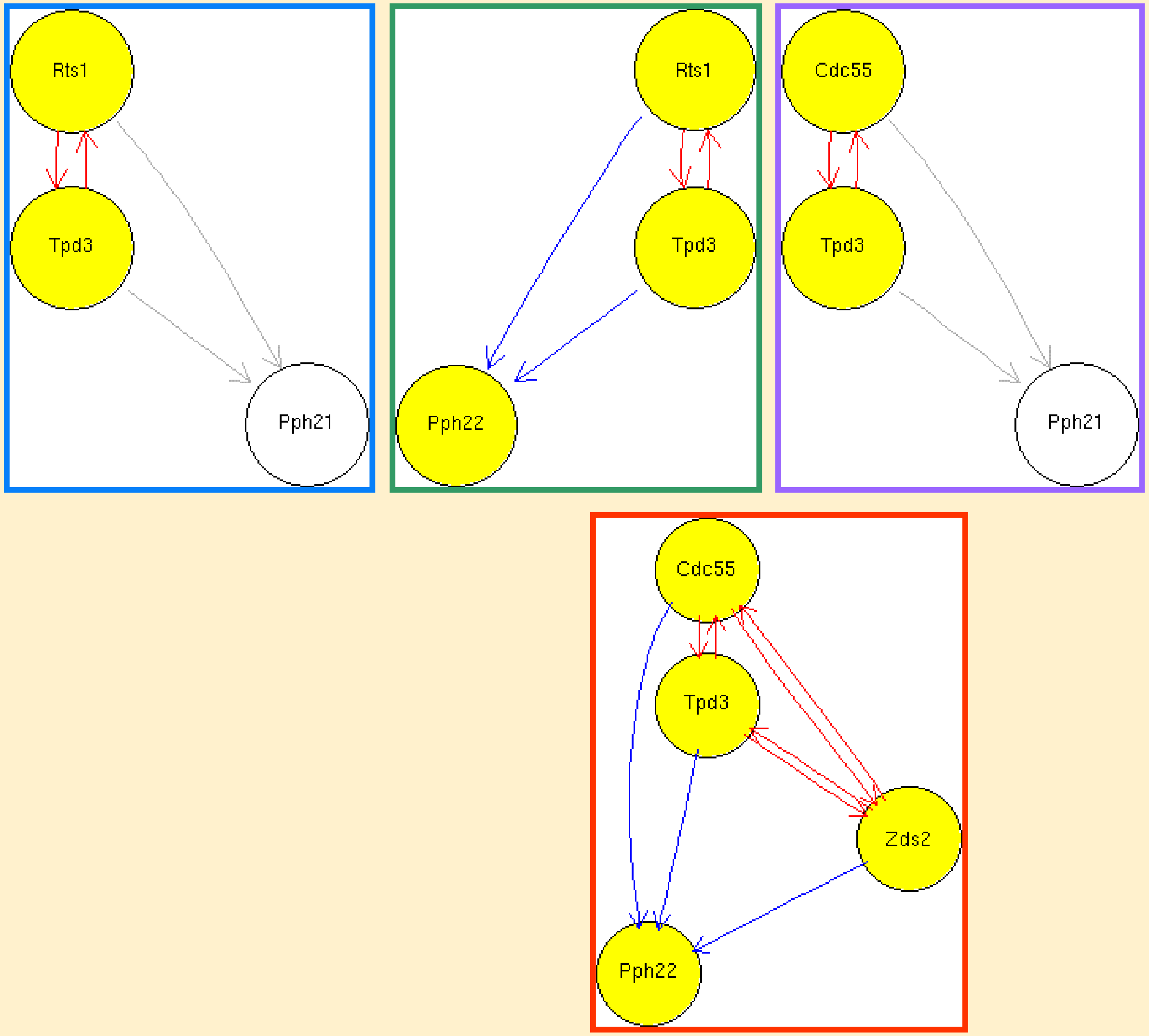
Tpd3
- regulatory A subunit

Rts1 *or* Cdc55
- regulatory B subunits

Pph21 *or* Pph22
- catalytic subunits

Jiang and Broach (1999). *EMBO.*

Our algorithm detects:



Zds1 and Zds2 (known cell-cycle regulators) only exist in complexes with the Cdc55-Pph22 trimer!

1. Some proteins participate in more than one complex

PP2A

Heterotrimeric complex consisting of:

Tpd3

- regulatory A subunit

Rts1 *or* Cdc55

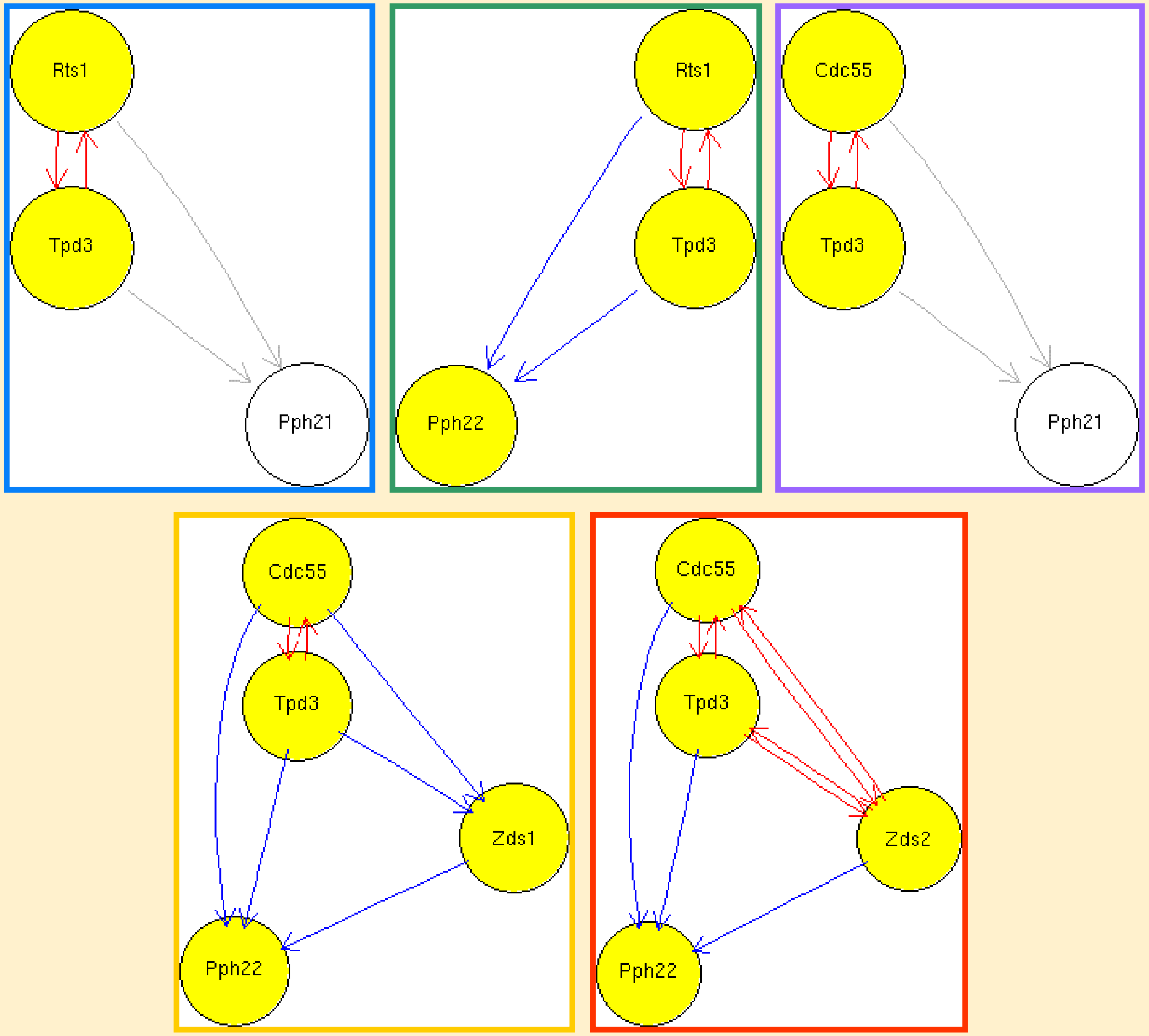
- regulatory B subunits

Pph21 *or* Pph22

- catalytic subunits

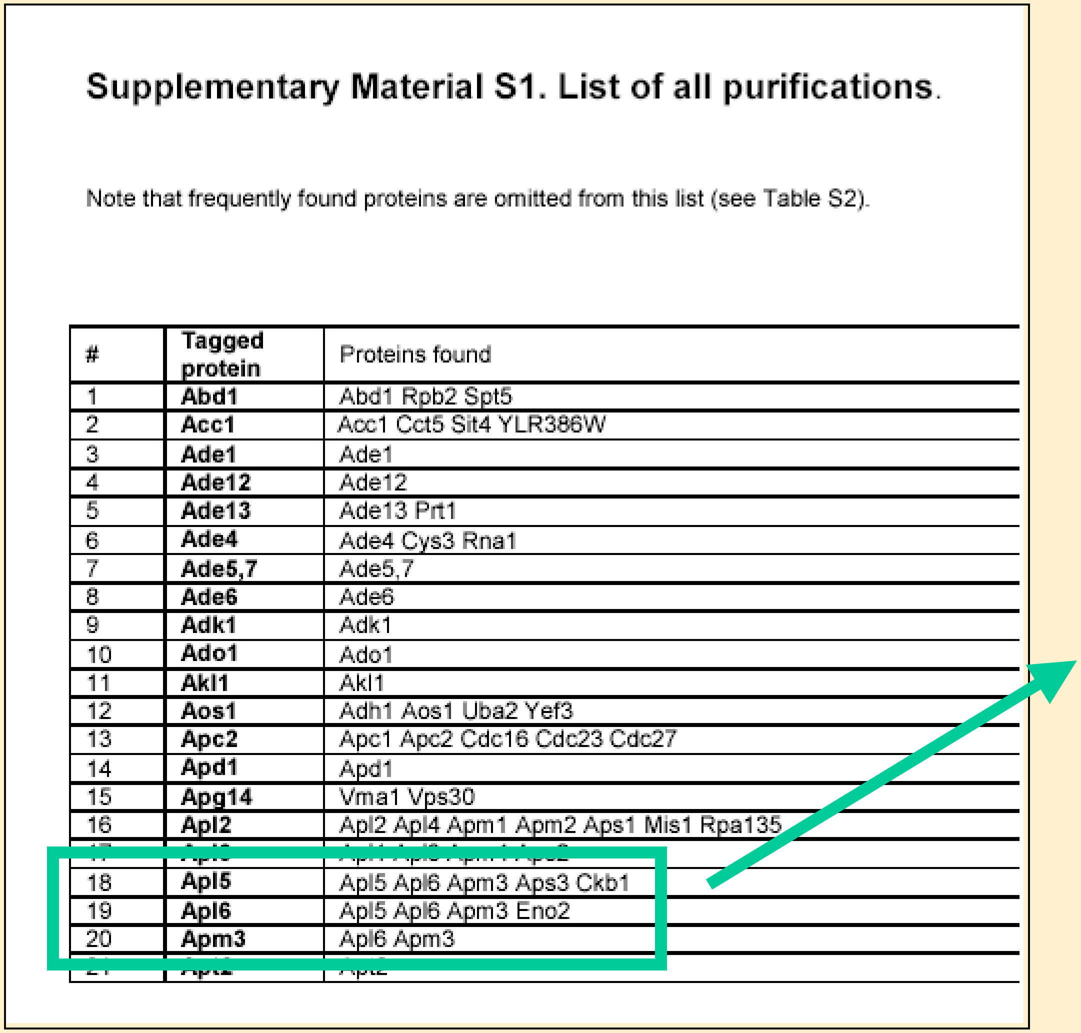
Jiang and Broach (1999). *EMBO.*

Our algorithm detects:



Zds1 and Zds2 (known cell-cycle regulators) only exist in complexes with the Cdc55-Pph22 trimer!

2. In an AP-MS experiment, some proteins are used as baits and some proteins are only ever found as hits

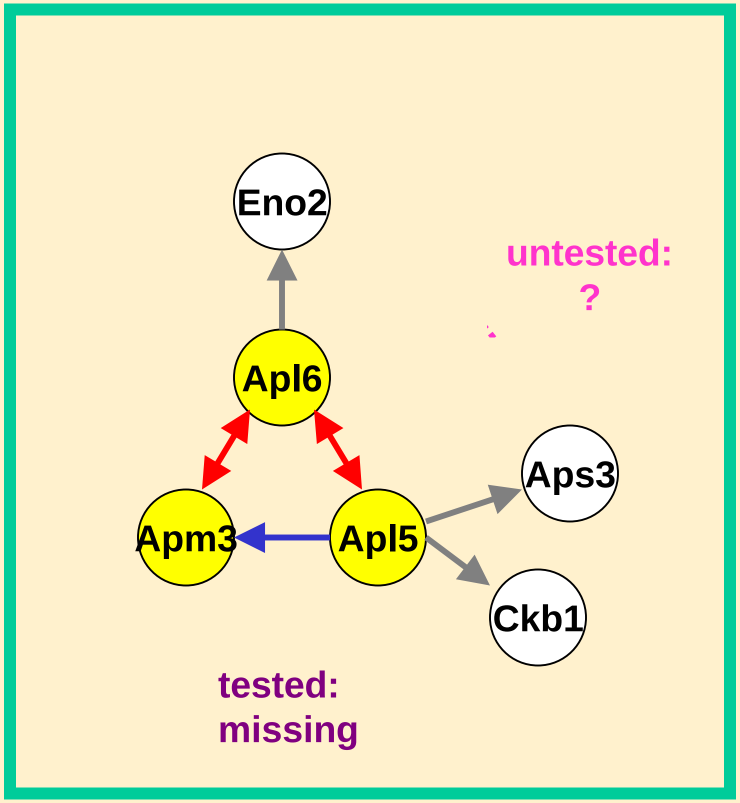


Supplementary Material S1. List of all purifications.

Note that frequently found proteins are omitted from this list (see Table S2).

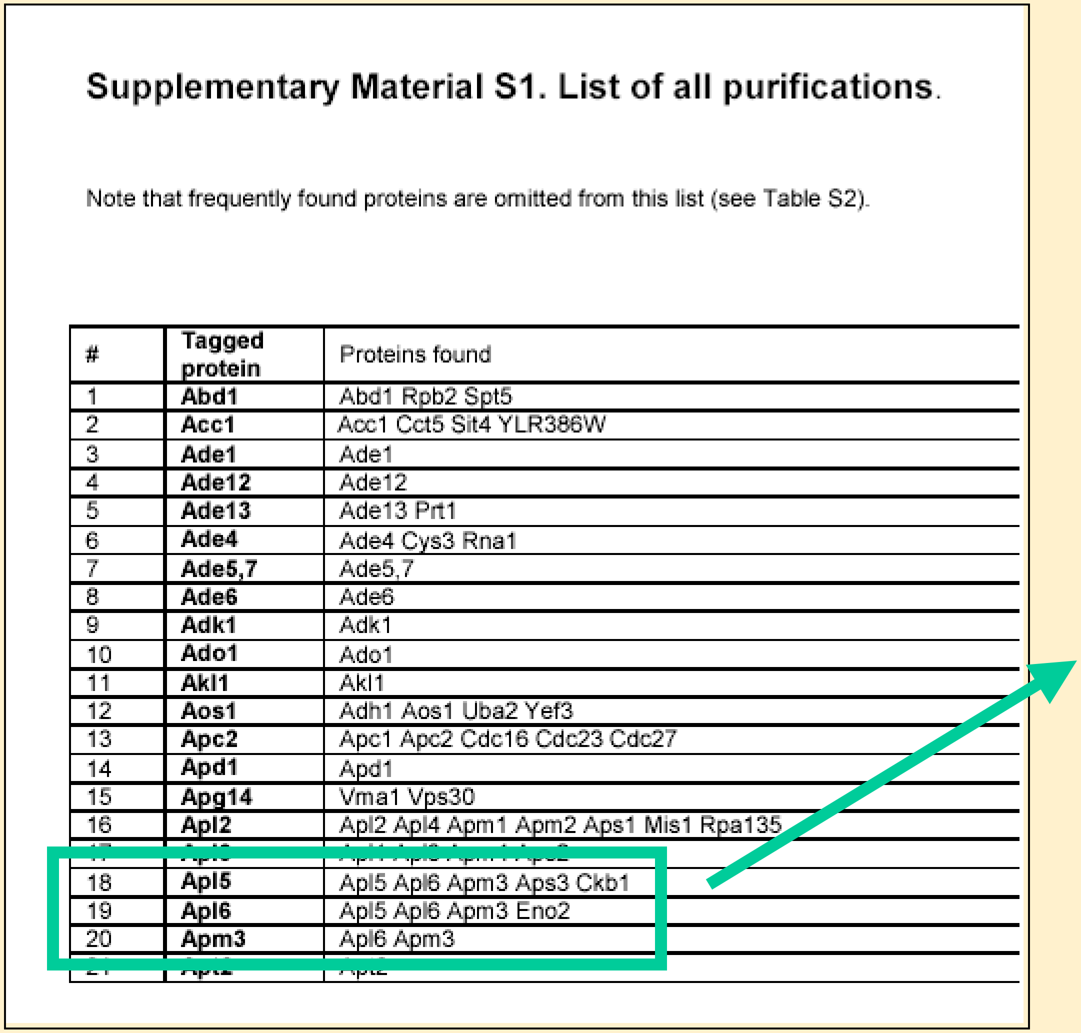
#	Tagged protein	Proteins found
1	**Abd1**	Abd1 Rpb2 Spt5
2	**Acc1**	Acc1 Cct5 Sit4 YLR386W
3	**Ade1**	Ade1
4	**Ade12**	Ade12
5	**Ade13**	Ade13 Prt1
6	**Ade4**	Ade4 Cys3 Rna1
7	**Ade5,7**	Ade5,7
8	**Ade6**	Ade6
9	**Adk1**	Adk1
10	**Ado1**	Ado1
11	**Akl1**	Akl1
12	**Aos1**	Adh1 Aos1 Uba2 Yef3
13	**Apc2**	Apc1 Apc2 Cdc16 Cdc23 Cdc27
14	**Apd1**	Apd1
15	**Apg14**	Vma1 Vps30
16	**Apl2**	Apl2 Apl4 Apm1 Apm2 Aps1 Mis1 Rpa135
17	**Apl3**	Apl1 Apl3 Apm4 Aps2
18	**Apl5**	Apl5 Apl6 Apm3 Aps3 Ckb1
19	**Apl6**	Apl5 Apl6 Apm3 Eno2
20	**Apm3**	Apl6 Apm3
21	**Apt2**	Apt2

Subgraph of *Z*



Raw TAP purifications (Gavin et al.)
Available at http://www.nature.com

2. In an AP-MS experiment, some proteins are used as baits and some proteins are only ever found as hits

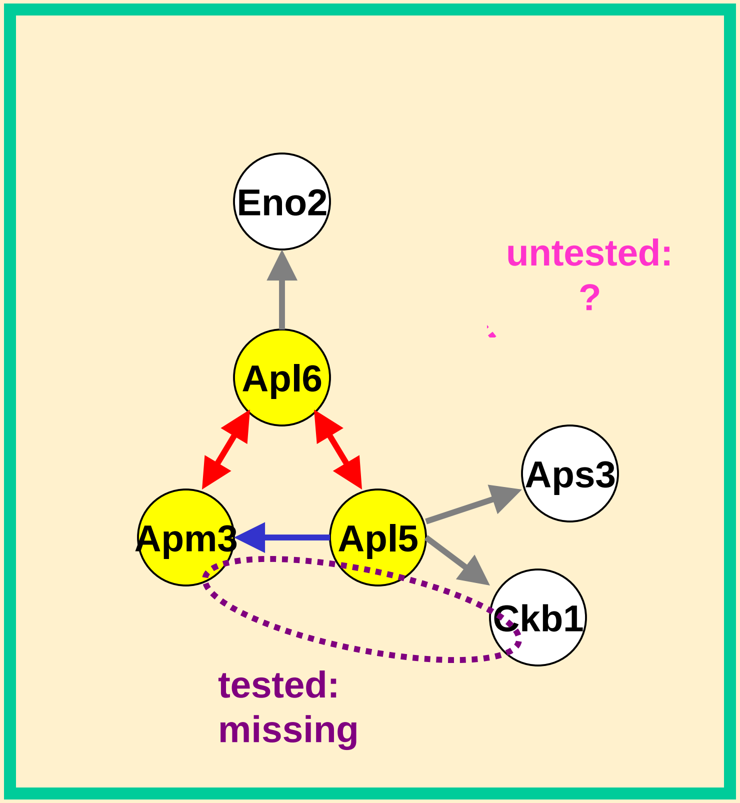


Supplementary Material S1. List of all purifications.

Note that frequently found proteins are omitted from this list (see Table S2).

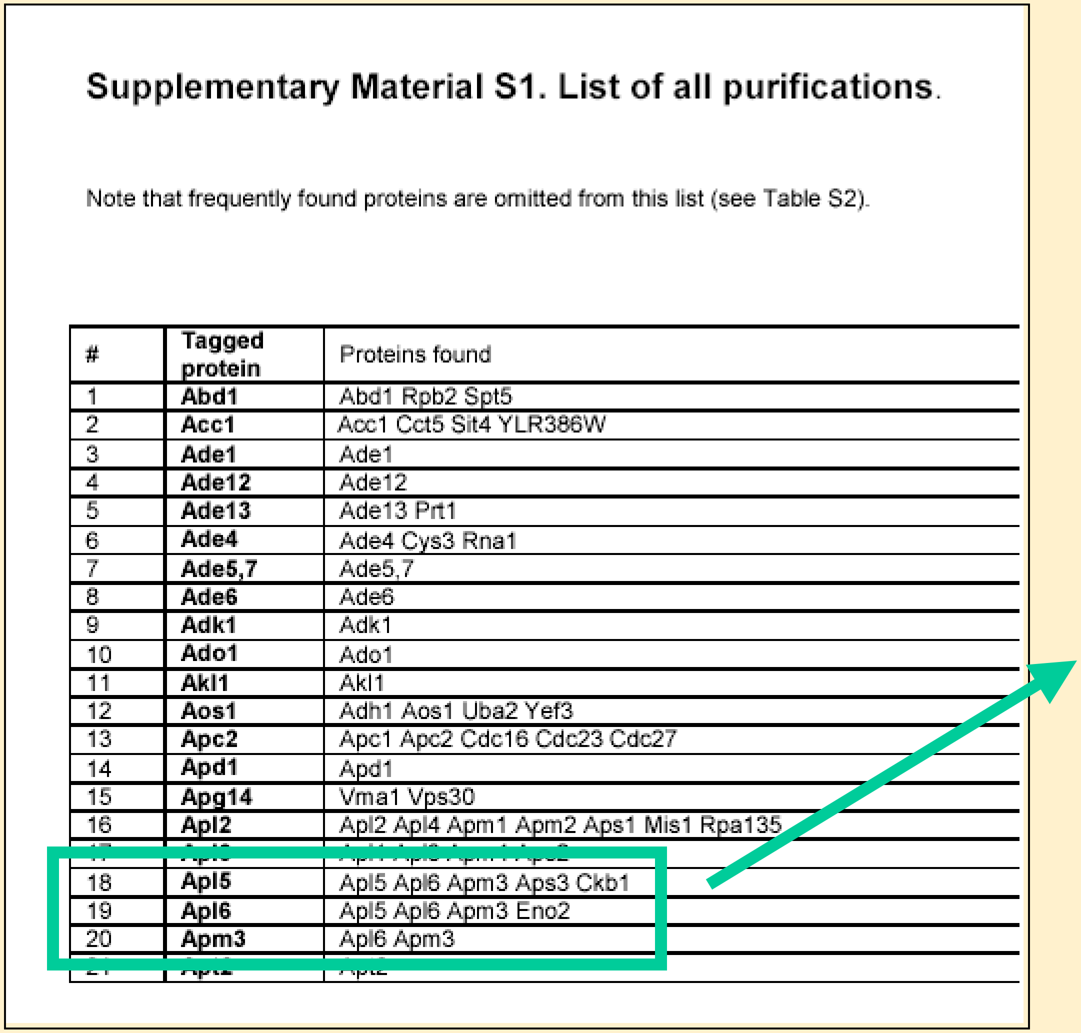
#	Tagged protein	Proteins found
1	**Abd1**	Abd1 Rpb2 Spt5
2	**Acc1**	Acc1 Cct5 Sit4 YLR386W
3	**Ade1**	Ade1
4	**Ade12**	Ade12
5	**Ade13**	Ade13 Prt1
6	**Ade4**	Ade4 Cys3 Rna1
7	**Ade5,7**	Ade5,7
8	**Ade6**	Ade6
9	**Adk1**	Adk1
10	**Ado1**	Ado1
11	**Akl1**	Akl1
12	**Aos1**	Adh1 Aos1 Uba2 Yef3
13	**Apc2**	Apc1 Apc2 Cdc16 Cdc23 Cdc27
14	**Apd1**	Apd1
15	**Apg14**	Vma1 Vps30
16	**Apl2**	Apl2 Apl4 Apm1 Apm2 Aps1 Mis1 Rpa135
17	**Apl3**	Apl1 Apl3 Apm4 Aps2
18	**Apl5**	Apl5 Apl6 Apm3 Aps3 Ckb1
19	**Apl6**	Apl5 Apl6 Apm3 Eno2
20	**Apm3**	Apl6 Apm3
21	**Apt2**	Apt2

Subgraph of *Z*



Raw TAP purifications (Gavin et al.)
Available at http://www.nature.com

2. In an AP-MS experiment, some proteins are used as baits and some proteins are only ever found as hits

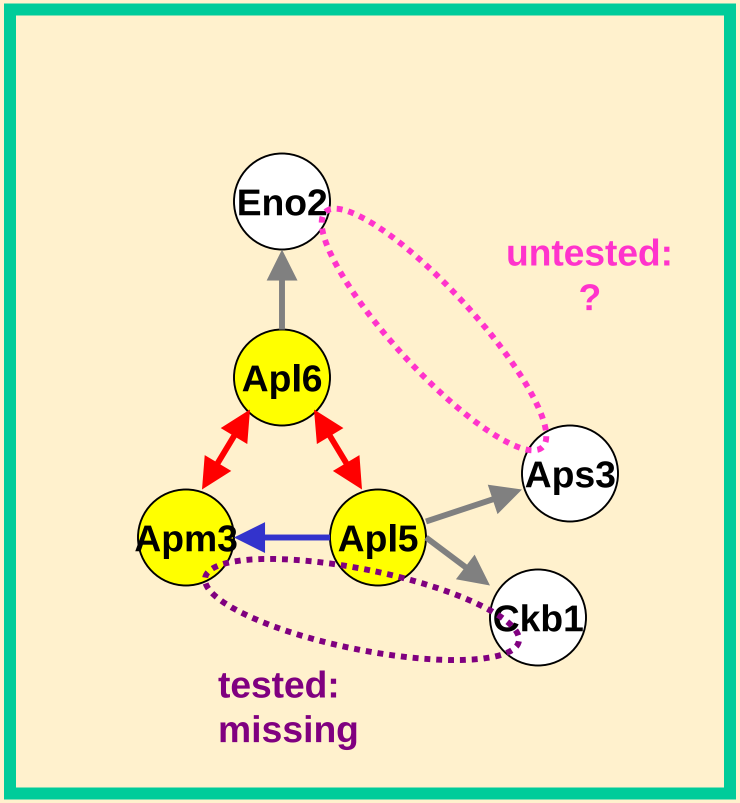


Supplementary Material S1. List of all purifications.

Note that frequently found proteins are omitted from this list (see Table S2).

#	Tagged protein	Proteins found
1	**Abd1**	Abd1 Rpb2 Spt5
2	**Acc1**	Acc1 Cct5 Sit4 YLR386W
3	**Ade1**	Ade1
4	**Ade12**	Ade12
5	**Ade13**	Ade13 Prt1
6	**Ade4**	Ade4 Cys3 Rna1
7	**Ade5,7**	Ade5,7
8	**Ade6**	Ade6
9	**Adk1**	Adk1
10	**Ado1**	Ado1
11	**Akl1**	Akl1
12	**Aos1**	Adh1 Aos1 Uba2 Yef3
13	**Apc2**	Apc1 Apc2 Cdc16 Cdc23 Cdc27
14	**Apd1**	Apd1
15	**Apg14**	Vma1 Vps30
16	**Apl2**	Apl2 Apl4 Apm1 Apm2 Aps1 Mis1 Rpa135
17	**Apl3**	Apl1 Apl3 Apm1 Aps2
18	**Apl5**	Apl5 Apl6 Apm3 Aps3 Ckb1
19	**Apl6**	Apl5 Apl6 Apm3 Eno2
20	**Apm3**	Apl6 Apm3
21	**Apt2**	Apt2

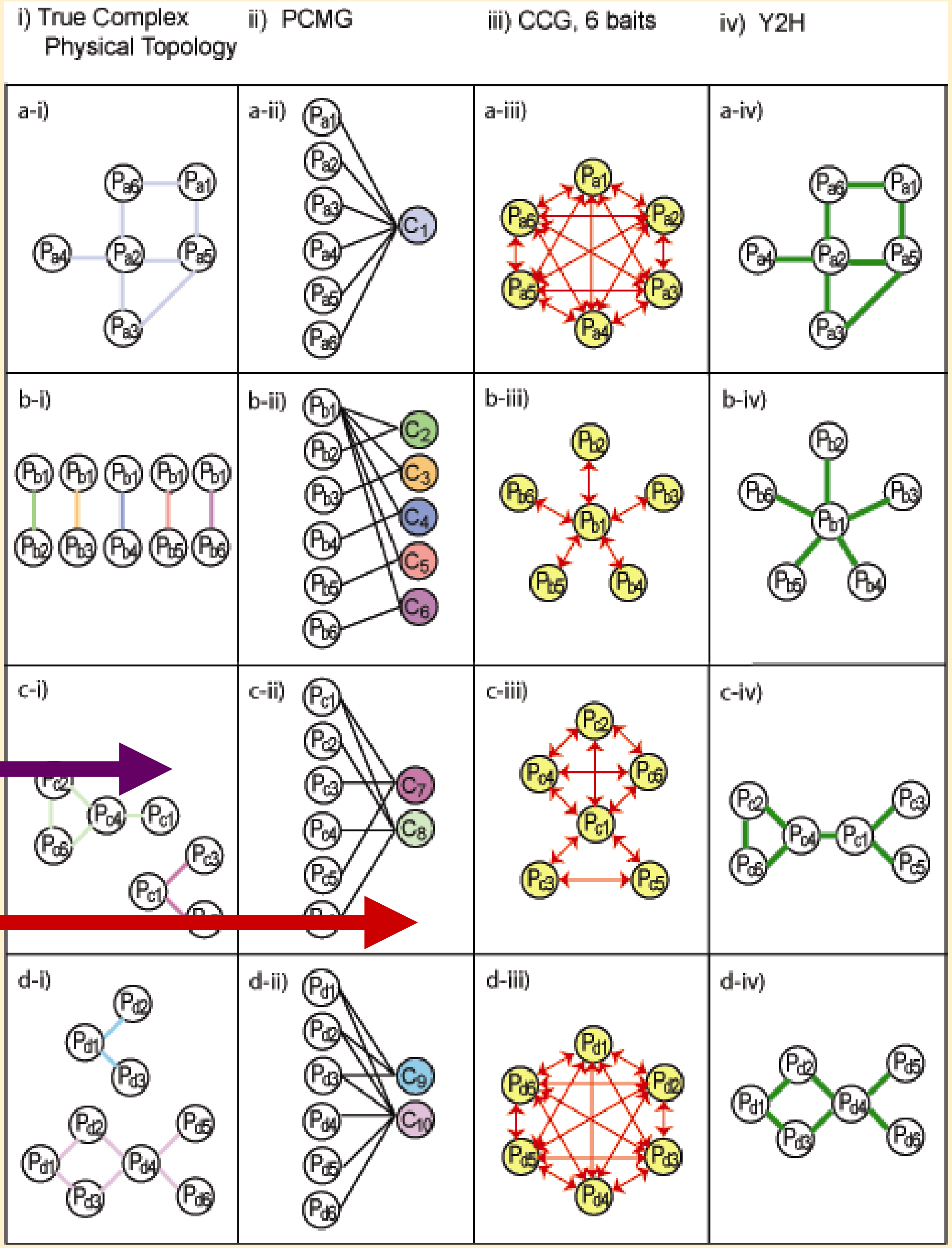
Subgraph of *Z*



Raw TAP purifications (Gavin et al.)
Available at http://www.nature.com

3. Graph theoretic paradigm to allow for succinct expression of constructs involved

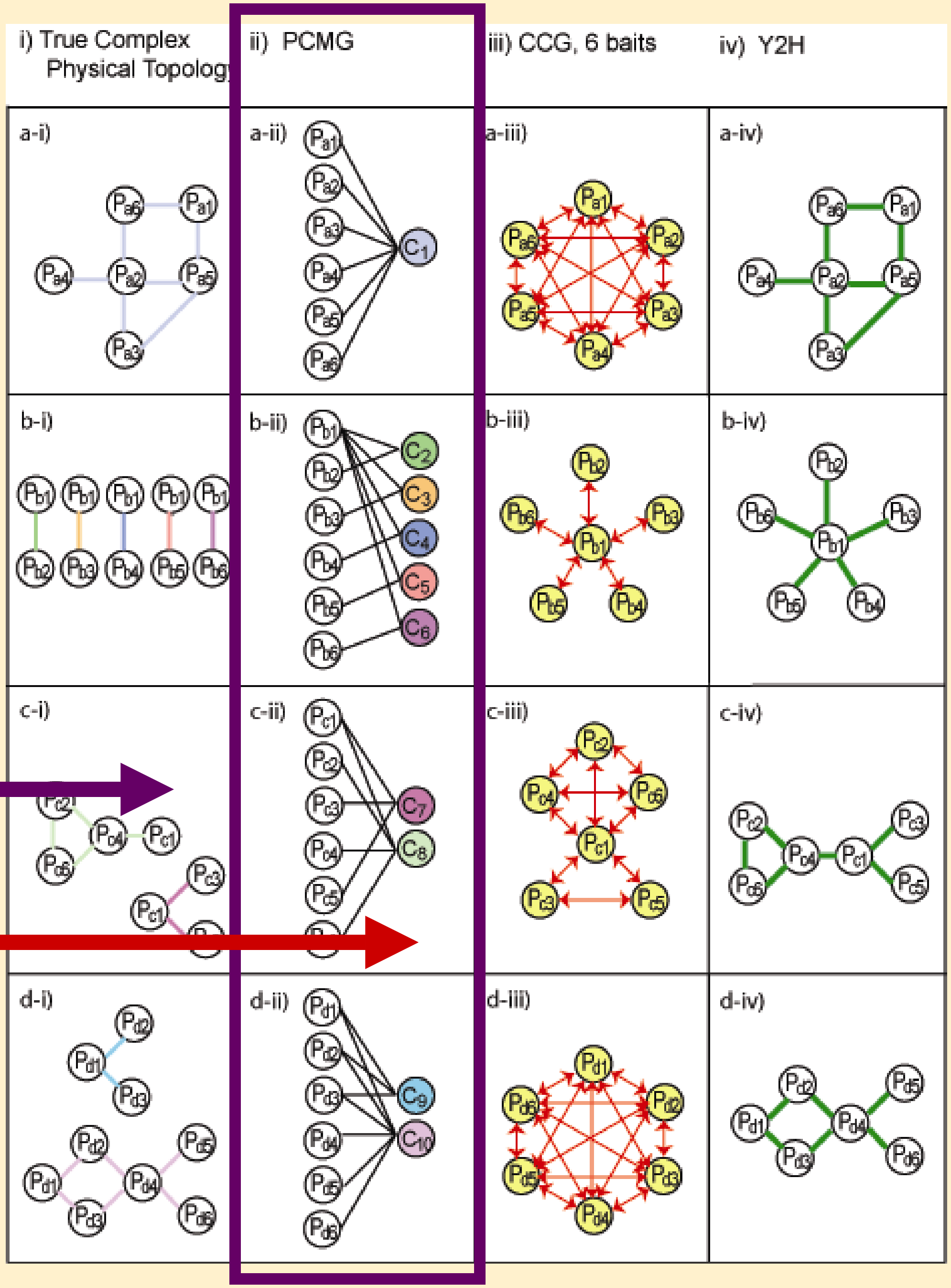
- *Bipartite graph for complex membership*
- *Relationship of complex membership (A) to complex comembership (Y) assayed in an AP-MS experiment (Z)*
- *AP-MS and Y2H are different technologies that measure different relationships between proteins*



We want to estimate A using AP-MS assays of Y.

3. Graph theoretic paradigm to allow for succinct expression of constructs involved

- *Bipartite graph for complex membership*
- *Relationship of complex membership (A) to complex comembership (Y) assayed in an AP-MS experiment (Z)*
- *AP-MS and Y2H are different technologies that measure different relationships between proteins*

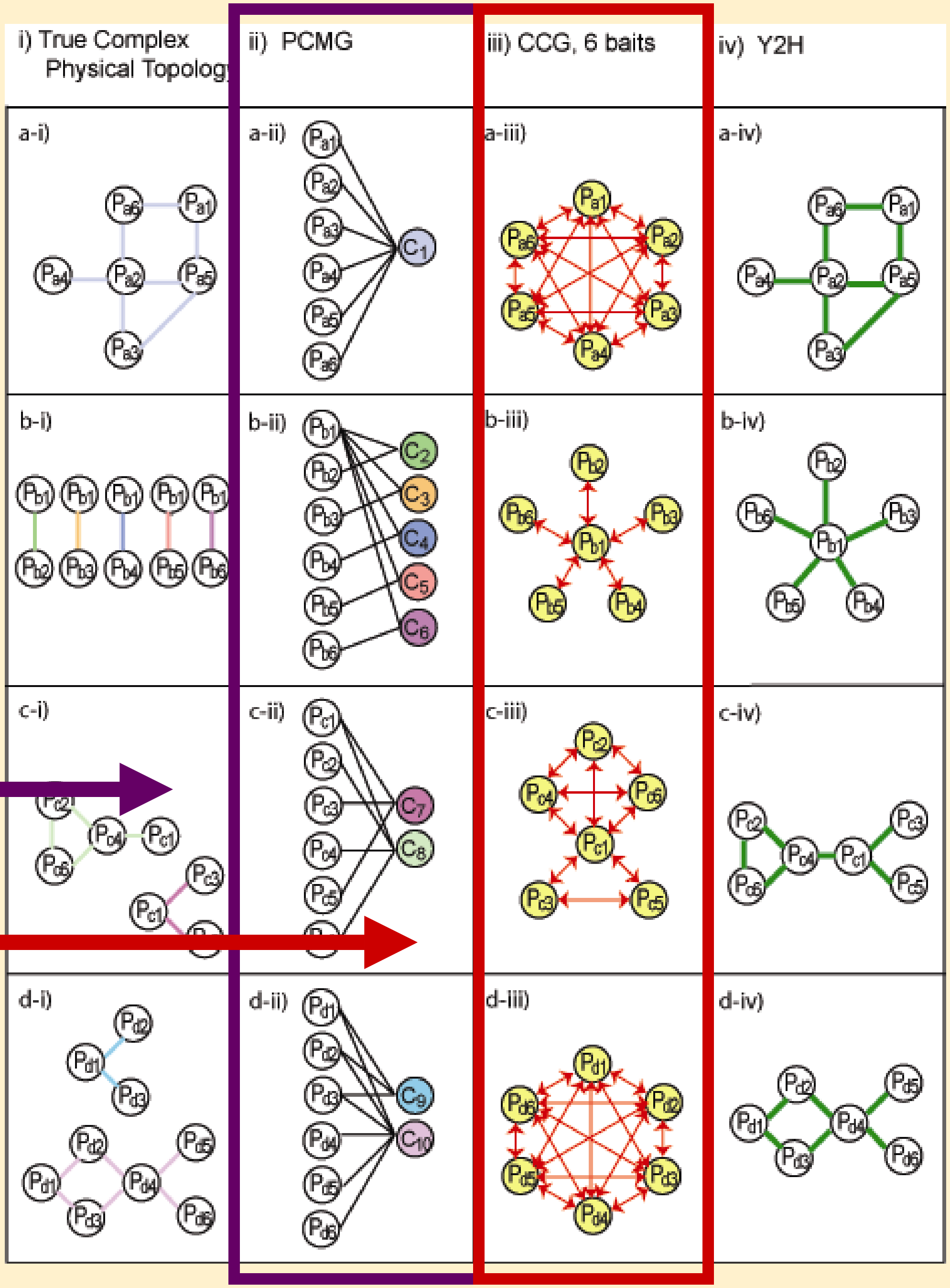


We want to estimate A using AP-MS assays of Y.

3. Graph theoretic paradigm to allow for succinct expression of constructs involved

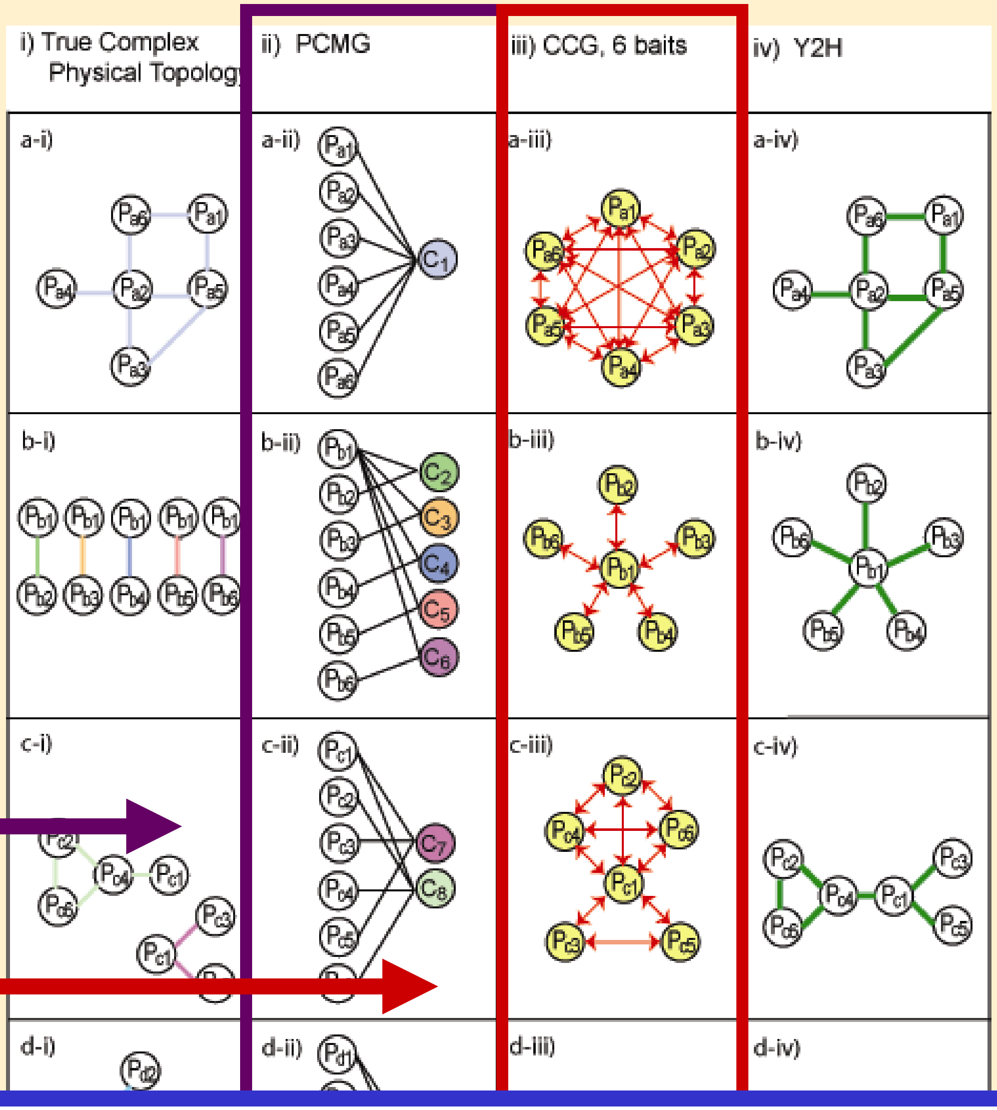
- *Bipartite graph for complex membership*
- *Relationship of complex membership (A) to complex comembership (Y) assayed in an AP-MS experiment (Z)*
- *AP-MS and Y2H are different technologies that measure different relationships between proteins*

We want to estimate A using AP-MS assays of Y.



3. Graph theoretic paradigm to allow for succinct expression of constructs involved

- *•Bipartite graph for complex membership*
- *•Relationship of complex membership (A) to complex comembership (Y) assayed in an AP-MS experiment (Z)*
- *•AP-MS and Y2H are different technologies that measure different relationships between proteins*



We want to estimate A using AP-MS assays of Y.

The Connection: Maximal Complete Subgraphs

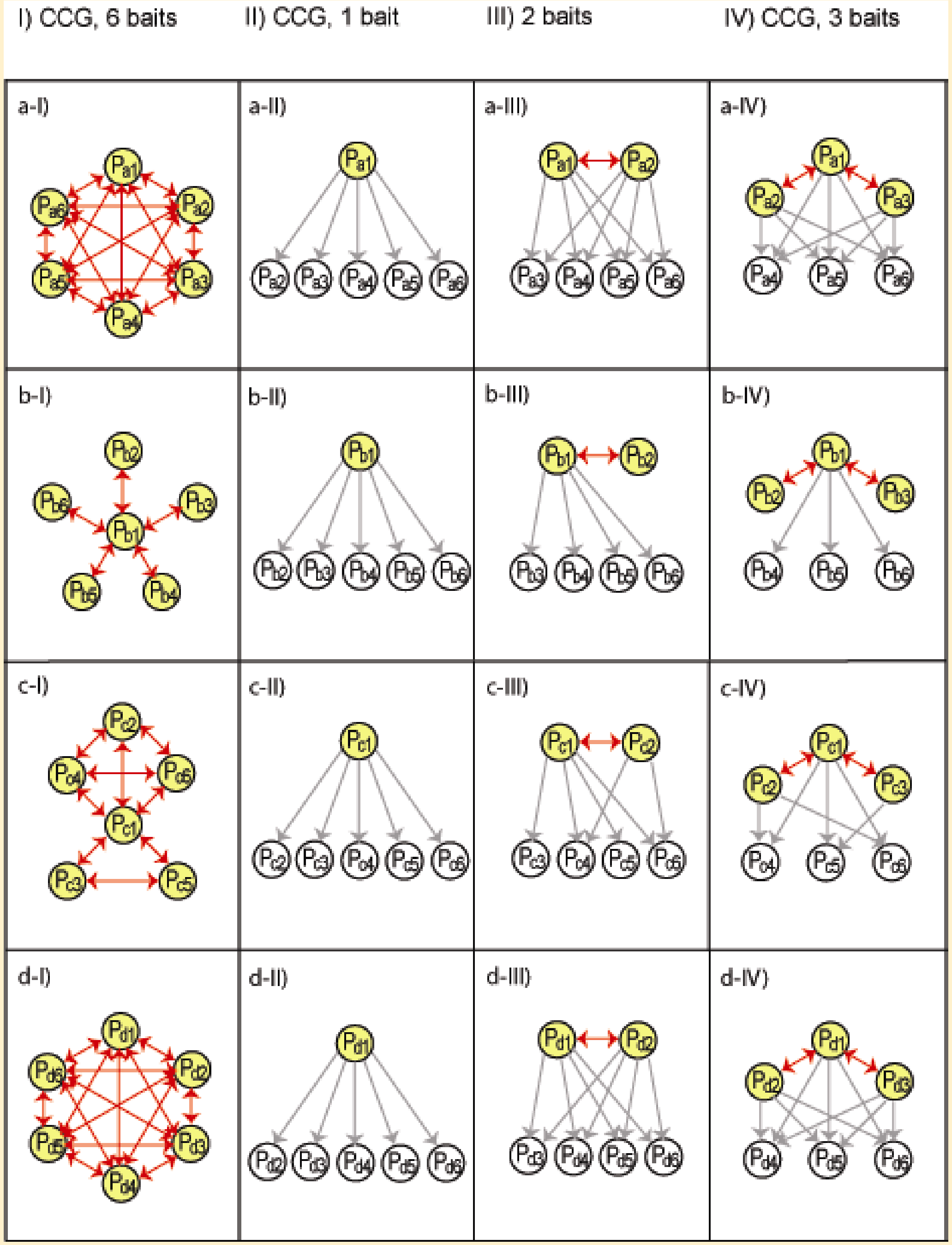
Complete Subgraph: set of *n* nodes for which all *n(n-1)* directed edges exist

Maximal Complete Subgraph: complete subgraph that is not contained in any other complete subgraph

3. Graph theoretic paradigm to allow for succinct expression of constructs involved

- *Relationship of complex membership (A) to complex comembership (Y) assayed in an AP-MS experiment (Z)*

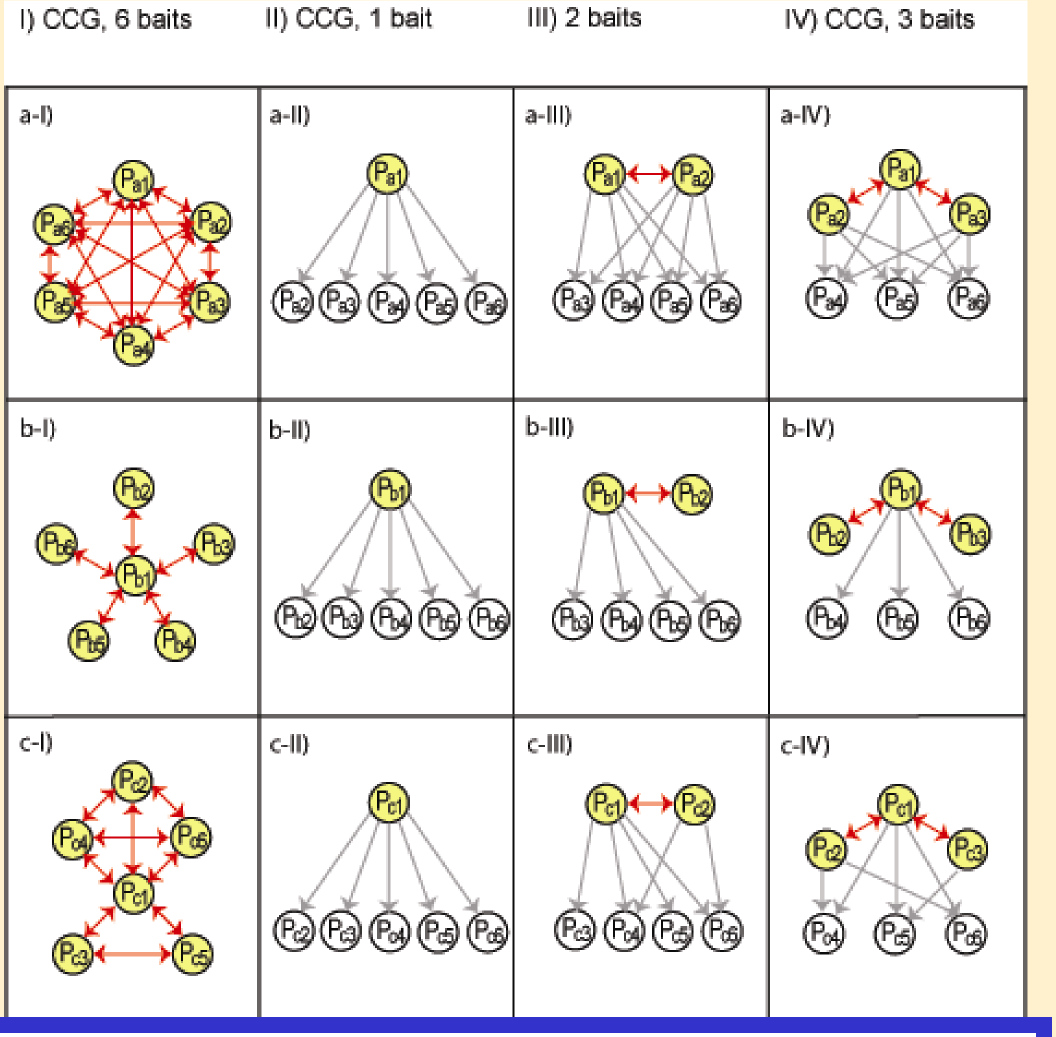
Y represents "ideal" complex comembership observations from perfectly sensitive and perfectly specific AP-MS technology. Y depends on the baits that are used in an experiment. Y is assayed by AP-MS technology.



3. Graph theoretic paradigm to allow for succinct expression of constructs involved

- *Relationship of complex membership (A) to complex comembership (Y) assayed in an AP-MS experiment (Z)*

Y represents "ideal" complex comembership observations from perfectly sensitive and perfectly specific AP-MS technology. Y depends on the baits that are used in an experiment. Y is assayed by AP-MS technology.



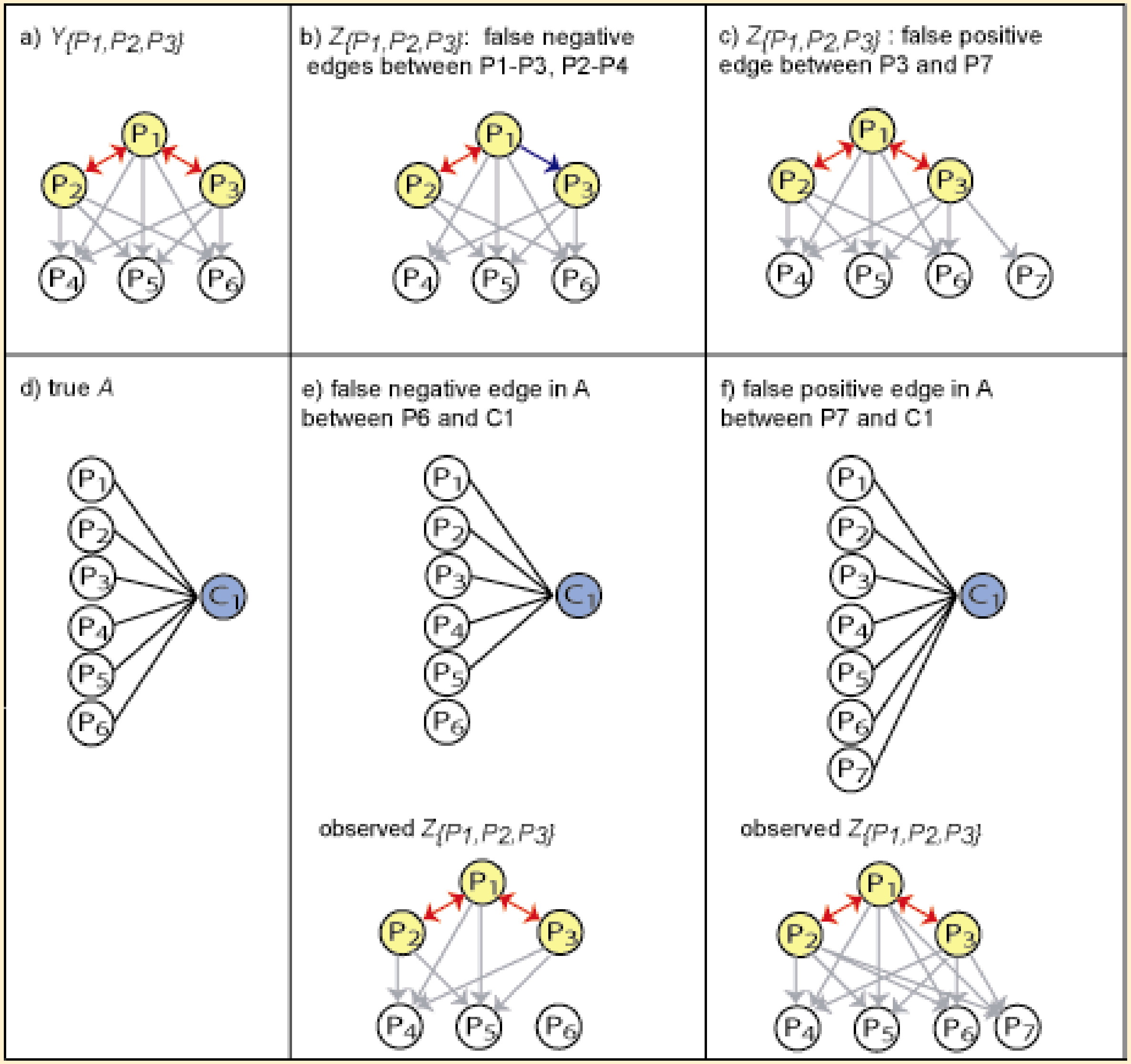
The Connection: Maximal BH-Complete Subgraphs

BH-Complete Subgraph: set of *n* bait nodes and *m* hit-only nodes for which all *n(n-1)+nm* directed edges exist

Maximal BH-Complete Subgraph: BH-complete subgraph that is not contained in any other complete subgraph

4. Statistical paradigm to allow for false positive and false negative observations

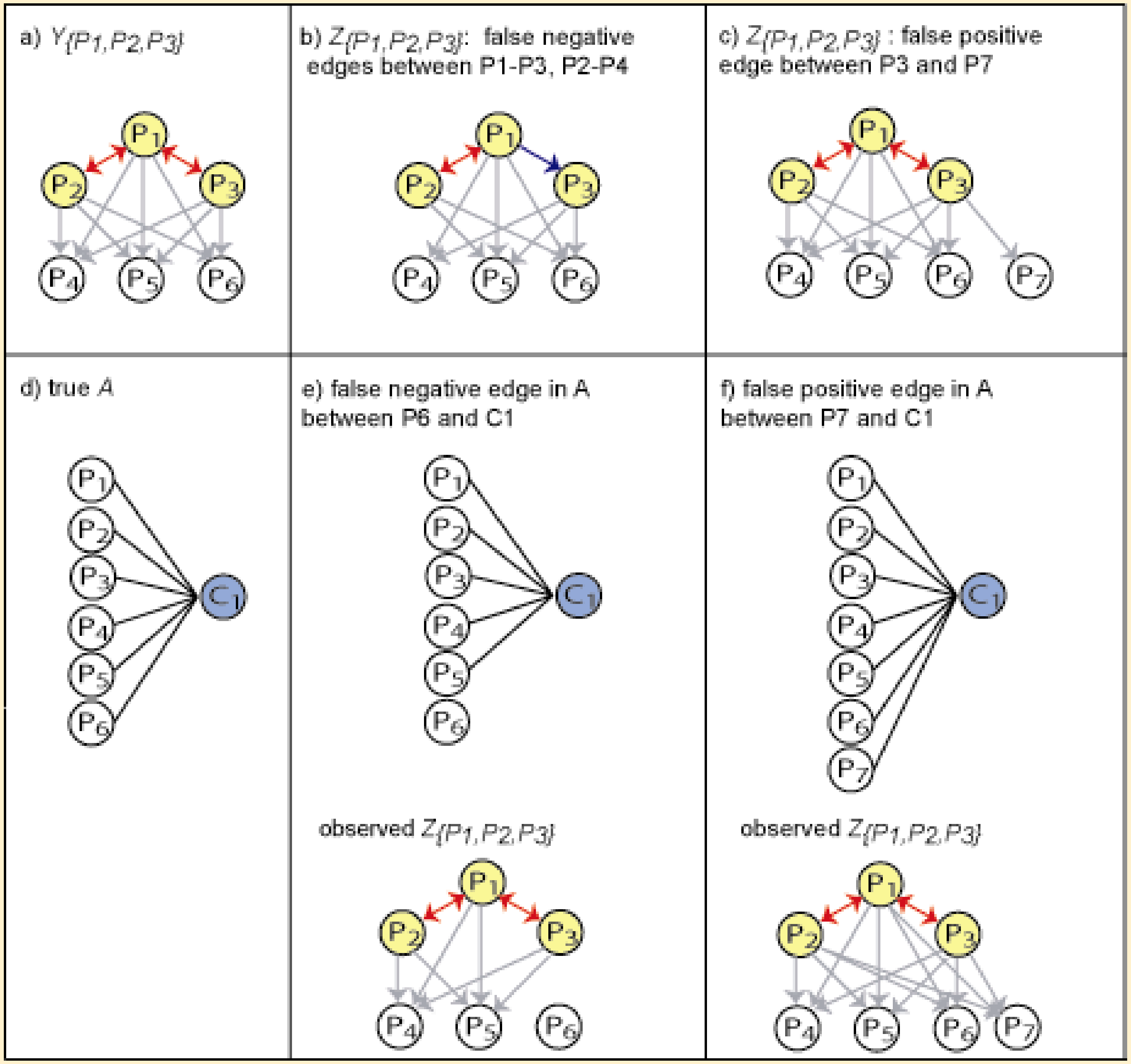
Z represents actual observations using AP-MS technology.


a) $Y_{\{P1,P2,P3\}}$
b) $Z_{\{P1,P2,P3\}}$: false negative edges between P1-P3, P2-P4
c) $Z_{\{P1,P2,P3\}}$: false positive edge between P3 and P7
d) true A
e) false negative edge in A between P6 and C1
f) false positive edge in A between P7 and C1
observed $Z_{\{P1,P2,P3\}}$
observed $Z_{\{P1,P2,P3\}}$
P_1
P_2
P_3
P_4
P_5
P_6
P_7
C_1

4. Statistical paradigm to allow for false positive and false negative observations

Z represents actual observations using AP-MS technology.

We will look for sets of proteins that form maximal BH-complete subgraphs with an allowance for false positive and false negative observations.



a) $Y_{\{P1,P2,P3\}}$

b) $Z_{\{P1,P2,P3\}}$: false negative edges between P1-P3, P2-P4

c) $Z_{\{P1,P2,P3\}}$: false positive edge between P3 and P7

d) true A

e) false negative edge in A between P6 and C1

observed $Z_{\{P1,P2,P3\}}$

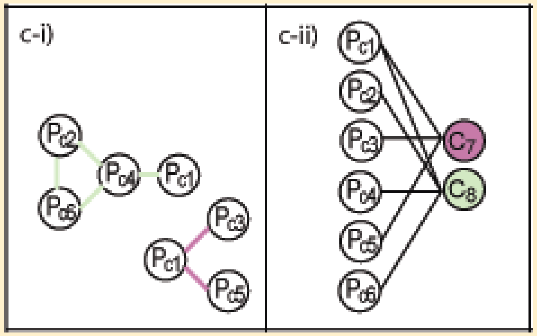
f) false positive edge in A between P7 and C1

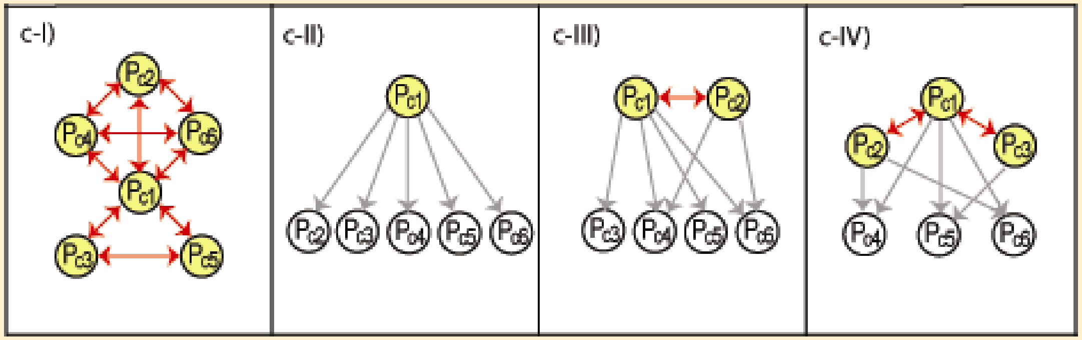
observed $Z_{\{P1,P2,P3\}}$

Our Goal

- for any (every) organism or tissue type we want to estimate the complex membership graph
- that is, the bipartite graph where one set of nodes are all proteins and the other are all complexes
- we are limited by the experimental data, experimental techniques and models

Graphs as Matrices





$$A = \begin{array}{c|cc|} & C_7 & C_8 \\ \hline P_{c1} & 1 & 1 \\ P_{c2} & 1 & 0 \\ P_{c3} & 0 & 1 \\ P_{c4} & 1 & 0 \\ P_{c5} & 0 & 1 \\ P_{c6} & 1 & 0 \\ \hline \end{array}$$

hits →

↓

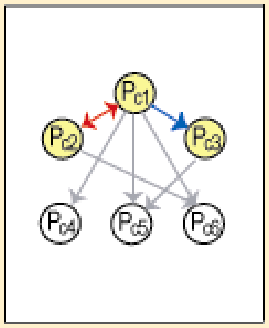
$$Y=A\otimes A' = \begin{array}{c|cccccc|} & P_{c1} & P_{c2} & P_{c3} & P_{c4} & P_{c5} & P_{c6} \\ \hline P_{c1} & 1 & 1 & 1 & 1 & 1 & 1 \\ P_{c2} & 1 & 1 & 0 & 1 & 0 & 1 \\ P_{c3} & 1 & 0 & 1 & 0 & 1 & 0 \\ P_{c4} & 1 & 1 & 0 & 1 & 0 & 1 \\ P_{c5} & 1 & 0 & 1 & 0 & 1 & 0 \\ P_{c6} & 1 & 1 & 0 & 1 & 0 & 1 \\ \hline \end{array}$$

Boolean Algebra:

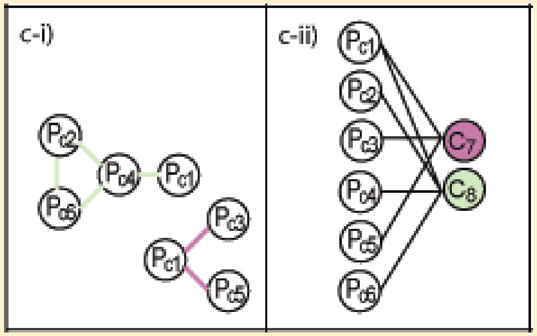
$0+0=0\bullet 1=1\bullet 0=0\bullet 0= 1^C= 0$

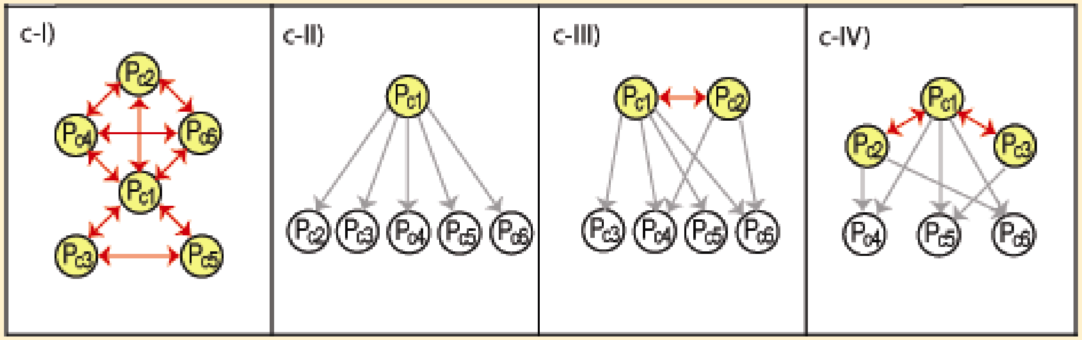
$0+1=1+0=1+1=1\bullet 1= 0^C= 1$

$$Z = \begin{array}{c|cccccc|} & P_{c1} & P_{c2} & P_{c3} & P_{c4} & P_{c5} & P_{c6} \\ \hline P_{c1} & 1 & 1 & 1 & 1 & 1 & 1 \\ P_{c2} & 1 & 1 & 0 & 0 & 0 & 1 \\ P_{c3} & 0 & 0 & 1 & 0 & 1 & 1 \\ \hline \end{array}$$



Graphs as Matrices





$$A = \begin{array}{c|cc|} & C_7 & C_8 \\ \hline P_{c1} & 1 & 1 \\ P_{c2} & 1 & 0 \\ P_{c3} & 0 & 1 \\ P_{c4} & 1 & 0 \\ P_{c5} & 0 & 1 \\ P_{c6} & 1 & 0 \\ \hline \end{array}$$

hits →

baits ↓

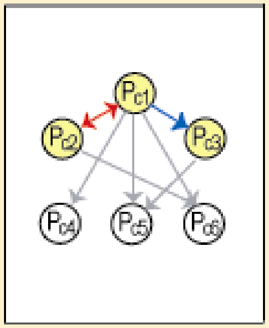
$$Y = A \otimes A' = \begin{array}{c|cccccc|} & P_{c1} & P_{c2} & P_{c3} & P_{c4} & P_{c5} & P_{c6} \\ \hline P_{c1} & 1 & 1 & 1 & 1 & 1 & 1 \\ P_{c2} & 1 & 1 & 0 & 1 & 0 & 1 \\ P_{c3} & 1 & 0 & 1 & 0 & 1 & 0 \\ P_{c4} & 1 & 1 & 0 & 1 & 0 & 1 \\ P_{c5} & 1 & 0 & 1 & 0 & 1 & 0 \\ P_{c6} & 1 & 1 & 0 & 1 & 0 & 1 \\ \hline \end{array}$$

Boolean Algebra:

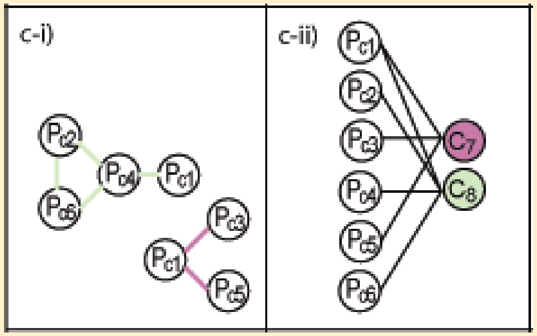
$0+0=0\bullet1=1\bullet0=0\bullet0= 1^{C}= 0$

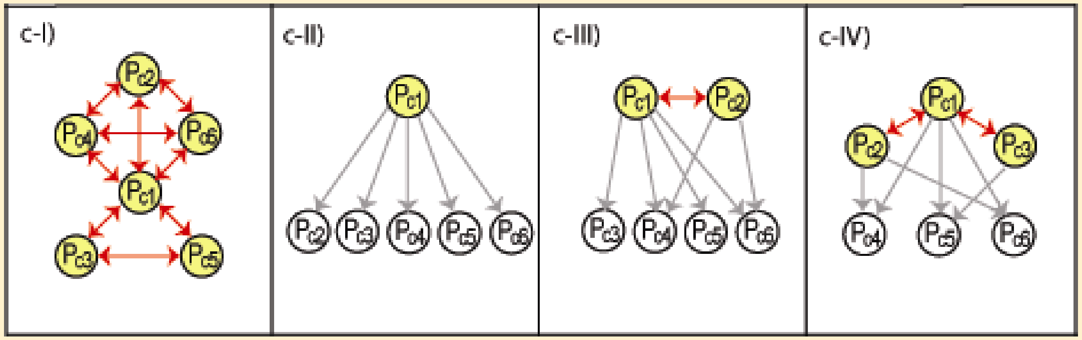
$0+1=1+0=1+1=1\bullet1= 0^{C}= 1$

$$Z = \begin{array}{c|cccccc|} & P_{c1} & P_{c2} & P_{c3} & P_{c4} & P_{c5} & P_{c6} \\ \hline P_{c1} & 1 & 1 & 1 & 1 & 1 & 1 \\ P_{c2} & 1 & 1 & 0 & 0 & 0 & 1 \\ P_{c3} & 0 & 0 & 1 & 0 & 1 & 1 \\ \hline \end{array}$$



Graphs as Matrices





$$A = \begin{array}{c|cc|} & C_7 & C_8 \\ \hline P_{c1} & 1 & 1 \\ P_{c2} & 1 & 0 \\ P_{c3} & 0 & 1 \\ P_{c4} & 1 & 0 \\ P_{c5} & 0 & 1 \\ P_{c6} & 1 & 0 \\ \hline \end{array}$$

hits →

baits ↓

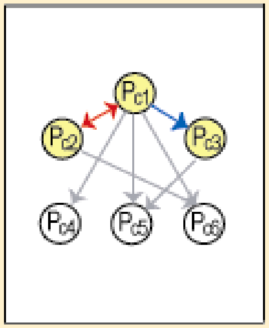
$$Y=A\otimes A' = \begin{array}{c|cccccc|} & P_{c1} & P_{c2} & P_{c3} & P_{c4} & P_{c5} & P_{c6} \\ \hline P_{c1} & 1 & 1 & 1 & 1 & 1 & 1 \\ \hline P_{c2} & 1 & 1 & 0 & 1 & 0 & 1 \\ P_{c3} & 1 & 0 & 1 & 0 & 1 & 0 \\ P_{c4} & 1 & 1 & 0 & 1 & 0 & 1 \\ P_{c5} & 1 & 0 & 1 & 0 & 1 & 0 \\ P_{c6} & 1 & 1 & 0 & 1 & 0 & 1 \\ \hline \end{array}$$

Boolean Algebra:

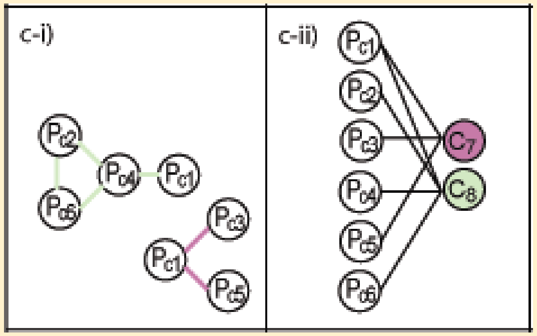
$0+0=0\bullet1=1\bullet0=0\bullet0= 1^C= 0$

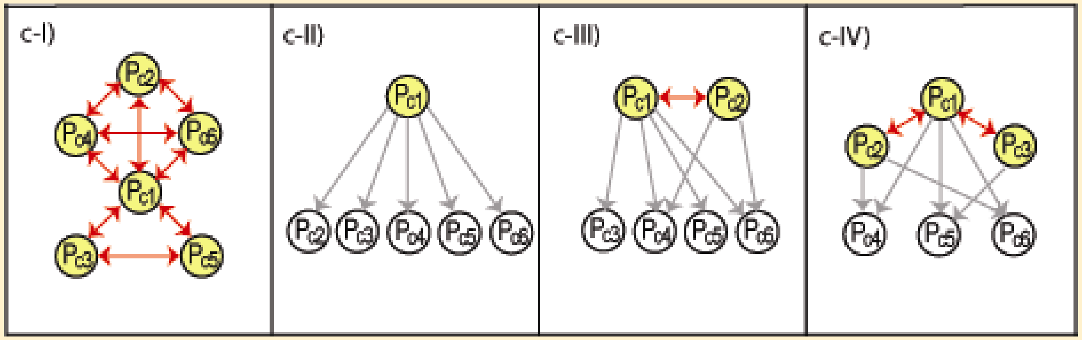
$0+1=1+0=1+1=1\bullet1= 0^C= 1$

$$Z = \begin{array}{c|cccccc|} & P_{c1} & P_{c2} & P_{c3} & P_{c4} & P_{c5} & P_{c6} \\ \hline P_{c1} & 1 & 1 & 1 & 1 & 1 & 1 \\ P_{c2} & 1 & 1 & 0 & 0 & 0 & 1 \\ P_{c3} & 0 & 0 & 1 & 0 & 1 & 1 \\ \hline \end{array}$$



Graphs as Matrices





$$A = \begin{array}{c|cc|} & C_7 & C_8 \\ \hline P_{c1} & 1 & 1 \\ P_{c2} & 1 & 0 \\ P_{c3} & 0 & 1 \\ P_{c4} & 1 & 0 \\ P_{c5} & 0 & 1 \\ P_{c6} & 1 & 0 \\ \hline \end{array}$$

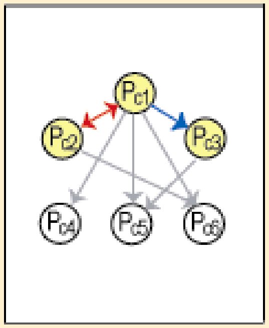
hits →

baits ↓

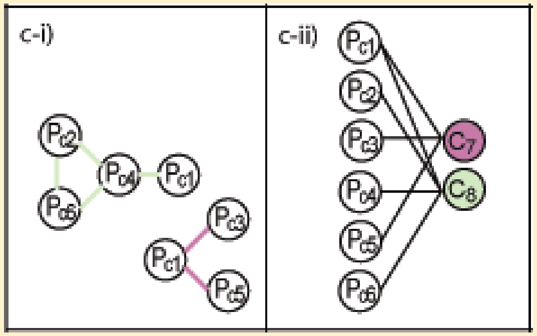
$$Y = A \otimes A' = \begin{array}{c|cccccc|} & P_{c1} & P_{c2} & P_{c3} & P_{c4} & P_{c5} & P_{c6} \\ \hline P_{c1} & 1 & 1 & 1 & 1 & 1 & 1 \\ P_{c2} & 1 & 1 & 0 & 1 & 0 & 1 \\ \hline P_{c3} & 1 & 0 & 1 & 0 & 1 & 0 \\ P_{c4} & 1 & 1 & 0 & 1 & 0 & 1 \\ P_{c5} & 1 & 0 & 1 & 0 & 1 & 0 \\ P_{c6} & 1 & 1 & 0 & 1 & 0 & 1 \\ \hline \end{array}$$

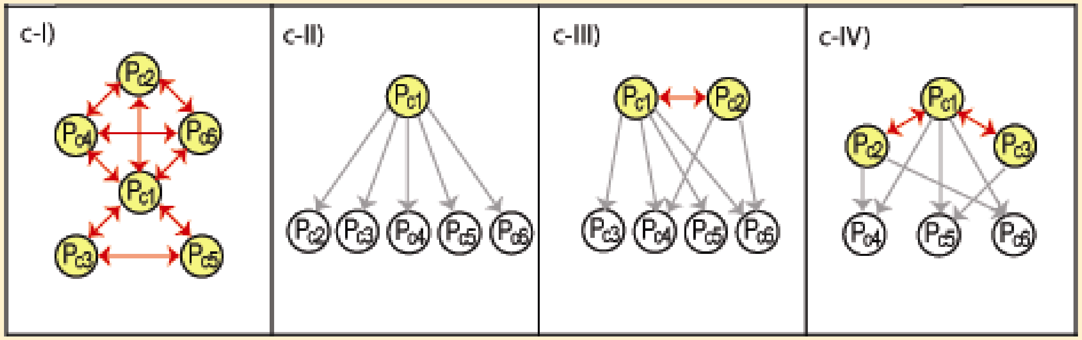
Boolean Algebra:
0+0=0•1=1•0=0•0= 1^{C}= 0
0+1=1+0=1+1=1•1= 0^{C}= 1

$$Z = \begin{array}{c|cccccc|} & P_{c1} & P_{c2} & P_{c3} & P_{c4} & P_{c5} & P_{c6} \\ \hline P_{c1} & 1 & 1 & 1 & 1 & 1 & 1 \\ P_{c2} & 1 & 1 & 0 & 0 & 0 & 1 \\ P_{c3} & 0 & 0 & 1 & 0 & 1 & 1 \\ \hline \end{array}$$



Graphs as Matrices





$$A = \begin{array}{c|cc|} & C_7 & C_8 \\ \hline P_{c1} & 1 & 1 \\ P_{c2} & 1 & 0 \\ P_{c3} & 0 & 1 \\ P_{c4} & 1 & 0 \\ P_{c5} & 0 & 1 \\ P_{c6} & 1 & 0 \\ \hline \end{array}$$

hits →

baits ↓

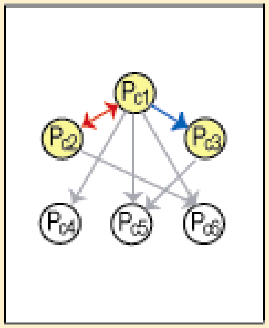
$$Y = A \otimes A' = \begin{array}{c|cccccc|} & P_{c1} & P_{c2} & P_{c3} & P_{c4} & P_{c5} & P_{c6} \\ \hline P_{c1} & 1 & 1 & 1 & 1 & 1 & 1 \\ P_{c2} & 1 & 1 & 0 & 1 & 0 & 1 \\ P_{c3} & 1 & 0 & 1 & 0 & 1 & 0 \\ \hline P_{c4} & 1 & 1 & 0 & 1 & 0 & 1 \\ P_{c5} & 1 & 0 & 1 & 0 & 1 & 0 \\ P_{c6} & 1 & 1 & 0 & 1 & 0 & 1 \\ \hline \end{array}$$

Boolean Algebra:

$0+0=0\bullet1=1\bullet0=0\bullet0= 1^C= 0$

$0+1=1+0=1+1=1\bullet1= 0^C= 1$

$$Z = \begin{array}{c|cccccc|} & P_{c1} & P_{c2} & P_{c3} & P_{c4} & P_{c5} & P_{c6} \\ \hline P_{c1} & 1 & 1 & 1 & 1 & 1 & 1 \\ P_{c2} & 1 & 1 & 0 & 0 & 0 & 1 \\ P_{c3} & 0 & 0 & 1 & 0 & 1 & 1 \\ \hline \end{array}$$



In summary…

$$A \xrightarrow[Y=A\otimes A']{} Y \xrightarrow[\substack{\text{AP}-\text{MS data for } N \text{ bait proteins} \\ \text{and } M \text{ hit-only proteins}}]{} Z$$

$$Z \xrightarrow[\text{estimation algorithm}]{} (\hat{Y}, \hat{A})$$

We start with an initial estimate for *A*, and then refine that estimate according to a two component probability measure:

$$P(Z|A,\mu,\alpha)=L(Z|Y=A\otimes A',\mu,\alpha)C\,(Z|A,\mu,\alpha)$$

$$P(Z|A,\mu,\alpha)=L(Z|Y=A\otimes A',\mu,\alpha)C\ (Z|A,\mu,\alpha)$$

L is the usual likelihood for independent Bernoulli observations of the existence of an edge under a logistic regression model with user-specified values of *μ* and *α*.

$$L(Z\,|\,A\otimes A',\mu,\alpha)=\prod_{i=1}^{N}\prod_{j=1,j\neq i}^{N}p_{ij}{}^{Z_{ij}}(1-p_{ij})^{(1-Z_{ij})}\prod_{l=1}^{N}\prod_{m=N+1}^{N+M}p_{lm}{}^{Z_{lm}}(1-p_{lm})^{(1-Z_{lm})}$$

doubly tested edges singly tested edges

$$p_{ij}=\Pr(Z_{ij}=1\,|\,\mu,\alpha,Y_{ij}),\quad \text{and}\quad \log\left(\frac{p_{ij}}{1-p_{ij}}\right)=\mu+\alpha Y_{ij}$$

$$\text{sensitivity}=\frac{e^{\mu}}{1+e^{\mu}},\qquad \text{specificity}=\frac{e^{\mu+\alpha}}{1+e^{\mu+\alpha}}$$

$$P(Z|A,\mu,\alpha)=L(Z|Y=A\otimes A',\mu,\alpha)C\ (Z|A,\mu,\alpha)$$

L is the usual likelihood for independent Bernoulli observations of the existence of an edge under a logistic regression model with user-specified values of *μ* and *α*.

$$L(Z\,|\,A\otimes A',\mu,\alpha)=\underbrace{\prod_{i=1}^{N}\prod_{j=1,j\neq i}^{N}{p_{ij}}^{Z_{ij}}(1-p_{ij})^{(1-Z_{ij})}}_{\text{doubly tested edges}}\underbrace{\prod_{l=1}^{N}\prod_{m=N+1}^{N+M}{p_{lm}}^{Z_{lm}}(1-p_{lm})^{(1-Z_{lm})}}_{\text{singly tested edges}}$$

$$p_{ij}=\Pr(Z_{ij}=1\,|\,\mu,\alpha,Y_{ij}),\quad \text{and}\quad \log\left(\frac{p_{ij}}{1-p_{ij}}\right)=\mu+\alpha Y_{ij}$$

$$\text{sensitivity}=\frac{e^{\mu}}{1+e^{\mu}},\qquad \text{specificity}=\frac{e^{\mu+\alpha}}{1+e^{\mu+\alpha}}$$

Using L, we can estimate Y_{ij}= 0 or 1 for i=1,…,N and j=1,…,N+M. For i=j, Y_{ij}=Y_{ji}.

$$P(Z|A,\mu,\alpha)=L(Z|Y=A\otimes A',\mu,\alpha)C\,(Z|A,\mu,\alpha)$$

Assumptions for μ and α in our analyses:

1) $\Pr(Z_{ij}=0|\ \mu,\alpha,Y_{ij}=0)>.5$ and $\Pr(Z_{ij}=1|\ \mu,\alpha,Y_{ij}=1)>.5$

-sensitivity and specificity are greater than .5

2) $\Pr(Z_{ij}=0|\ \mu,\alpha,Y_{ij}=1)>\Pr(Z_{ij}=1|\ \mu,\alpha,Y_{ij}=0)$

-false negative probability is greater than false positive probability

Under these assumptions for μ and α, L is easily maximized.

For singly tested bait-hit pairs, $\boxed{\hat{Y}_{ij}=Z_{ij}.}$

For doubly tested bait-bait pairs, $\boxed{(\hat{Y}_{ij},\hat{Y}_{ji})=\max(Z_{ij},Z_{ji}).}$

$$P(Z|A,\mu,\alpha)=L(Z|Y=A\otimes A',\mu,\alpha)C\ (Z|A,\mu,\alpha)$$

Assumptions for μ and α in our analyses:

1) $\Pr(Z_{ij}=0|\ \mu,\alpha,Y_{ij}=0)>.5$ and $\Pr(Z_{ij}=1|\ \mu,\alpha,Y_{ij}=1)>.5$
   -sensitivity and specificity are greater than .5

2) $\Pr(Z_{ij}=0|\ \mu,\alpha,Y_{ij}=1)>\Pr(Z_{ij}=1|\ \mu,\alpha,Y_{ij}=0)$
   -false negative probability is greater than false positive probability

Under these assumptions for μ and α, L is easily maximized.

For singly tested bait-hit pairs, $\boxed{\hat{Y}_{ij}=Z_{ij}.}$

For doubly tested bait-bait pairs, $\boxed{(\hat{Y}_{ij},\hat{Y}_{ji})=\max(Z_{ij},Z_{ji}).}$

We have an estimate for Y, but our goal is to estimate A.
We use the transformation $Y=A\otimes A'$ and maximal BH-complete subgraphs.

Given Y, What is A? Identifiability

$$Y=A\otimes A'$$

Y is uniquely determined by *A*,
but *A* is not uniquely determined by *Y*.

One Trimer

$$A = \begin{array}{c|c|} & C_1 \\ \hline P_1 & 1 \\ P_2 & 1 \\ P_3 & 1 \\ \hline \end{array}$$

One Trimer with a Dimer Subcomplex

$$A = \begin{array}{c|cc|} & C_1 & C_2 \\ \hline P_1 & 1 & 1 \\ P_2 & 1 & 1 \\ P_3 & 1 & 0 \\ \hline \end{array}$$

Three Dimers

$$A = \begin{array}{c|ccc|} & C_1 & C_2 & C_3 \\ \hline P_1 & 1 & 1 & 0 \\ P_2 & 1 & 0 & 1 \\ P_3 & 0 & 1 & 1 \\ \hline \end{array}$$

Identical *Y*

$$Y=A\otimes A' = \begin{array}{c|ccc|} & P_1 & P_2 & P_3 \\ \hline P_1 & 1 & 1 & 1 \\ P_2 & 1 & 1 & 1 \\ P_3 & 1 & 1 & 1 \\ \hline \end{array}$$

Given Y, What is A? Identifiability

$$Y = A \otimes A'$$

Y is uniquely determined by *A*,
but *A* is not uniquely determined by *Y*.

One Trimer

$$A = \begin{array}{c|c|} & C_1 \\ \hline P_1 & 1 \\ P_2 & 1 \\ P_3 & 1 \\ \hline \end{array}$$

One Trimer with a Dimer Subcomplex

$$A = \begin{array}{c|cc|} & C_1 & C_2 \\ \hline P_1 & 1 & 1 \\ P_2 & 1 & 1 \\ P_3 & 1 & 0 \\ \hline \end{array}$$

Three Dimers

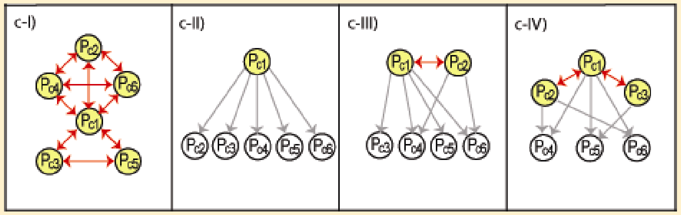
$$A = \begin{array}{c|ccc|} & C_1 & C_2 & C_3 \\ \hline P_1 & 1 & 1 & 0 \\ P_2 & 1 & 0 & 1 \\ P_3 & 0 & 1 & 1 \\ \hline \end{array}$$

Identical *Y*

$$Y = A \otimes A' = \begin{array}{c|ccc|} & P_1 & P_2 & P_3 \\ \hline P_1 & 1 & 1 & 1 \\ P_2 & 1 & 1 & 1 \\ P_3 & 1 & 1 & 1 \\ \hline \end{array}$$

A is identifiable if it assumed to consist of maximal subgraphs of *Y*. I.e., given the *Y* above, we would find the "one trimer" version of *A*.

Initial Estimate of *A*



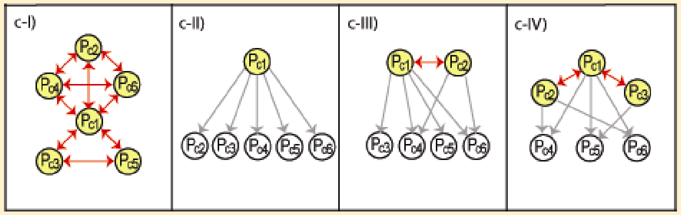
hits →

$$Y=A\otimes A' = \begin{array}{c|cccccc} & P_{c1} & P_{c2} & P_{c3} & P_{c4} & P_{c5} & P_{c6} \\ \hline P_{c1} & 1 & 1 & 1 & 1 & 1 & 1 \\ P_{c2} & 1 & 1 & 0 & 1 & 0 & 1 \\ P_{c3} & 1 & 0 & 1 & 0 & 1 & 0 \\ P_{c4} & 1 & 1 & 0 & 1 & 0 & 1 \\ P_{c5} & 1 & 0 & 1 & 0 & 1 & 0 \\ P_{c6} & 1 & 1 & 0 & 1 & 0 & 1 \end{array}$$

$$Y= \begin{array}{c|cccccc} & P_{c1} & P_{c2} & P_{c3} & P_{c4} & P_{c5} & P_{c6} \\ \hline P_{c1} & 1 & 1 & 1 & 1 & 1 & 1 \\ P_{c2} & 1 & 1 & 0 & 1 & 0 & 1 \\ P_{c3} & 1 & 0 & 1 & 1 & 1 & 1 \\ P_{c4} & 1 & 1 & 1 & 1 & 1 & 1 \\ P_{c5} & 1 & 0 & 1 & 1 & 1 & 1 \\ P_{c6} & 1 & 1 & 1 & 1 & 1 & 1 \end{array} \Rightarrow A= \begin{array}{c|cc} & C_1 & C_2 \\ \hline P_{c1} & 1 & 1 \\ P_{c2} & 1 & 0 \\ P_{c3} & 0 & 1 \\ P_{c4} & 1 & 1 \\ P_{c5} & 0 & 1 \\ P_{c6} & 1 & 1 \end{array}$$

$$Y= \begin{array}{c|cccccc} & P_{c1} & P_{c2} & P_{c3} & P_{c4} & P_{c5} & P_{c6} \\ \hline P_{c1} & 1 & 1 & 1 & 1 & 1 & 1 \\ P_{c2} & 1 & 1 & 0 & 1 & 0 & 1 \\ P_{c3} & 1 & 0 & 1 & 0 & 1 & 0 \\ P_{c4} & 1 & 1 & 0 & 1 & 1 & 1 \\ P_{c5} & 1 & 0 & 1 & 1 & 1 & 1 \\ P_{c6} & 1 & 1 & 0 & 1 & 1 & 1 \end{array} \Rightarrow A= \begin{array}{c|ccc} & C_1 & C_2 & C_3 \\ \hline P_{c1} & 1 & 1 & 1 \\ P_{c2} & 1 & 0 & 0 \\ P_{c3} & 0 & 1 & 0 \\ P_{c4} & 1 & 0 & 1 \\ P_{c5} & 0 & 1 & 1 \\ P_{c6} & 1 & 0 & 1 \end{array}$$

Initial Estimate of *A*



hits →

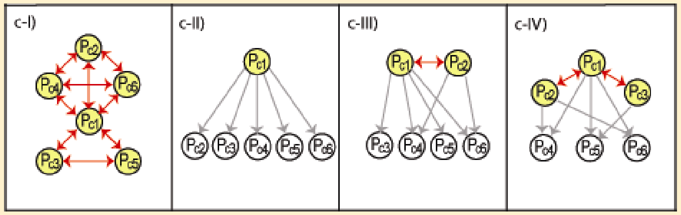
baits ↓

$$Y = A \otimes A' = \begin{array}{c|cccccc} & P_{c1} & P_{c2} & P_{c3} & P_{c4} & P_{c5} & P_{c6} \\ \hline P_{c1} & 1 & 1 & 1 & 1 & 1 & 1 \\ \hline P_{c2} & 1 & 1 & 0 & 1 & 0 & 1 \\ P_{c3} & 1 & 0 & 1 & 0 & 1 & 0 \\ P_{c4} & 1 & 1 & 0 & 1 & 0 & 1 \\ P_{c5} & 1 & 0 & 1 & 0 & 1 & 0 \\ P_{c6} & 1 & 1 & 0 & 1 & 0 & 1 \end{array}$$

$$Y = \begin{array}{c|cc|cccc} & P_{c1} & P_{c2} & P_{c3} & P_{c4} & P_{c5} & P_{c6} \\ \hline P_{c1} & 1 & 1 & 1 & 1 & 1 & 1 \\ P_{c2} & 1 & 1 & 0 & 1 & 0 & 1 \\ \hline P_{c3} & 1 & 0 & 1 & 1 & 1 & 1 \\ P_{c4} & 1 & 1 & 1 & 1 & 1 & 1 \\ P_{c5} & 1 & 0 & 1 & 1 & 1 & 1 \\ P_{c6} & 1 & 1 & 1 & 1 & 1 & 1 \end{array} \Rightarrow A = \begin{array}{c|cc} & C_1 & C_2 \\ \hline P_{c1} & 1 & 1 \\ P_{c2} & 1 & 0 \\ P_{c3} & 0 & 1 \\ P_{c4} & 1 & 1 \\ P_{c5} & 0 & 1 \\ P_{c6} & 1 & 1 \end{array}$$

$$Y = \begin{array}{c|ccc|ccc} & P_{c1} & P_{c2} & P_{c3} & P_{c4} & P_{c5} & P_{c6} \\ \hline P_{c1} & 1 & 1 & 1 & 1 & 1 & 1 \\ P_{c2} & 1 & 1 & 0 & 1 & 0 & 1 \\ P_{c3} & 1 & 0 & 1 & 0 & 1 & 0 \\ \hline P_{c4} & 1 & 1 & 0 & 1 & 1 & 1 \\ P_{c5} & 1 & 0 & 1 & 1 & 1 & 1 \\ P_{c6} & 1 & 1 & 0 & 1 & 1 & 1 \end{array} \Rightarrow A = \begin{array}{c|ccc} & C_1 & C_2 & C_3 \\ \hline P_{c1} & 1 & 1 & 1 \\ P_{c2} & 1 & 0 & 0 \\ P_{c3} & 0 & 1 & 0 \\ P_{c4} & 1 & 0 & 1 \\ P_{c5} & 0 & 1 & 1 \\ P_{c6} & 1 & 0 & 1 \end{array}$$

Initial Estimate of *A*



hits →

baits ↓

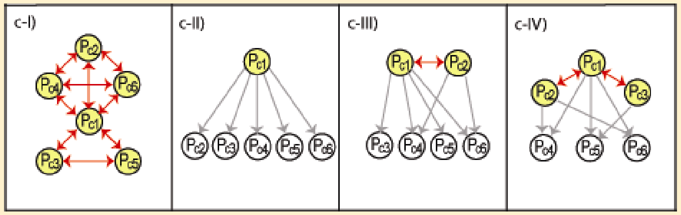
$$Y=A\otimes A' = \begin{array}{c|cccccc} & P_{c1} & P_{c2} & P_{c3} & P_{c4} & P_{c5} & P_{c6} \\ \hline P_{c1} & 1 & 1 & 1 & 1 & 1 & 1 \\ \hline P_{c2} & 1 & 1 & 0 & 1 & 0 & 1 \\ P_{c3} & 1 & 0 & 1 & 0 & 1 & 0 \\ P_{c4} & 1 & 1 & 0 & 1 & 0 & 1 \\ P_{c5} & 1 & 0 & 1 & 0 & 1 & 0 \\ P_{c6} & 1 & 1 & 0 & 1 & 0 & 1 \end{array}$$

$$Y= \begin{array}{c|c|ccccc} & P_{c1} & P_{c2} & P_{c3} & P_{c4} & P_{c5} & P_{c6} \\ \hline P_{c1} & 1 & 1 & 1 & 1 & 1 & 1 \\ \hline P_{c2} & 1 & 1 & 1 & 1 & 1 & 1 \\ P_{c3} & 1 & 1 & 1 & 1 & 1 & 1 \\ P_{c4} & 1 & 1 & 1 & 1 & 1 & 1 \\ P_{c5} & 1 & 1 & 1 & 1 & 1 & 1 \\ P_{c6} & 1 & 1 & 1 & 1 & 1 & 1 \end{array} \Rightarrow A= \begin{array}{c|c} & C_1 \\ \hline P_{c1} & 1 \\ P_{c2} & 1 \\ P_{c3} & 1 \\ P_{c4} & 1 \\ P_{c5} & 1 \\ P_{c6} & 1 \end{array}$$

$$Y= \begin{array}{c|cc|cccc} & P_{c1} & P_{c2} & P_{c3} & P_{c4} & P_{c5} & P_{c6} \\ \hline P_{c1} & 1 & 1 & 1 & 1 & 1 & 1 \\ P_{c2} & 1 & 1 & 0 & 1 & 0 & 1 \\ \hline P_{c3} & 1 & 0 & 1 & 1 & 1 & 1 \\ P_{c4} & 1 & 1 & 1 & 1 & 1 & 1 \\ P_{c5} & 1 & 0 & 1 & 1 & 1 & 1 \\ P_{c6} & 1 & 1 & 1 & 1 & 1 & 1 \end{array} \Rightarrow A= \begin{array}{c|cc} & C_1 & C_2 \\ \hline P_{c1} & 1 & 1 \\ P_{c2} & 1 & 0 \\ P_{c3} & 0 & 1 \\ P_{c4} & 1 & 1 \\ P_{c5} & 0 & 1 \\ P_{c6} & 1 & 1 \end{array}$$

$$Y= \begin{array}{c|ccc|ccc} & P_{c1} & P_{c2} & P_{c3} & P_{c4} & P_{c5} & P_{c6} \\ \hline P_{c1} & 1 & 1 & 1 & 1 & 1 & 1 \\ P_{c2} & 1 & 1 & 0 & 1 & 0 & 1 \\ P_{c3} & 1 & 0 & 1 & 0 & 1 & 0 \\ \hline P_{c4} & 1 & 1 & 0 & 1 & 1 & 1 \\ P_{c5} & 1 & 0 & 1 & 1 & 1 & 1 \\ P_{c6} & 1 & 1 & 0 & 1 & 1 & 1 \end{array} \Rightarrow A= \begin{array}{c|ccc} & C_1 & C_2 & C_3 \\ \hline P_{c1} & 1 & 1 & 1 \\ P_{c2} & 1 & 0 & 0 \\ P_{c3} & 0 & 1 & 0 \\ P_{c4} & 1 & 0 & 1 \\ P_{c5} & 0 & 1 & 1 \\ P_{c6} & 1 & 0 & 1 \end{array}$$

Initial Estimate of *A*



hits →

baits ↓

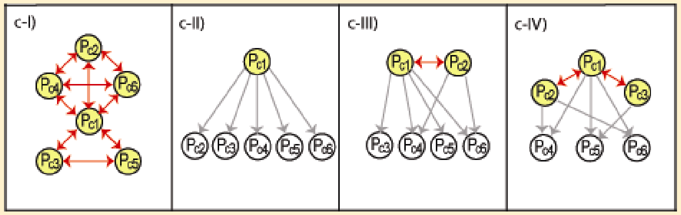
$$Y=A\otimes A' = \begin{array}{c|cccccc} & P_{c1} & P_{c2} & P_{c3} & P_{c4} & P_{c5} & P_{c6} \\ \hline P_{c1} & 1 & 1 & 1 & 1 & 1 & 1 \\ P_{c2} & 1 & 1 & 0 & 1 & 0 & 1 \\ P_{c3} & 1 & 0 & 1 & 0 & 1 & 0 \\ P_{c4} & 1 & 1 & 0 & 1 & 0 & 1 \\ P_{c5} & 1 & 0 & 1 & 0 & 1 & 0 \\ P_{c6} & 1 & 1 & 0 & 1 & 0 & 1 \end{array}$$

$$Y= \begin{array}{c|cccccc} & P_{c1} & P_{c2} & P_{c3} & P_{c4} & P_{c5} & P_{c6} \\ \hline P_{c1} & 1 & 1 & 1 & 1 & 1 & 1 \\ P_{c2} & 1 & 1 & 1 & 1 & 1 & 1 \\ P_{c3} & 1 & 1 & 1 & 1 & 1 & 1 \\ P_{c4} & 1 & 1 & 1 & 1 & 1 & 1 \\ P_{c5} & 1 & 1 & 1 & 1 & 1 & 1 \\ P_{c6} & 1 & 1 & 1 & 1 & 1 & 1 \end{array} \Rrightarrow A= \begin{array}{c|c} & C_1 \\ \hline P_{c1} & 1 \\ P_{c2} & 1 \\ P_{c3} & 1 \\ P_{c4} & 1 \\ P_{c5} & 1 \\ P_{c6} & 1 \end{array}$$

$$Y= \begin{array}{c|cccccc} & P_{c1} & P_{c2} & P_{c3} & P_{c4} & P_{c5} & P_{c6} \\ \hline P_{c1} & 1 & 1 & 1 & 1 & 1 & 1 \\ P_{c2} & 1 & 1 & 0 & 1 & 0 & 1 \\ P_{c3} & 1 & 0 & 1 & 1 & 1 & 1 \\ P_{c4} & 1 & 1 & 1 & 1 & 1 & 1 \\ P_{c5} & 1 & 0 & 1 & 1 & 1 & 1 \\ P_{c6} & 1 & 1 & 1 & 1 & 1 & 1 \end{array} \Rrightarrow A= \begin{array}{c|cc} & C_1 & C_2 \\ \hline P_{c1} & 1 & 1 \\ P_{c2} & 1 & 0 \\ P_{c3} & 0 & 1 \\ P_{c4} & 1 & 1 \\ P_{c5} & 0 & 1 \\ P_{c6} & 1 & 1 \end{array}$$

$$Y= \begin{array}{c|cccccc} & P_{c1} & P_{c2} & P_{c3} & P_{c4} & P_{c5} & P_{c6} \\ \hline P_{c1} & 1 & 1 & 1 & 1 & 1 & 1 \\ P_{c2} & 1 & 1 & 0 & 1 & 0 & 1 \\ P_{c3} & 1 & 0 & 1 & 0 & 1 & 0 \\ P_{c4} & 1 & 1 & 0 & 1 & 1 & 1 \\ P_{c5} & 1 & 0 & 1 & 1 & 1 & 1 \\ P_{c6} & 1 & 1 & 0 & 1 & 1 & 1 \end{array} \Rrightarrow A= \begin{array}{c|ccc} & C_1 & C_2 & C_3 \\ \hline P_{c1} & 1 & 1 & 1 \\ P_{c2} & 1 & 0 & 0 \\ P_{c3} & 0 & 1 & 0 \\ P_{c4} & 1 & 0 & 1 \\ P_{c5} & 0 & 1 & 1 \\ P_{c6} & 1 & 0 & 1 \end{array}$$

Initial Estimate of *A*


c-I)
P_{c2} P_{c4} P_{c6} P_{c1} P_{c3} P_{c5}
c-II)
P_{c1}
P_{c2} P_{c3} P_{c4} P_{c5} P_{c6}
c-III)
P_{c1} P_{c2}
P_{c3} P_{c4} P_{c5} P_{c6}
c-IV)
P_{c1} P_{c2} P_{c3}
P_{c4} P_{c5} P_{c6}

hits →

baits ↓

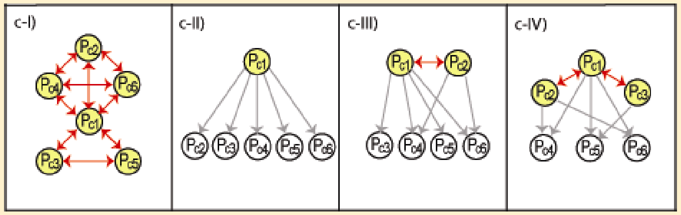
$$Y=A\otimes A' = \begin{array}{c|cccccc} & P_{c1} & P_{c2} & P_{c3} & P_{c4} & P_{c5} & P_{c6} \\ \hline P_{c1} & 1 & 1 & 1 & 1 & 1 & 1 \\ P_{c2} & 1 & 1 & 0 & 1 & 0 & 1 \\ P_{c3} & 1 & 0 & 1 & 0 & 1 & 0 \\ P_{c4} & 1 & 1 & 0 & 1 & 0 & 1 \\ P_{c5} & 1 & 0 & 1 & 0 & 1 & 0 \\ P_{c6} & 1 & 1 & 0 & 1 & 0 & 1 \end{array}$$

$$Y= \begin{array}{c|cccccc} & P_{c1} & P_{c2} & P_{c3} & P_{c4} & P_{c5} & P_{c6} \\ \hline P_{c1} & 1 & 1 & 1 & 1 & 1 & 1 \\ P_{c2} & 1 & 1 & 1 & 1 & 1 & 1 \\ P_{c3} & 1 & 1 & 1 & 1 & 1 & 1 \\ P_{c4} & 1 & 1 & 1 & 1 & 1 & 1 \\ P_{c5} & 1 & 1 & 1 & 1 & 1 & 1 \\ P_{c6} & 1 & 1 & 1 & 1 & 1 & 1 \end{array} \Rightarrow A= \begin{array}{c|c} & C_1 \\ \hline P_{c1} & 1 \\ P_{c2} & 1 \\ P_{c3} & 1 \\ P_{c4} & 1 \\ P_{c5} & 1 \\ P_{c6} & 1 \end{array}$$

$$Y= \begin{array}{c|cccccc} & P_{c1} & P_{c2} & P_{c3} & P_{c4} & P_{c5} & P_{c6} \\ \hline P_{c1} & 1 & 1 & 1 & 1 & 1 & 1 \\ P_{c2} & 1 & 1 & 0 & 1 & 0 & 1 \\ P_{c3} & 1 & 0 & 1 & 1 & 1 & 1 \\ P_{c4} & 1 & 1 & 1 & 1 & 1 & 1 \\ P_{c5} & 1 & 0 & 1 & 1 & 1 & 1 \\ P_{c6} & 1 & 1 & 1 & 1 & 1 & 1 \end{array} \Rightarrow A= \begin{array}{c|cc} & C_1 & C_2 \\ \hline P_{c1} & 1 & 1 \\ P_{c2} & 1 & 0 \\ P_{c3} & 0 & 1 \\ P_{c4} & 1 & 1 \\ P_{c5} & 0 & 1 \\ P_{c6} & 1 & 1 \end{array}$$

$$Y= \begin{array}{c|cccccc} & P_{c1} & P_{c2} & P_{c3} & P_{c4} & P_{c5} & P_{c6} \\ \hline P_{c1} & 1 & 1 & 1 & 1 & 1 & 1 \\ P_{c2} & 1 & 1 & 0 & 1 & 0 & 1 \\ P_{c3} & 1 & 0 & 1 & 0 & 1 & 0 \\ P_{c4} & 1 & 1 & 0 & 1 & 1 & 1 \\ P_{c5} & 1 & 0 & 1 & 1 & 1 & 1 \\ P_{c6} & 1 & 1 & 0 & 1 & 1 & 1 \end{array} \Rightarrow A= \begin{array}{c|ccc} & C_1 & C_2 & C_3 \\ \hline P_{c1} & 1 & 1 & 1 \\ P_{c2} & 1 & 0 & 0 \\ P_{c3} & 0 & 1 & 0 \\ P_{c4} & 1 & 0 & 1 \\ P_{c5} & 0 & 1 & 1 \\ P_{c6} & 1 & 0 & 1 \end{array}$$

Initial Estimate of *A*



hits →

baits ↓

$$Y=A\otimes A' = \begin{array}{ccccccc} & P_{c1} & P_{c2} & P_{c3} & P_{c4} & P_{c5} & P_{c6} \\ P_{c1} & 1 & 1 & 1 & 1 & 1 & 1 \\ P_{c2} & 1 & 1 & 0 & 1 & 0 & 1 \\ P_{c3} & 1 & 0 & 1 & 0 & 1 & 0 \\ P_{c4} & 1 & 1 & 0 & 1 & 0 & 1 \\ P_{c5} & 1 & 0 & 1 & 0 & 1 & 0 \\ P_{c6} & 1 & 1 & 0 & 1 & 0 & 1 \end{array}$$

$$Y= \begin{array}{ccccccc} & P_{c1} & P_{c2} & P_{c3} & P_{c4} & P_{c5} & P_{c6} \\ P_{c1} & 1 & 1 & 1 & 1 & 1 & 1 \\ P_{c2} & 1 & 1 & 1 & 1 & 1 & 1 \\ P_{c3} & 1 & 1 & 1 & 1 & 1 & 1 \\ P_{c4} & 1 & 1 & 1 & 1 & 1 & 1 \\ P_{c5} & 1 & 1 & 1 & 1 & 1 & 1 \\ P_{c6} & 1 & 1 & 1 & 1 & 1 & 1 \end{array} \Rightarrow A= \begin{array}{cc} & C_1 \\ P_{c1} & 1 \\ P_{c2} & 1 \\ P_{c3} & 1 \\ P_{c4} & 1 \\ P_{c5} & 1 \\ P_{c6} & 1 \end{array}$$

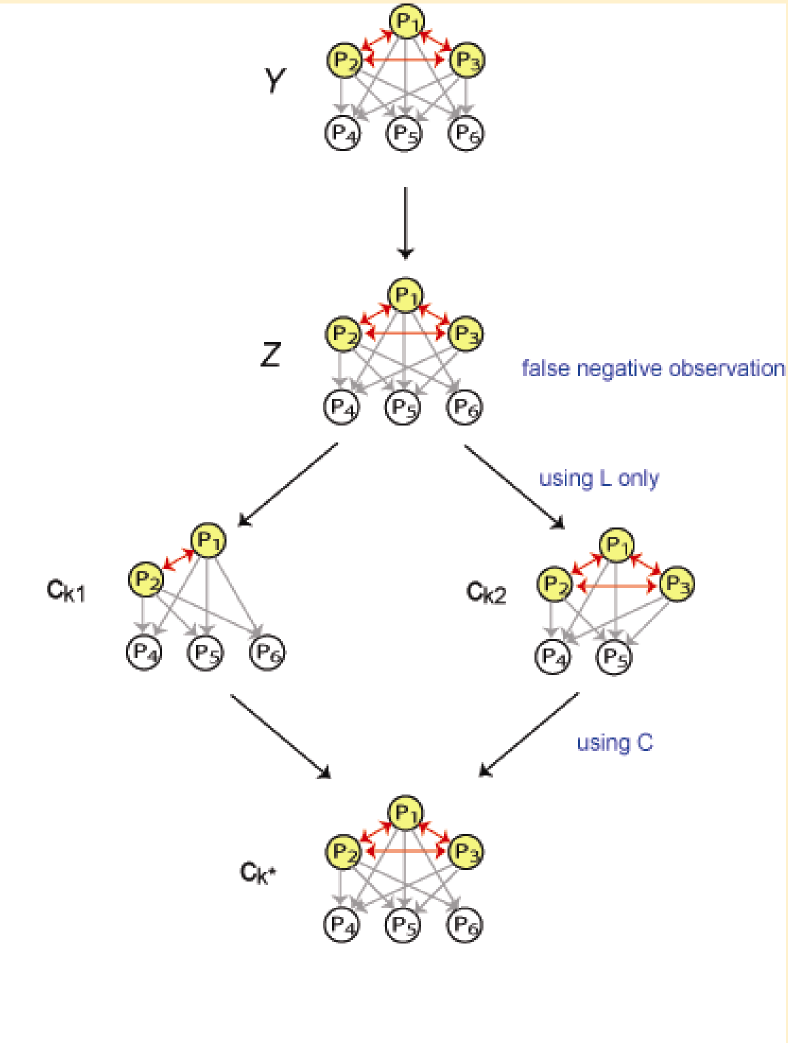
$$Y= \begin{array}{ccccccc} & P_{c1} & P_{c2} & P_{c3} & P_{c4} & P_{c5} & P_{c6} \\ P_{c1} & 1 & 1 & 1 & 1 & 1 & 1 \\ P_{c2} & 1 & 1 & 0 & 1 & 0 & 1 \\ P_{c3} & 1 & 0 & 1 & 1 & 1 & 1 \\ P_{c4} & 1 & 1 & 1 & 1 & 1 & 1 \\ P_{c5} & 1 & 0 & 1 & 1 & 1 & 1 \\ P_{c6} & 1 & 1 & 1 & 1 & 1 & 1 \end{array} \Rightarrow A= \begin{array}{ccc} & C_1 & C_2 \\ P_{c1} & 1 & 1 \\ P_{c2} & 1 & 0 \\ P_{c3} & 0 & 1 \\ P_{c4} & 1 & 1 \\ P_{c5} & 0 & 1 \\ P_{c6} & 1 & 1 \end{array}$$

$$Y= \begin{array}{ccccccc} & P_{c1} & P_{c2} & P_{c3} & P_{c4} & P_{c5} & P_{c6} \\ P_{c1} & 1 & 1 & 1 & 1 & 1 & 1 \\ P_{c2} & 1 & 1 & 0 & 1 & 0 & 1 \\ P_{c3} & 1 & 0 & 1 & 0 & 1 & 0 \\ P_{c4} & 1 & 1 & 0 & 1 & 1 & 1 \\ P_{c5} & 1 & 0 & 1 & 1 & 1 & 1 \\ P_{c6} & 1 & 1 & 0 & 1 & 1 & 1 \end{array} \Rightarrow A= \begin{array}{cccc} & C_1 & C_2 & C_3 \\ P_{c1} & 1 & 1 & 1 \\ P_{c2} & 1 & 0 & 0 \\ P_{c3} & 0 & 1 & 0 \\ P_{c4} & 1 & 0 & 1 \\ P_{c5} & 0 & 1 & 1 \\ P_{c6} & 1 & 0 & 1 \end{array}$$

Since we only use a subset of the proteins as baits, we cannot identify maximal complete subgraphs in Y. Instead, the initial estimate of A based on Y consists of the maximal BH-complete subgraphs in Y.

Why *C*?
Why isn't *L* enough?

- At most, each edge is tested twice, and independent errors are made in the observation of all edges.
- A false negative observation from a bait to a hit would break one complex into two estimated complexes.
- Effectively, *C* relaxes the maximal BH-complete subgraph requirement for the initial complex estimates to accommodate a proportion of false negative observations in accordance with the sensitivity of the AP-MS technology.



$P(Z|A,\mu,\alpha)=L(Z|Y=A\otimes A',\mu,\alpha)C\ (Z|A,\mu,\alpha)$

C is designed to allow combinations of the complexes in the estimated *A* that increase *C* in favor of small decreases in *L*.

$$C(Z\mid A,\mu,\alpha)=\prod_{k=1}^{K}\Phi(c_k)\Gamma(c_k) \qquad (K=\text{total \# of complexes})$$

c_k is a complex estimate with n_k bait proteins and m_k hit - only proteins

$\Phi(c_k)=$ cumulative probability of observing a particular missing edge pattern or something more extreme for the edges in complex c_k, i.e. two - sided p - value from Fisher's exact test on node indegree

$$\Gamma(c_k)=\begin{pmatrix}t_k\\x_k\end{pmatrix}\frac{e^{x_k(\mu+\alpha)}}{\left(1+e^{(\mu+\alpha)}\right)^{t_k}}, \qquad \left(\frac{e^{(\mu+\alpha)}}{1+e^{(\mu+\alpha)}}=\text{sensitivity}\right)$$

$t_k=n_k(n_k+m_k-1)=$ number of tested edges in BH - complete subgraph for c_k

$x_k=$ number of observed edges in BH - complete subgraph for c_k

$P(Z|A,\mu,\alpha)=L(Z|Y=A\otimes A',\mu,\alpha)C\ (Z|A,\mu,\alpha)$

C is designed to allow combinations of the complexes in the estimated *A* that increase *C* in favor of small decreases in *L*.

$$C(Z\,|\,A,\mu,\alpha)=\prod_{k=1}^{K}\Phi(c_k)\Gamma(c_k)\qquad(K=\text{total \# of complexes})$$

c_k is a complex estimate with n_k bait proteins and m_k hit-only proteins

$\Phi(c_k)=$ cumulative probability of observing a particular missing edge pattern or something more extreme for the edges in complex c_k,

i.e. two-sided p-value from Fisher's exact test on node indegree

$$\Gamma(c_k)=\binom{t_k}{x_k}\frac{e^{x_k(\mu+\alpha)}}{\left(1+e^{(\mu+\alpha)}\right)^{t_k}},\qquad\left(\frac{e^{(\mu+\alpha)}}{1+e^{(\mu+\alpha)}}=\text{sensitivity}\right)$$

$t_k=n_k(n_k+m_k-1)=$ number of tested edges in BH-complete subgraph for c_k

$x_k=$ number of observed edges in BH-complete subgraph for c_k

Since the thousands of individual edges in Y are tested at most twice, an estimate of A based solely on L may not be accurate. C offers a second criteria to further refine A.

Combining Complex Estimates

For two complex estimates, c_{k1} and c_{k2}, we check to see if they increase P when treated as one complex c_{k^*}.

Specifically, if $\log P_{k^*} \text{-} \log P_{k1,k2} > 0$, we combine c_{k1} and c_{k2} a new c_{k^*}.

$$\begin{aligned}\log P_{k^*} \text{-} \log P_{k1,k2} &= \log \Phi(c_{k^*}) - \log \Phi(c_{k1}) - \log \Phi(c_{k2}) \\ &+ \log \Gamma(c_{k^*}) - \log \Gamma(c_{k1}) - \log \Gamma(c_{k2}) \\ &+ \sum_{S_{new}} \left[\alpha z_{gh} - \log(1+e^{\mu+\alpha}) + \log(1+e^{\mu}) \right]\end{aligned}$$

where S_{new} = set of all edges between proteins g and h that are being changed from "absent" to "present"

Combining Complex Estimates

For two complex estimates, c_{k1} and c_{k2}, we check to see if they increase P when treated as one complex c_{k*}.

Specifically, if $\log P_{k*} - \log P_{k1,k2} > 0$, we combine c_{k1} and c_{k2} a new c_{k*}.

$$\begin{aligned}\log P_{k*} - \log P_{k1,k2} &= \log \Phi(c_{k*}) - \log \Phi(c_{k1}) - \log \Phi(c_{k2}) \\ &+ \log \Gamma(c_{k*}) - \log \Gamma(c_{k1}) - \log \Gamma(c_{k2}) \\ &+ \sum_{S_{new}} \left[\alpha z_{gh} - \log(1+e^{\mu+\alpha}) + \log(1+e^{\mu}) \right]\end{aligned}$$

where S_{new} = set of all edges between proteins g and h that are being changed from "absent" to "present"

In general, P increases for a smaller number of complexes that are both reflective of approximate maximal BH-complete subgraph structure and consistent with the observed data.

Complex Estimation Algorithm

Complex Estimation Algorithm

1. Find the MLE for *Y* using *Z.*

Complex Estimation Algorithm

1. Find the MLE for *Y* using *Z.*
2. Find the initial estimate for *A* by constructing maximal BH-complete subgraphs in *Y*.

Complex Estimation Algorithm

1. Find the MLE for *Y* using *Z.*
2. Find the initial estimate for *A* by constructing maximal BH-complete subgraphs in *Y*.
3. Order the columns of *A* according to the number of baits.

Complex Estimation Algorithm

1. Find the MLE for *Y* using *Z.*
2. Find the initial estimate for *A* by constructing maximal BH-complete subgraphs in *Y*.
3. Order the columns of *A* according to the number of baits.
4. Set *k*=1 and *K*=number of columns of *A*.

Complex Estimation Algorithm

1. Find the MLE for *Y* using *Z.*
2. Find the initial estimate for *A* by constructing maximal BH-complete subgraphs in *Y*.
3. Order the columns of *A* according to the number of baits.
4. Set *k*=1 and *K*=number of columns of *A*.
5. For c_k, find the set A_k of columns of *A*, excluding c_k, that share at least one common entry of “1”. Calculate log P_{k*} -log $\mathrm{P}_{\mathrm{k1,k2}}$ for c_k paired with all elements in A_k.

Complex Estimation Algorithm

1. Find the MLE for *Y* using *Z.*
2. Find the initial estimate for *A* by constructing maximal BH-complete subgraphs in *Y*.
3. Order the columns of *A* according to the number of baits.
4. Set *k*=1 and *K*=number of columns of *A*.
5. For c_k, find the set A_k of columns of *A*, excluding c_k, that share at least one common entry of "1". Calculate log P_{k*} -log $\mathrm{P}_{\mathrm{k1,k2}}$ for c_k paired with all elements in A_k.
6. If at least one value of log P_{k*} -log $\mathrm{P}_{\mathrm{k1,k2}}$ is greater than 0, replace c_k with the union of c_k and c_{Akmax},the element of A_k giving the largest value of log P_{k*} -log $\mathrm{P}_{\mathrm{k1,k2}}$. Remove c_{Akmax} and any columns that are strictly less than c_k Uc_{Akmax}. Set *K*=number of columns of *A*.

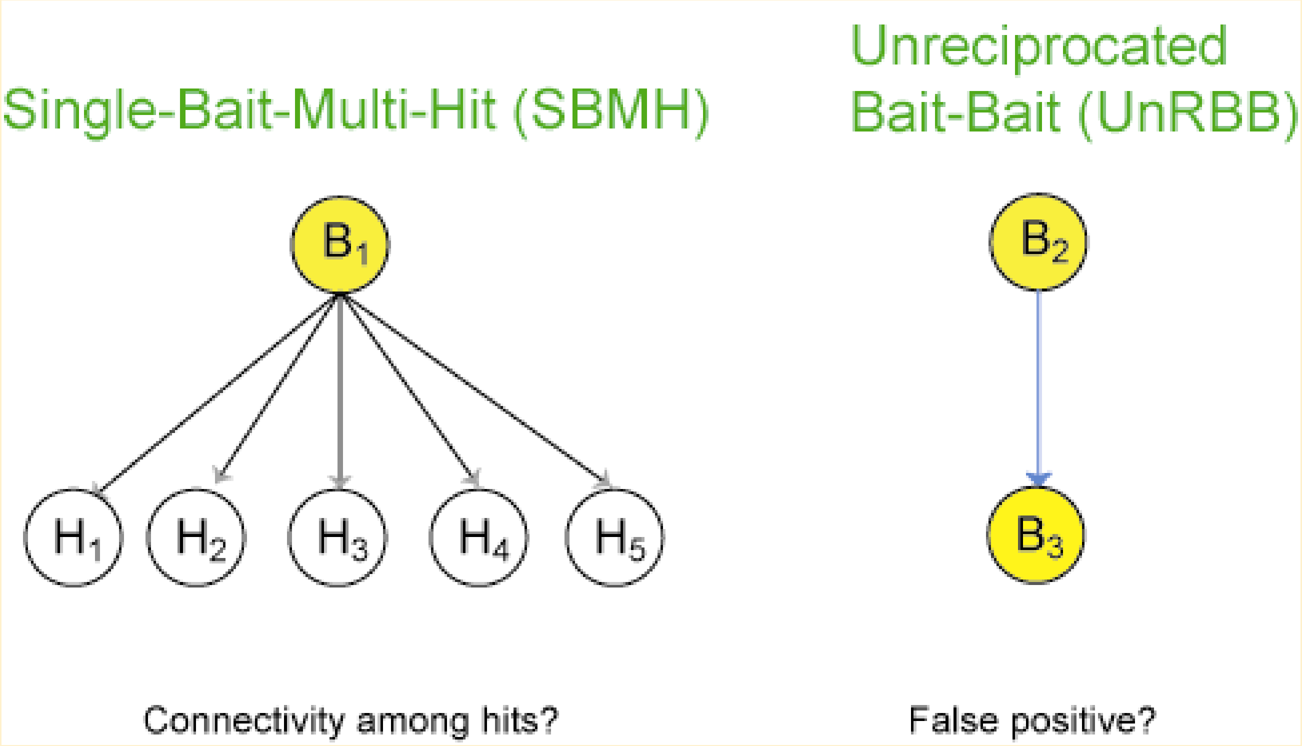
Complex Estimation Algorithm

1. Find the MLE for Y using Z.
2. Find the initial estimate for A by constructing maximal BH-complete subgraphs in Y.
3. Order the columns of A according to the number of baits.
4. Set k=1 and K=number of columns of A.
5. For c_k, find the set A_k of columns of A, excluding c_k, that share at least one common entry of “1”. Calculate log P_{k*} -log $\mathrm{P}_{k1,k2}$ for c_k paired with all elements in A_k.
6. If at least one value of log P_{k*} -log $\mathrm{P}_{k1,k2}$ is greater than 0, replace c_k with the union of c_k and c_{Akmax},the element of A_k giving the largest value of log P_{k*} -log $\mathrm{P}_{k1,k2}$. Remove c_{Akmax} and any columns that are strictly less than c_k Uc_{Akmax}. Set K=number of columns of A.
7. If none of the values of log P_{k*} -log $\mathrm{P}_{k1,k2}$ are greater than 0, set k=k+1, and return to step 5.

Complex Estimation Algorithm

1. Find the MLE for *Y* using *Z.*
2. Find the initial estimate for *A* by constructing maximal BH-complete subgraphs in *Y*.
3. Order the columns of *A* according to the number of baits.
4. Set k=1 and K=number of columns of *A*.
5. For c_k, find the set A_k of columns of *A*, excluding c_k, that share at least one common entry of “1”. Calculate log P_{k*} -log $\mathrm{P}_{\mathrm{k1,k2}}$ for c_k paired with all elements in A_k.
6. If at least one value of log P_{k*} -log $\mathrm{P}_{\mathrm{k1,k2}}$ is greater than 0, replace c_k with the union of c_k and c_{Akmax},the element of A_k giving the largest value of log P_{k*} -log $\mathrm{P}_{\mathrm{k1,k2}}$. Remove c_{Akmax} and any columns that are strictly less than c_k Uc_{Akmax}. Set K=number of columns of *A*.
7. If none of the values of log P_{k*} -log $\mathrm{P}_{\mathrm{k1,k2}}$ are greater than 0, set k=k+1, and return to step 5.
8. Repeat until k=K.

Two types of complex estimates to interpret with care



Single-Bait-Multi-Hit (SBMH)

Unreciprocated Bait-Bait (UnRBB)

Connectivity among hits?

False positive?

TAP data analysis

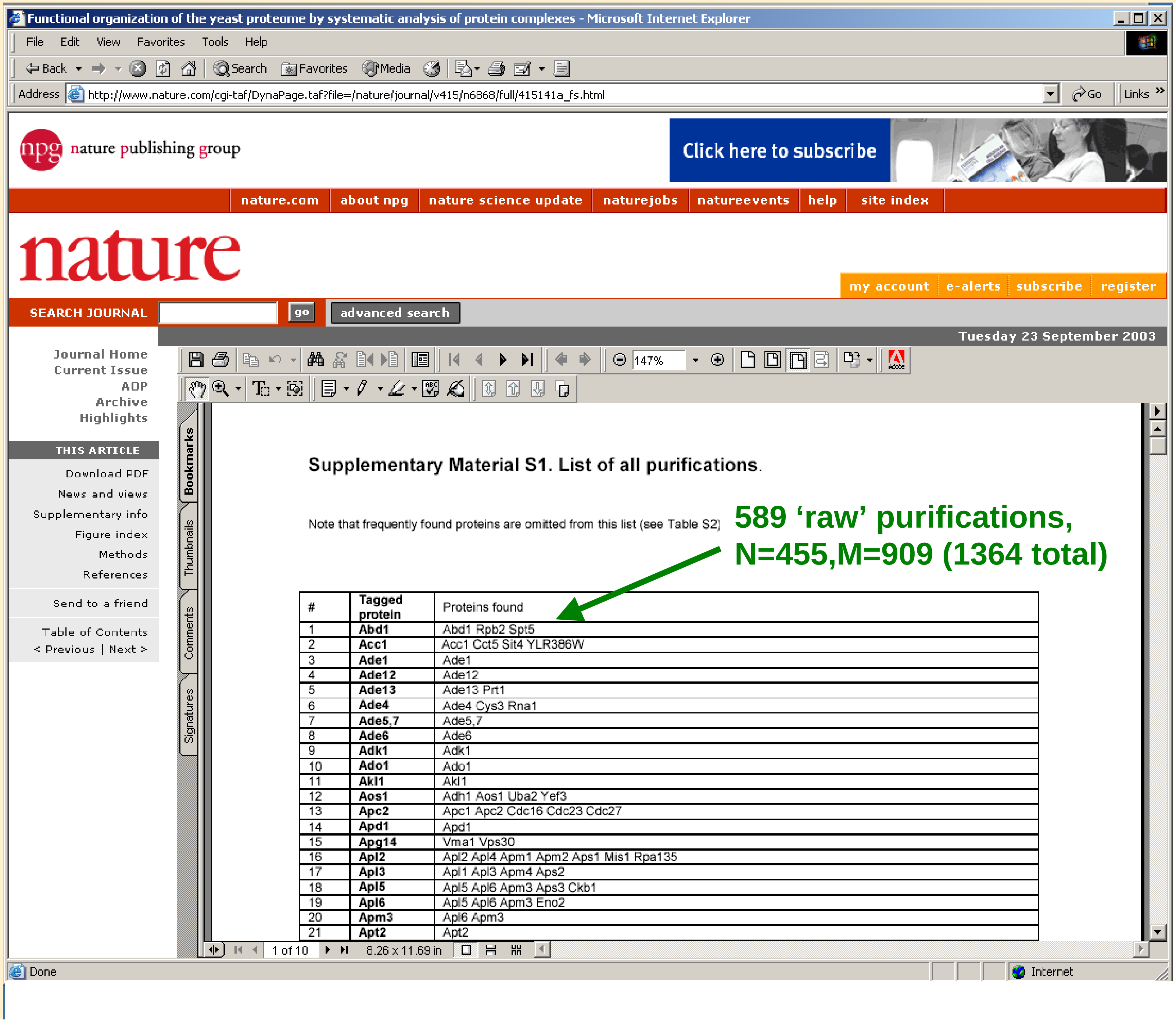
- Sensitivity=.75, Specificity=.001
- Gene Ontology (GO) cellular component-based similarity measure in an extended logistic regression model
  - Purpose is to increase the probability that two proximally located proteins are complex comembers even if there is not an edge between them
- 720 complexes total
  - 123 UnRBB
  - 331 SBMH
  - ***266* multi-bait complexes with at least 2 proteins and at least 2 edges**
- Compared these ***266*** complexes to the ***232 yTAP*** complexes (Gavin et al. 2002) through both a large scale comparison, and complex-by-complex for several complexes.

Large Scale Comparison to Known Complexes

- Similarity measure: $\omega=\min(i/a,i/b)$
  - a = # proteins in complex A, b = # proteins in complex B
  - i = # proteins in both A and B

- Munich Information Center for Protein Sequences (**MIPS**) reports a list of 267 curated protein complexes , **129** of which involved 595 proteins contained in the TAP data.

- Using $\omega>.70$ as a mapping criteria and the common subset of 595 proteins, we mapped **85 of our complexes to 65 MIPS complexes** and **40 yTAP complexes to 32 MIPS complexes**.


Functional organization of the yeast proteome by systematic analysis of protein complexes - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.nature.com/cgi-taf/DynaPage.taf?file=/nature/journal/v415/n6868/full/415141a_fs.html
Go Links
npg nature publishing group
Click here to subscribe
nature.com about npg nature science update naturejobs natureevents help site index
nature
my account e-alerts subscribe register
SEARCH JOURNAL go advanced search
Tuesday 23 September 2003
Journal Home
Current Issue
AOP
Archive
Highlights
THIS ARTICLE
Download PDF
News and views
Supplementary info
Figure index
Methods
References
Send to a friend
Table of Contents
< Previous | Next >
147%
Bookmarks
Thumbnails
Comments
Signatures
1 of 10
8.26 x 11.69 in
Done
Internet

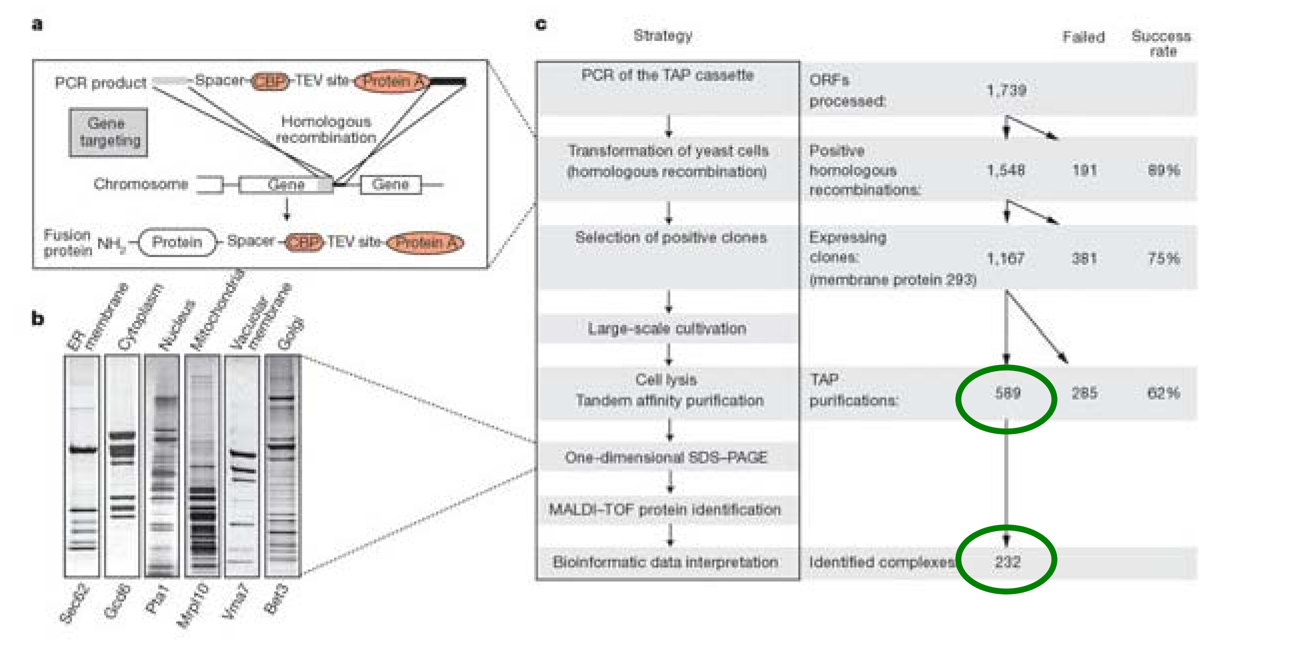
Supplementary Material S1. List of all purifications.

Note that frequently found proteins are omitted from this list (see Table S2)

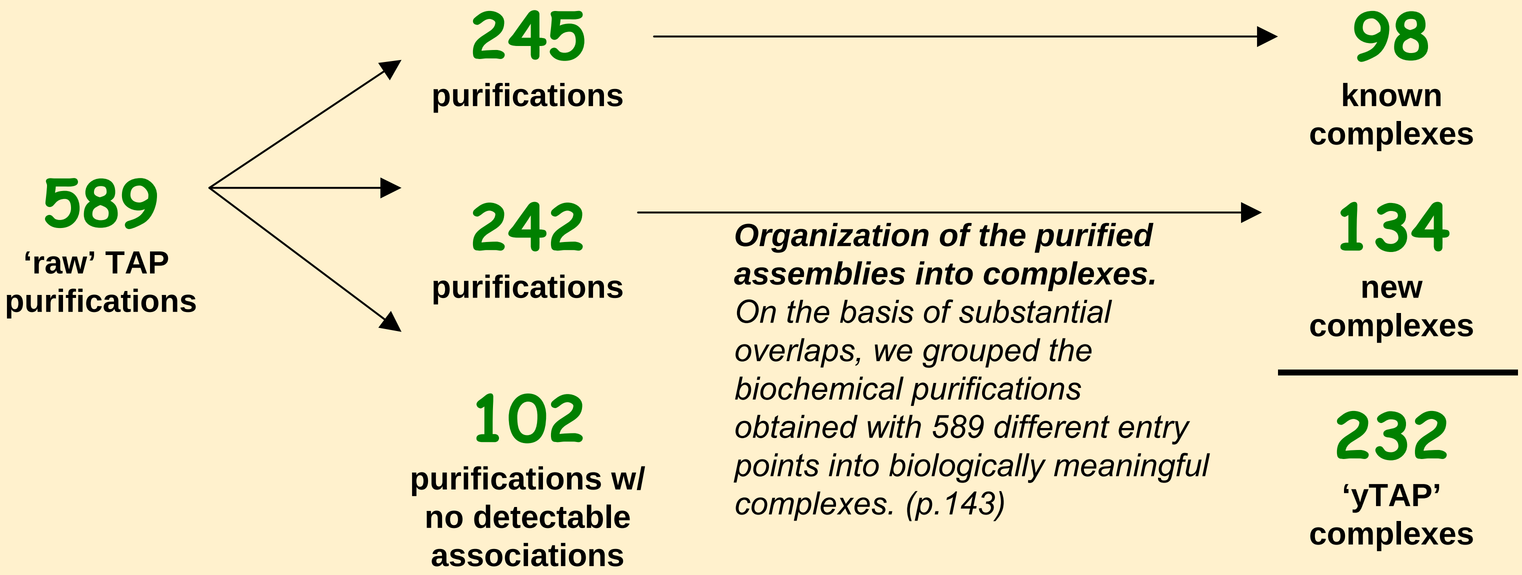
589 ‘raw’ purifications, N=455,M=909 (1364 total)

#	Tagged protein	Proteins found
1	**Abd1**	Abd1 Rpb2 Spt5
2	**Acc1**	Acc1 Cct5 Sit4 YLR386W
3	**Ade1**	Ade1
4	**Ade12**	Ade12
5	**Ade13**	Ade13 Prt1
6	**Ade4**	Ade4 Cys3 Rna1
7	**Ade5,7**	Ade5,7
8	**Ade6**	Ade6
9	**Adk1**	Adk1
10	**Ado1**	Ado1
11	**Akl1**	Akl1
12	**Aos1**	Adh1 Aos1 Uba2 Yef3
13	**Apc2**	Apc1 Apc2 Cdc16 Cdc23 Cdc27
14	**Apd1**	Apd1
15	**Apg14**	Vma1 Vps30
16	**Apl2**	Apl2 Apl4 Apm1 Apm2 Aps1 Mis1 Rpa135
17	**Apl3**	Apl1 Apl3 Apm4 Aps2
18	**Apl5**	Apl5 Apl6 Apm3 Aps3 Ckb1
19	**Apl6**	Apl5 Apl6 Apm3 Eno2
20	**Apm3**	Apl6 Apm3
21	**Apt2**	Apt2

Figure 1 from Gavin, et al.



Strategy			Failed	Success rate
PCR of the TAP cassette	ORFs processed:	1,739		
Transformation of yeast cells (homologous recombination)	Positive homologous recombinations:	1,548	191	89%
Selection of positive clones	Expressing clones: (membrane protein 293)	1,167	381	75%
Large-scale cultivation				
Cell lysis Tandem affinity purification	TAP purifications:	589	285	62%
One-dimensional SDS–PAGE				
MALDI–TOF protein identification				
Bioinformatic data interpretation	Identified complexes	232		



589
'raw' TAP purifications

245
purifications

242
purifications

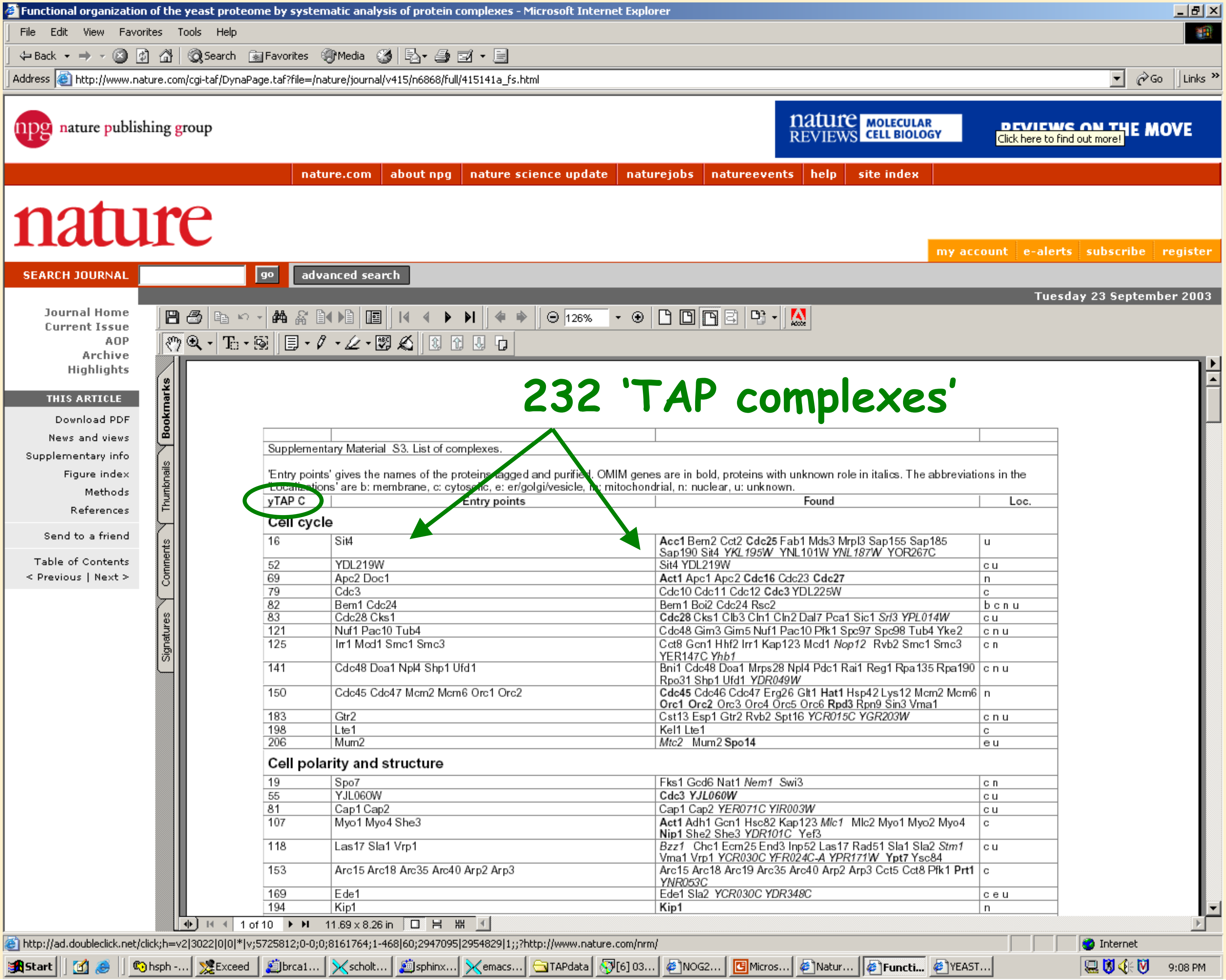
102
purifications w/ no detectable associations

98
known complexes

134
new complexes

232
'yTAP' complexes

Organization of the purified assemblies into complexes. *On the basis of substantial overlaps, we grouped the biochemical purifications obtained with 589 different entry points into biologically meaningful complexes. (p.143)*

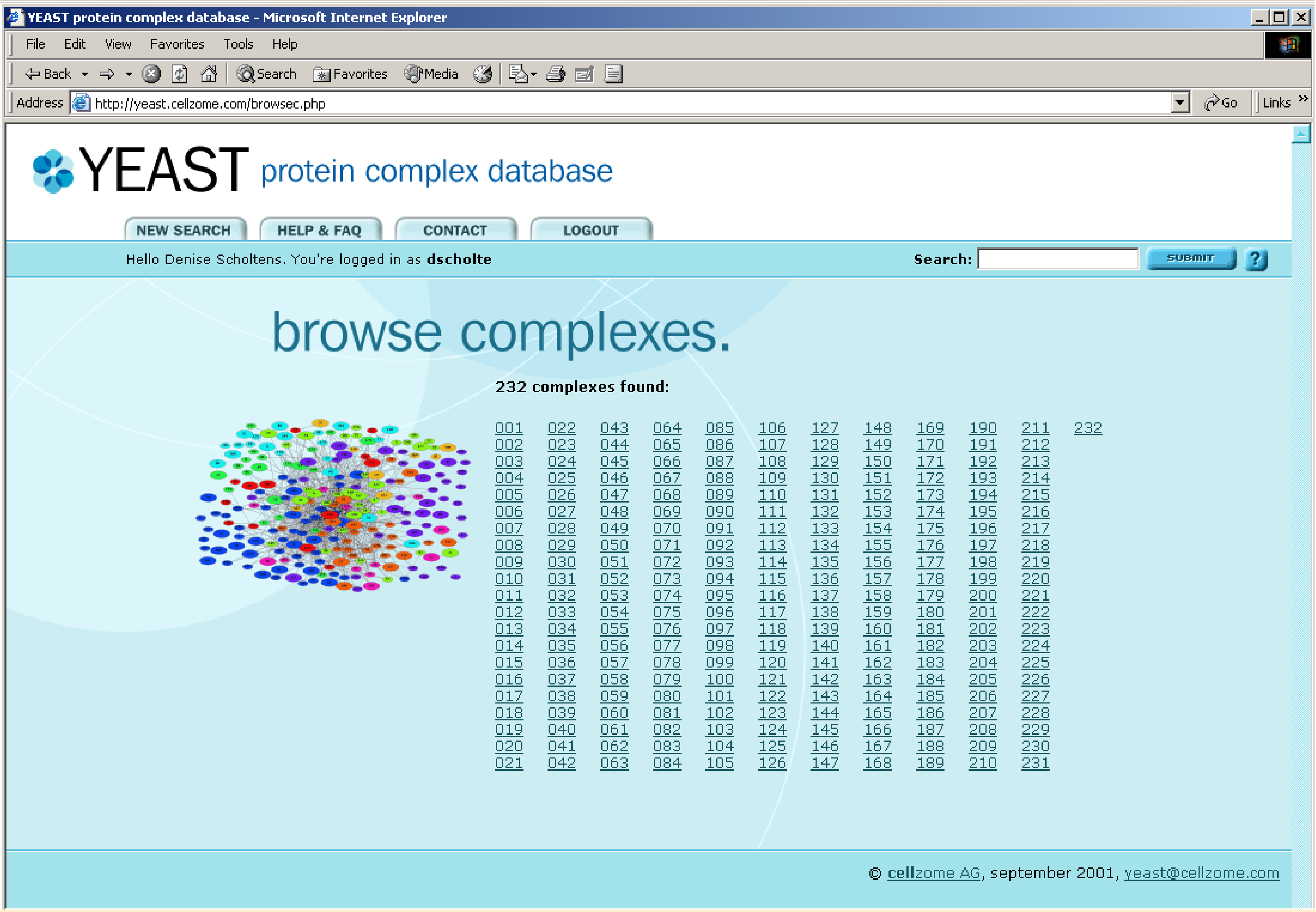

Functional organization of the yeast proteome by systematic analysis of protein complexes - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.nature.com/cgi-taf/DynaPage.taf?file=/nature/journal/v415/n6868/full/415141a_fs.html
Go Links
npg nature publishing group
nature REVIEWS MOLECULAR CELL BIOLOGY
REVIEWS ON THE MOVE
Click here to find out more!
nature.com about npg nature science update naturejobs natureevents help site index
nature
my account e-alerts subscribe register
SEARCH JOURNAL go advanced search
Tuesday 23 September 2003
Journal Home
Current Issue
AOP
Archive
Highlights
THIS ARTICLE
Download PDF
News and views
Supplementary info
Figure index
Methods
References
Send to a friend
Table of Contents
< Previous | Next >
126%
Bookmarks
Thumbnails
Comments
Signatures
1 of 10
11.69 x 8.26 in
http://ad.doubleclick.net/click;h=v2|3022|0|0|*|v;5725812;0-0;0;8161764;1-468|60;2947095|2954829|1;;?http://www.nature.com/nrm/
Internet
Start hsph -... Exceed brca1... scholt... sphinx... emacs... TAPdata [6] 03... NOG2... Micros... Natur... Functi... YEAST...
9:08 PM

232 'TAP complexes'

Supplementary Material S3. List of complexes.

'Entry points' gives the names of the proteins tagged and purified. OMIM genes are in bold, proteins with unknown role in italics. The abbreviations in the 'Localizations' are b: membrane, c: cytosolic, e: er/golgi/vesicle, m: mitochondrial, n: nuclear, u: unknown.

yTAP C	Entry points	Found	Loc.
Cell cycle			
16	Sit4	**Acc1** Bem2 Cct2 **Cdc25** Fab1 Mds3 Mrpl3 Sap155 Sap185 Sap190 Sit4 *YKL195W* YNL101W *YNL187W* YOR267C	u
52	YDL219W	Sit4 YDL219W	c u
69	Apc2 Doc1	**Act1** Apc1 Apc2 **Cdc16** Cdc23 **Cdc27**	n
79	Cdc3	Cdc10 Cdc11 Cdc12 **Cdc3** YDL225W	c
82	Bem1 Cdc24	Bem1 Boi2 Cdc24 Rsc2	b c n u
83	Cdc28 Cks1	**Cdc28** Cks1 Clb3 Cln1 Cln2 Dal7 Pca1 Sic1 *Srl3 YPL014W*	c u
121	Nuf1 Pac10 Tub4	Cdc48 Gim3 Gim5 Nuf1 Pac10 Pfk1 Spc97 Spc98 Tub4 Yke2	c n u
125	Irr1 Mcd1 Smc1 Smc3	Cct8 Gcn1 Hhf2 Irr1 Kap123 Mcd1 *Nop12* Rvb2 Smc1 Smc3 YER147C *Yhb1*	c n
141	Cdc48 Doa1 Npl4 Shp1 Ufd1	Bni1 Cdc48 Doa1 Mrps28 Npl4 Pdc1 Rai1 Reg1 Rpa135 Rpa190 Rpo31 Shp1 Ufd1 *YDR049W*	c n u
150	Cdc45 Cdc47 Mcm2 Mcm6 Orc1 Orc2	**Cdc45** Cdc46 Cdc47 Erg26 Glt1 **Hat1** Hsp42 Lys12 Mcm2 Mcm6 **Orc1 Orc2** Orc3 Orc4 Orc5 Orc6 **Rpd3** Rpn9 Sin3 Vma1	n
183	Gtr2	Cst13 Esp1 Gtr2 Rvb2 Spt16 *YCR015C YGR203W*	c n u
198	Lte1	Kel1 Lte1	c
206	Mum2	*Mtc2* Mum2 **Spo14**	e u
Cell polarity and structure			
19	Spo7	Fks1 Gcd6 Nat1 *Nem1* Swi3	c n
55	YJL060W	**Cdc3** ***YJL060W***	c u
81	Cap1 Cap2	Cap1 Cap2 *YER071C YIR003W*	c u
107	Myo1 Myo4 She3	**Act1** Adh1 Gcn1 Hsc82 Kap123 *Mlc1* Mlc2 Myo1 Myo2 Myo4 **Nip1** She2 She3 *YDR101C* Yef3	c
118	Las17 Sla1 Vrp1	*Bzz1* Chc1 Ecm25 End3 Inp52 Las17 Rad51 Sla1 Sla2 *Stm1* Vma1 Vrp1 *YCR030C YFR024C-A YPR171W* **Ypt7** Ysc84	c u
153	Arc15 Arc18 Arc35 Arc40 Arp2 Arp3	Arc15 Arc18 Arc19 Arc35 Arc40 Arp2 Arp3 Cct5 Cct8 Pfk1 **Prt1** *YNR053C*	c
169	Ede1	Ede1 Sla2 *YCR030C YDR348C*	c e u
194	Kip1	**Kip1**	n



YEAST protein complex database - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://yeast.cellzome.com/browsec.php
Go Links
YEAST protein complex database
NEW SEARCH
HELP & FAQ
CONTACT
LOGOUT
Search:
SUBMIT
?

Hello Denise Scholtens. You're logged in as **dscholte**

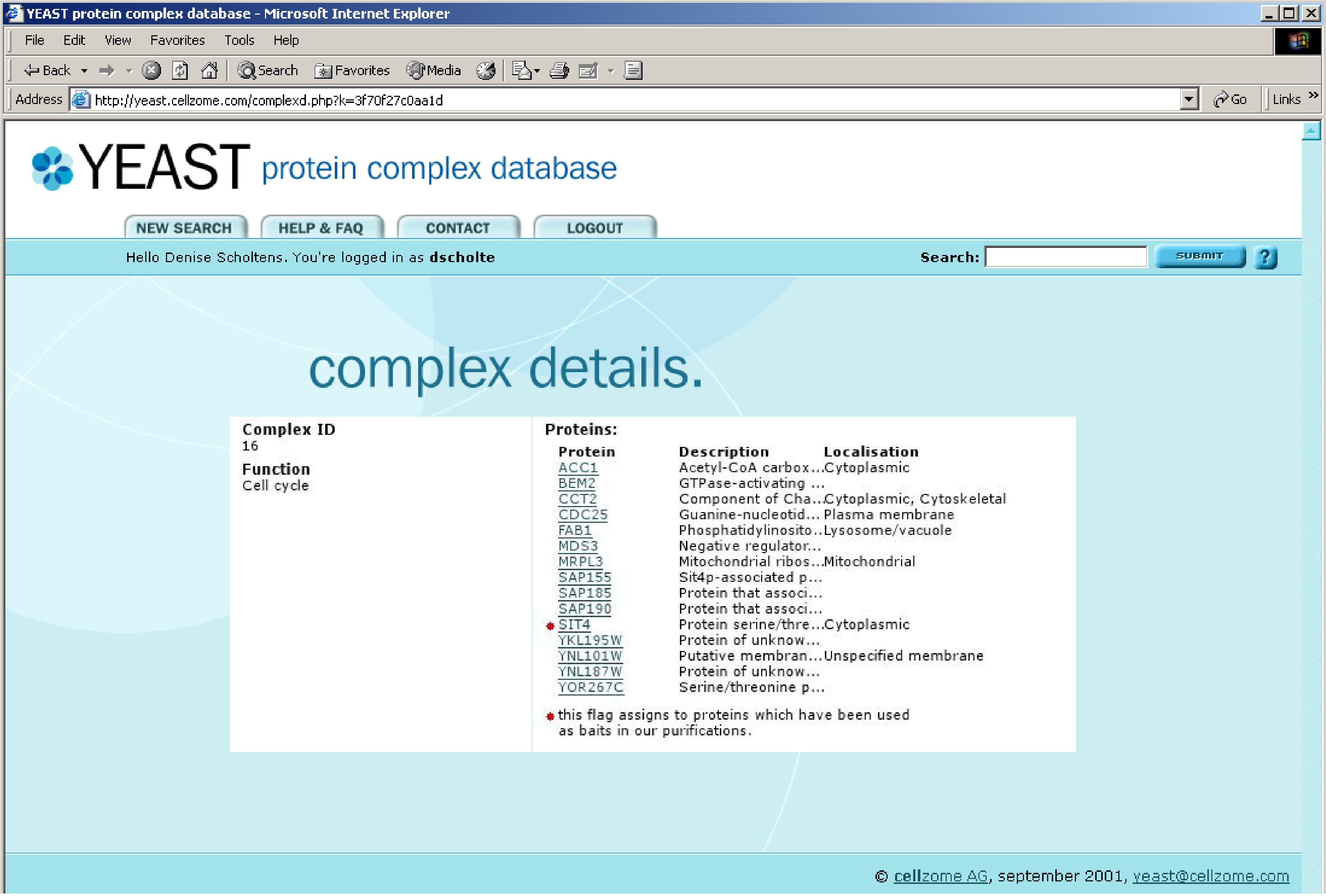
browse complexes.

232 complexes found:

001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021
022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042
043 044 045 046 047 048 049 050 051 052 053 054 055 056 057 058 059 060 061 062 063
064 065 066 067 068 069 070 071 072 073 074 075 076 077 078 079 080 081 082 083 084
085 086 087 088 089 090 091 092 093 094 095 096 097 098 099 100 101 102 103 104 105
106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126
127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147
148 149 150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168
169 170 171 172 173 174 175 176 177 178 179 180 181 182 183 184 185 186 187 188 189
190 191 192 193 194 195 196 197 198 199 200 201 202 203 204 205 206 207 208 209 210
211 212 213 214 215 216 217 218 219 220 221 222 223 224 225 226 227 228 229 230 231
232

© cellzome AG, september 2001, yeast@cellzome.com

http://yeast.cellzome.com


YEAST protein complex database - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://yeast.cellzome.com/complexd.php?k=3f70f27c0aa1d
Go Links
YEAST protein complex database
NEW SEARCH
HELP & FAQ
CONTACT
LOGOUT
Search:
SUBMIT
?

Hello Denise Scholtens. You're logged in as **dscholte**

complex details.

Complex ID
16

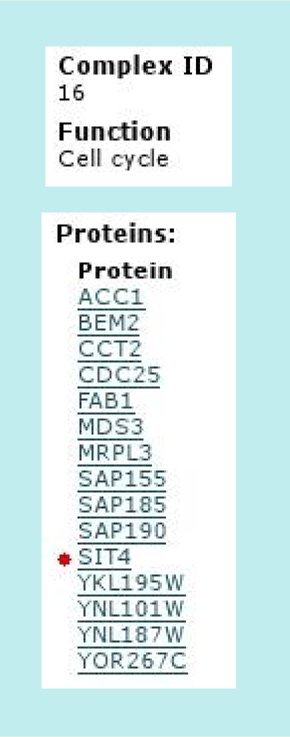
Function
Cell cycle

Proteins:

Protein	Description	Localisation
ACC1	Acetyl-CoA carbox...	Cytoplasmic
BEM2	GTPase-activating ...	
CCT2	Component of Cha...	Cytoplasmic, Cytoskeletal
CDC25	Guanine-nucleotid...	Plasma membrane
FAB1	Phosphatidylinosito..	Lysosome/vacuole
MDS3	Negative regulator...	
MRPL3	Mitochondrial ribos...	Mitochondrial
SAP155	Sit4p-associated p...	
SAP185	Protein that associ...	
SAP190	Protein that associ...	
● SIT4	Protein serine/thre...	Cytoplasmic
YKL195W	Protein of unknow...	
YNL101W	Putative membran...	Unspecified membrane
YNL187W	Protein of unknow...	
YOR267C	Serine/threonine p...	

● this flag assigns to proteins which have been used as baits in our purifications.

© cellzome AG, september 2001, yeast@cellzome.com

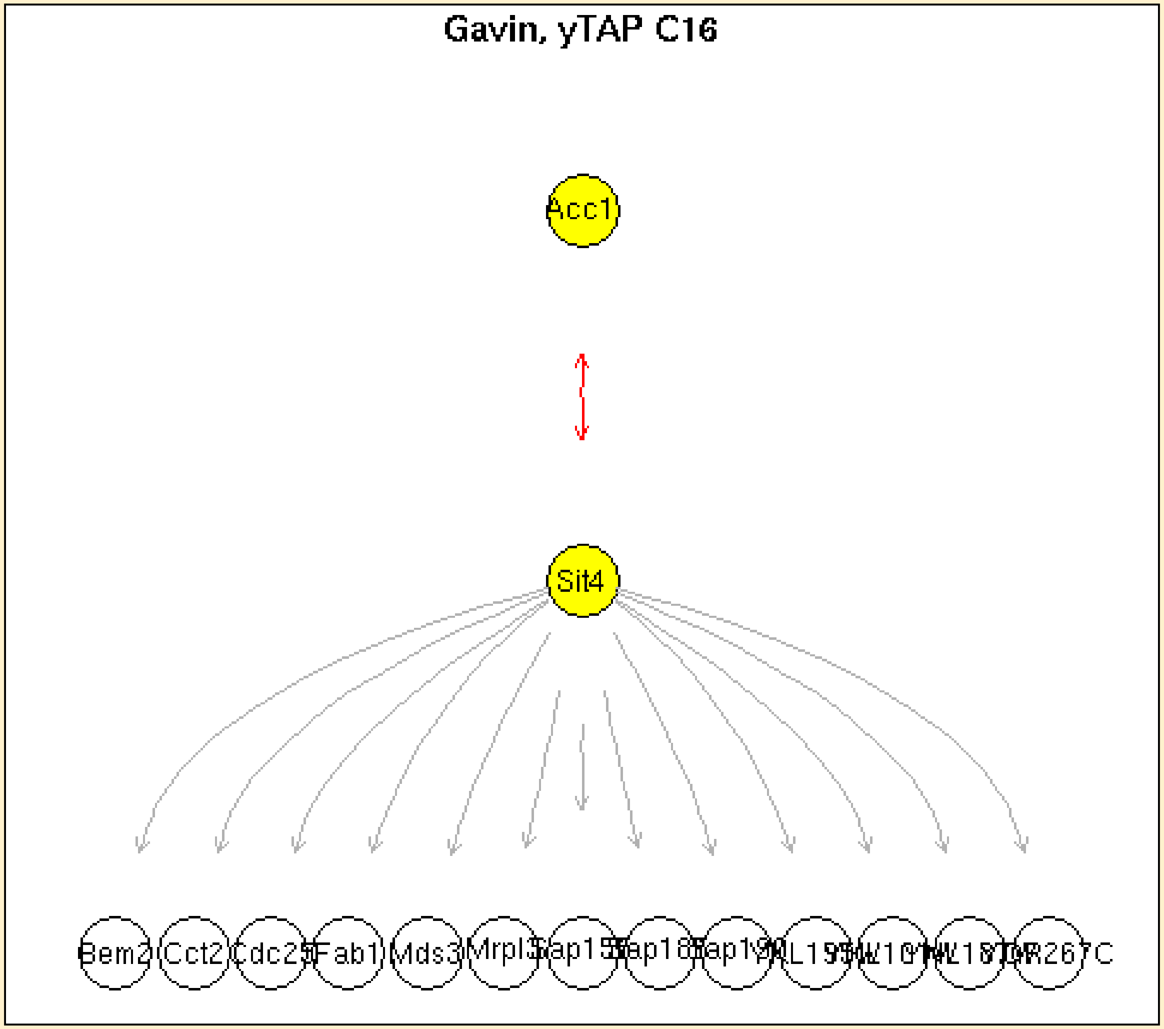


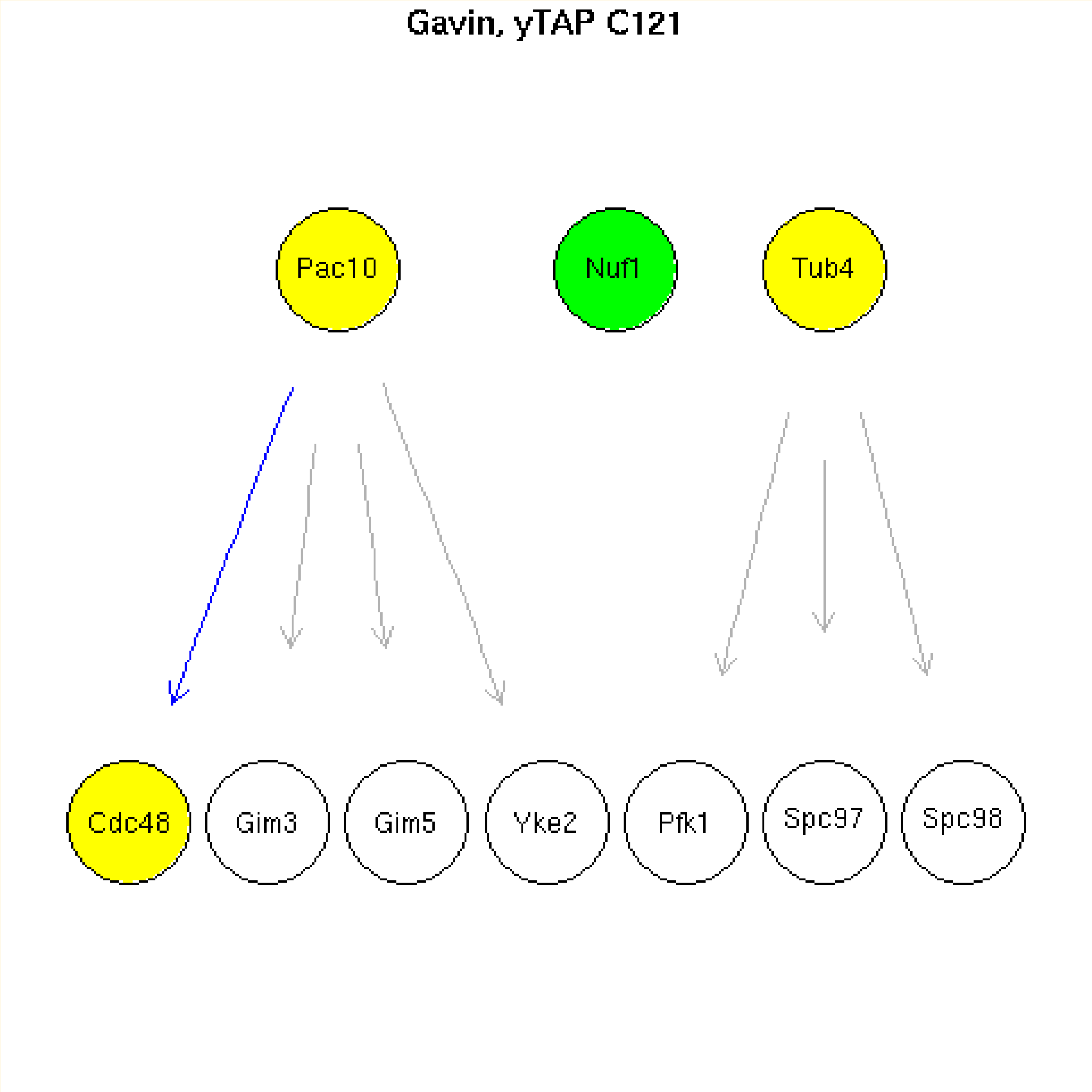
Complex ID
16

Function
Cell cycle

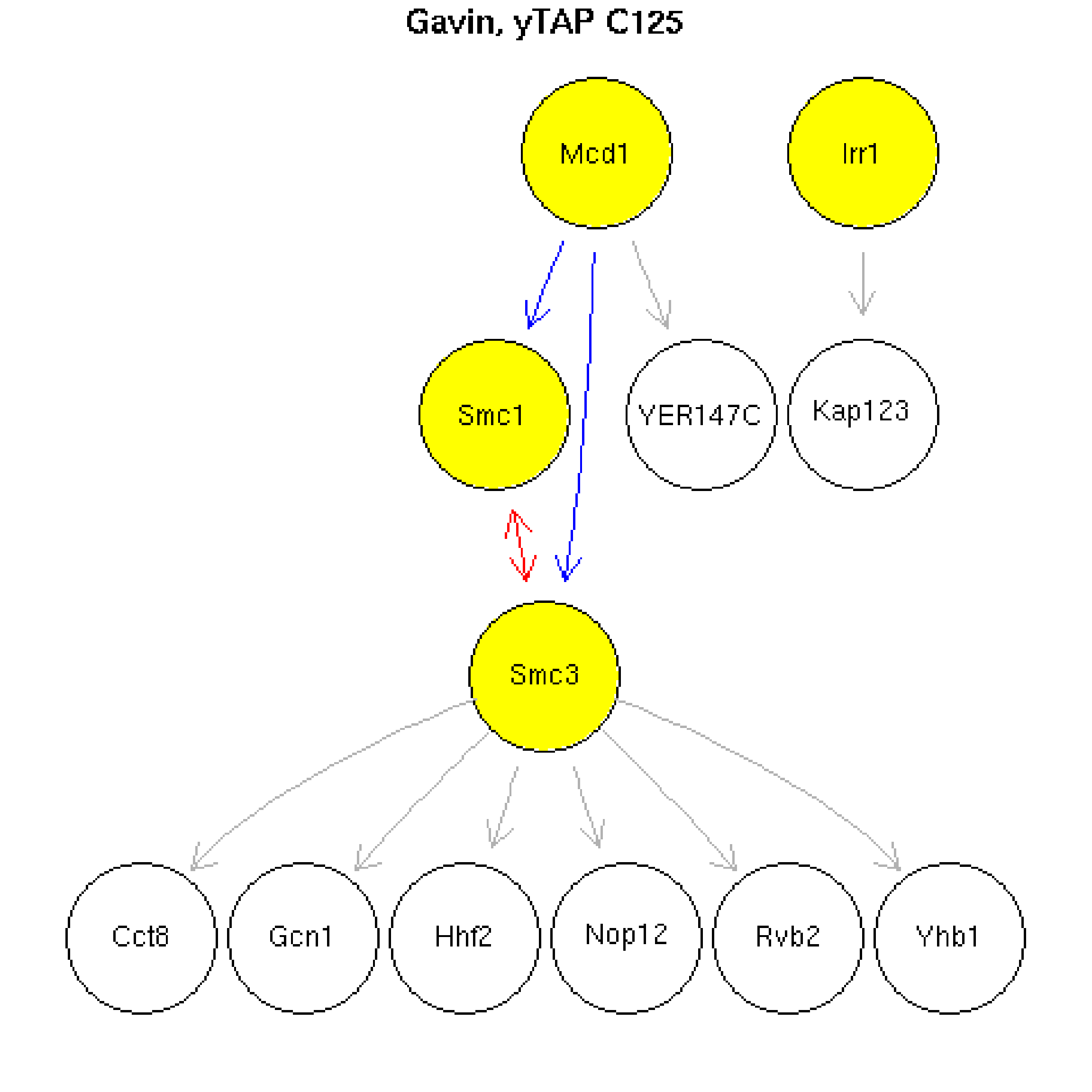
Proteins:

Protein
ACC1
BEM2
CCT2
CDC25
FAB1
MDS3
MRPL3
SAP155
SAP185
SAP190
● SIT4
YKL195W
YNL101W
YNL187W
YOR267C





Example of unconnected complex, yTAP C121



Example of unconnected complex, yTAP C125

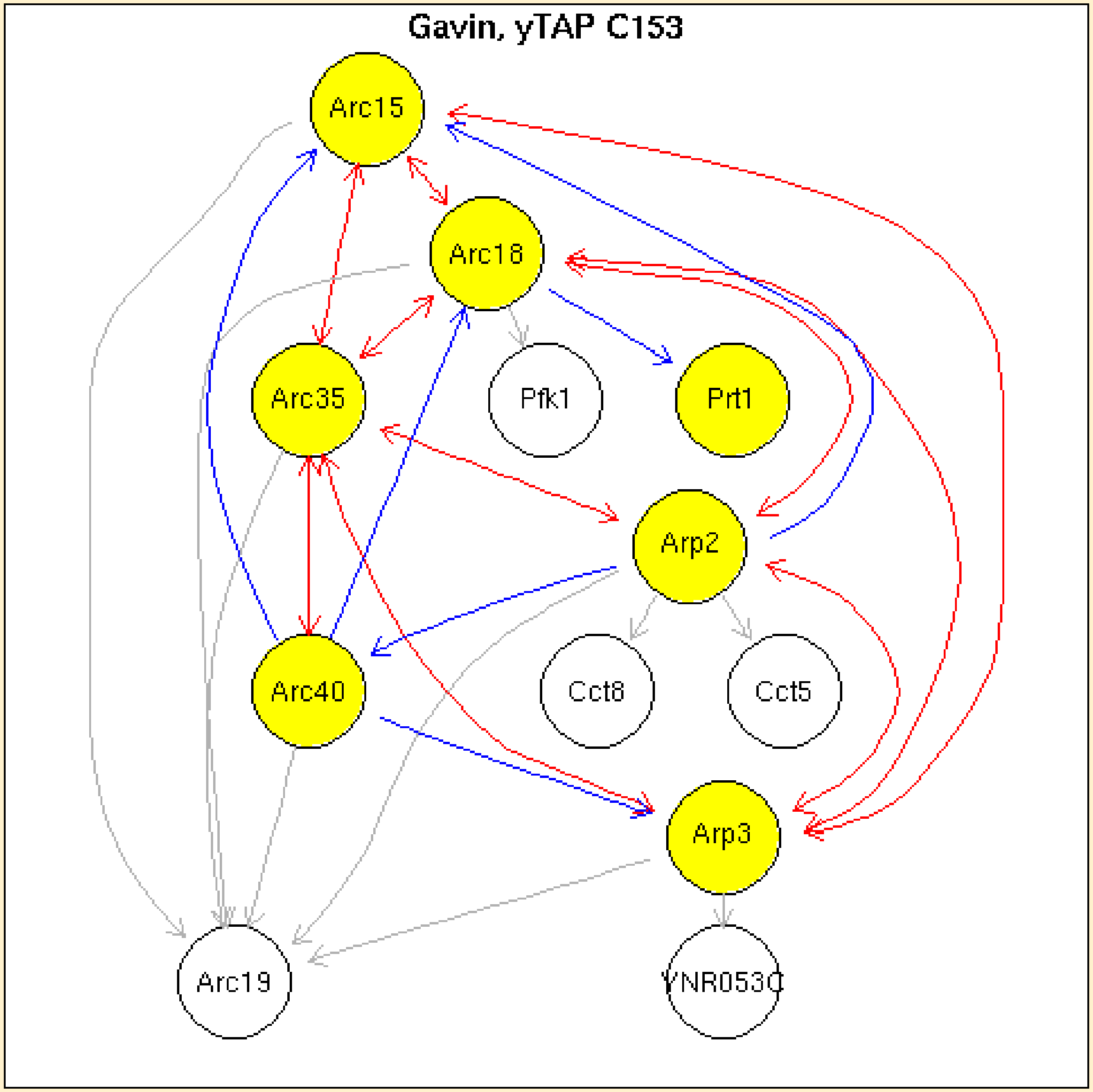
Arp2/3

Arp2/3 complex:

- Arp2
- Arp3
- Arc15
- Arc18
- Arc19
- Arc35
- Arc40

‘The Arp2/3 complex is a stable multiprotein assembly required for the nucleation of actin filaments in all eukaryotic cells and consists of seven proteins in human and yeast.’

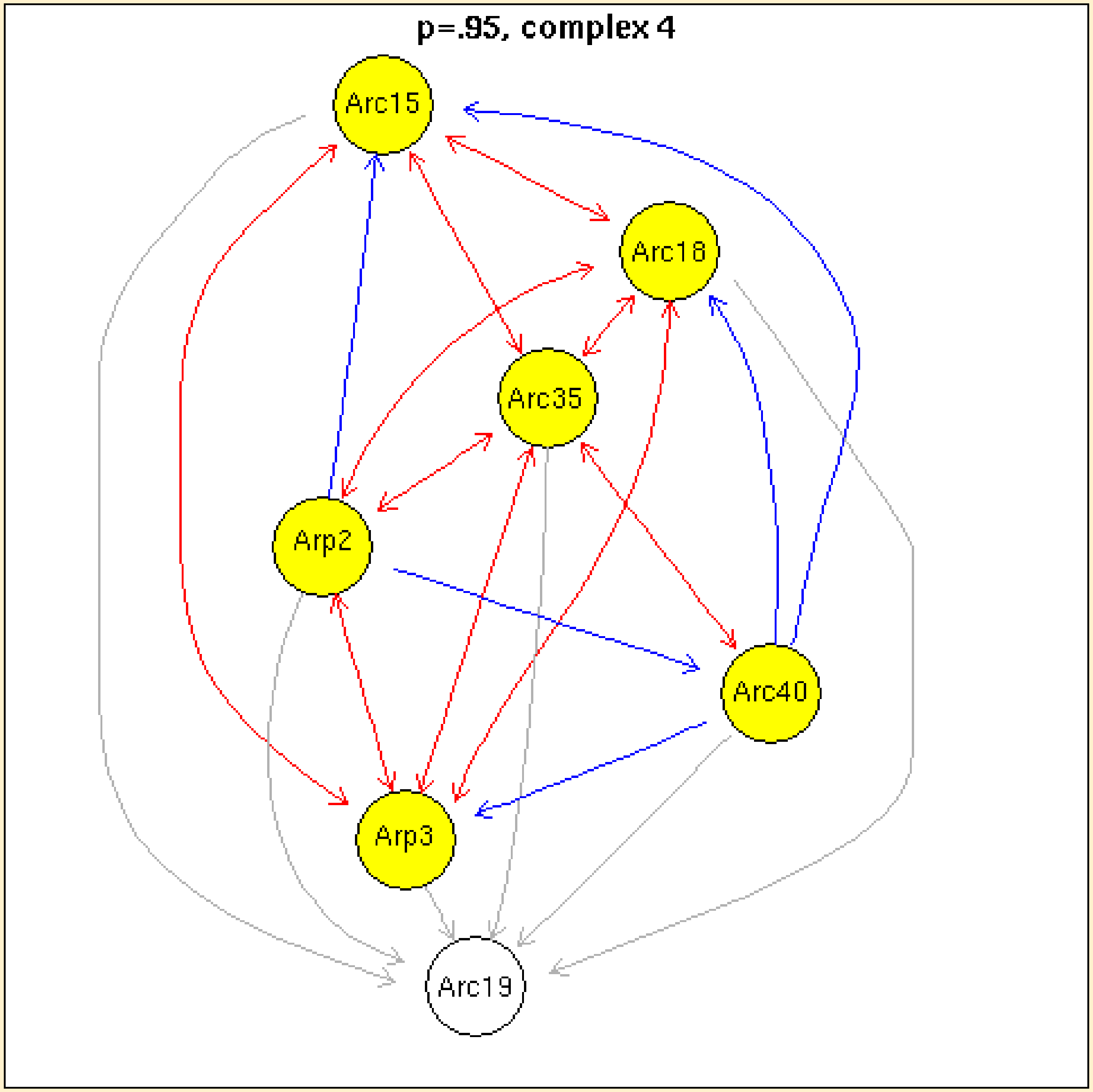
Winter, et al (1997). *Curr Biol.*
Higgs and Pollard (2001). *Annu Rev Biochem.*



Arp2/3

Arp2/3 complex:

- Arp2
- Arp3
- Arc15
- Arc18
- Arc19
- Arc35
- Arc40

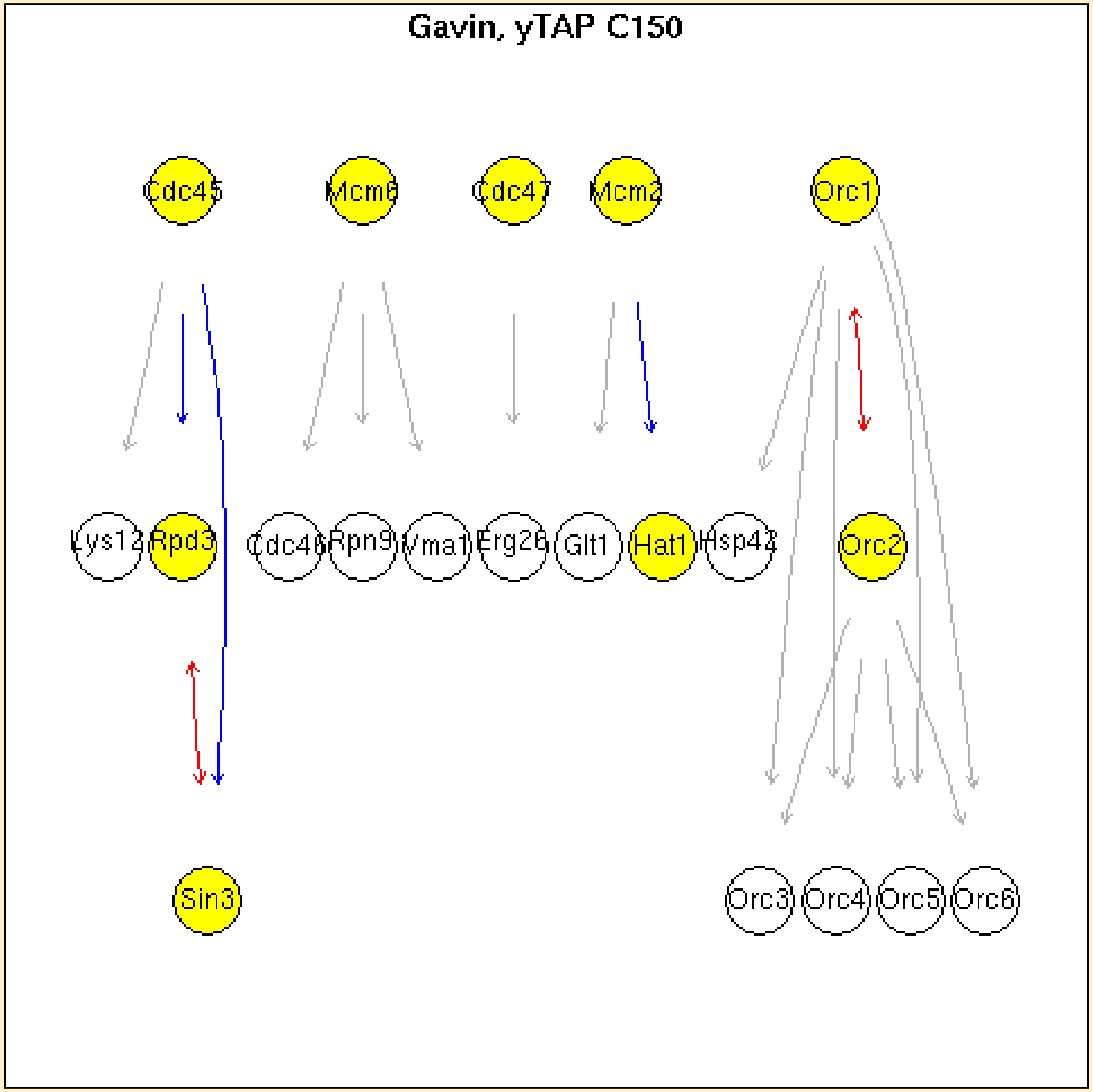


Origin Recognition Complex

Origin Recognition Complex:

Orc1
Orc2
Orc3
Orc4
Orc5
Orc6

Dutta and Bell (1997). *Annu Rev Cell Dev Biol.*

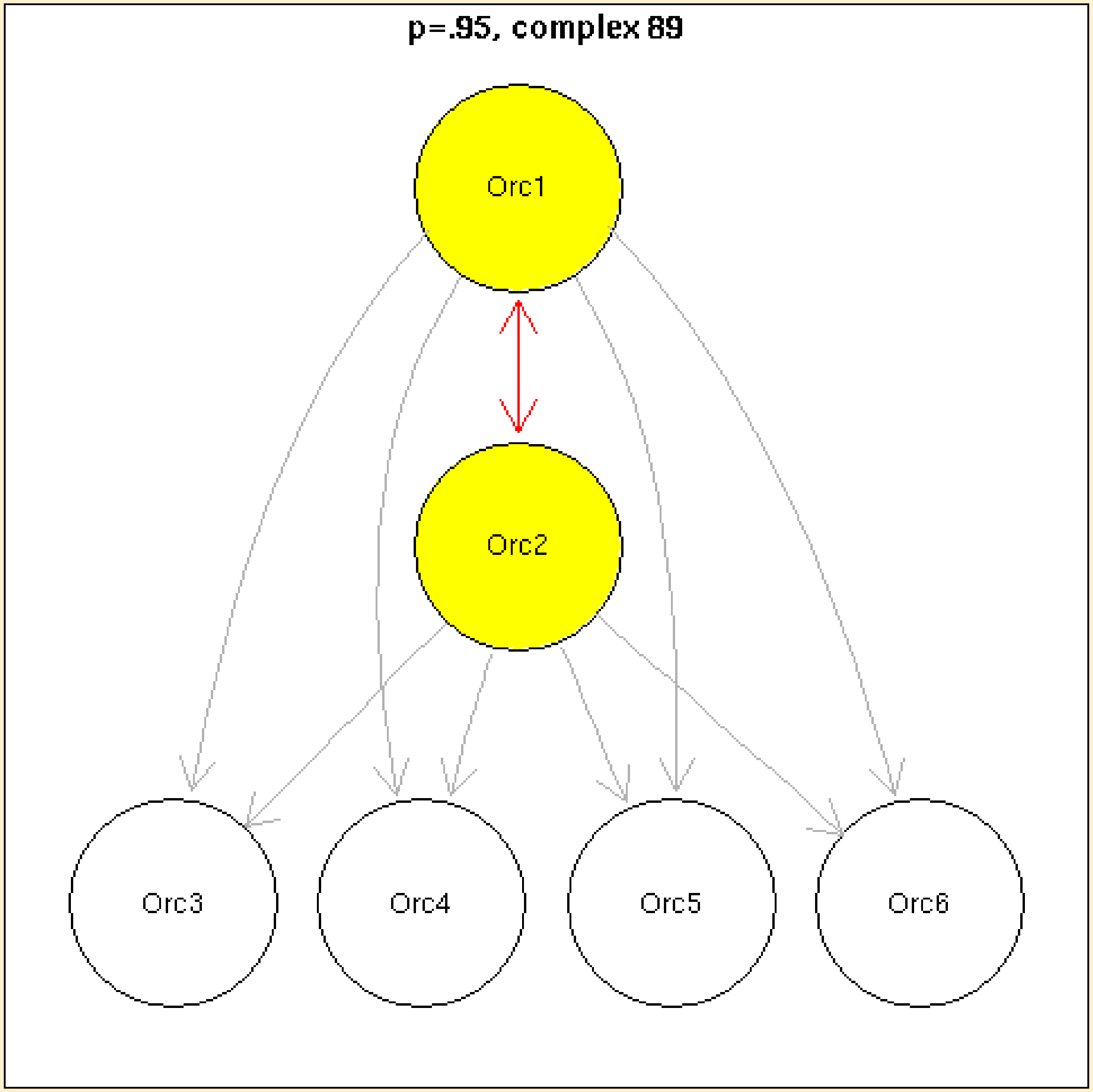


Origin Recognition Complex

Origin Recognition Complex:

Orc1
Orc2
Orc3
Orc4
Orc5
Orc6

Dutta and Bell (1997). *Annu Rev Cell Dev Biol.*

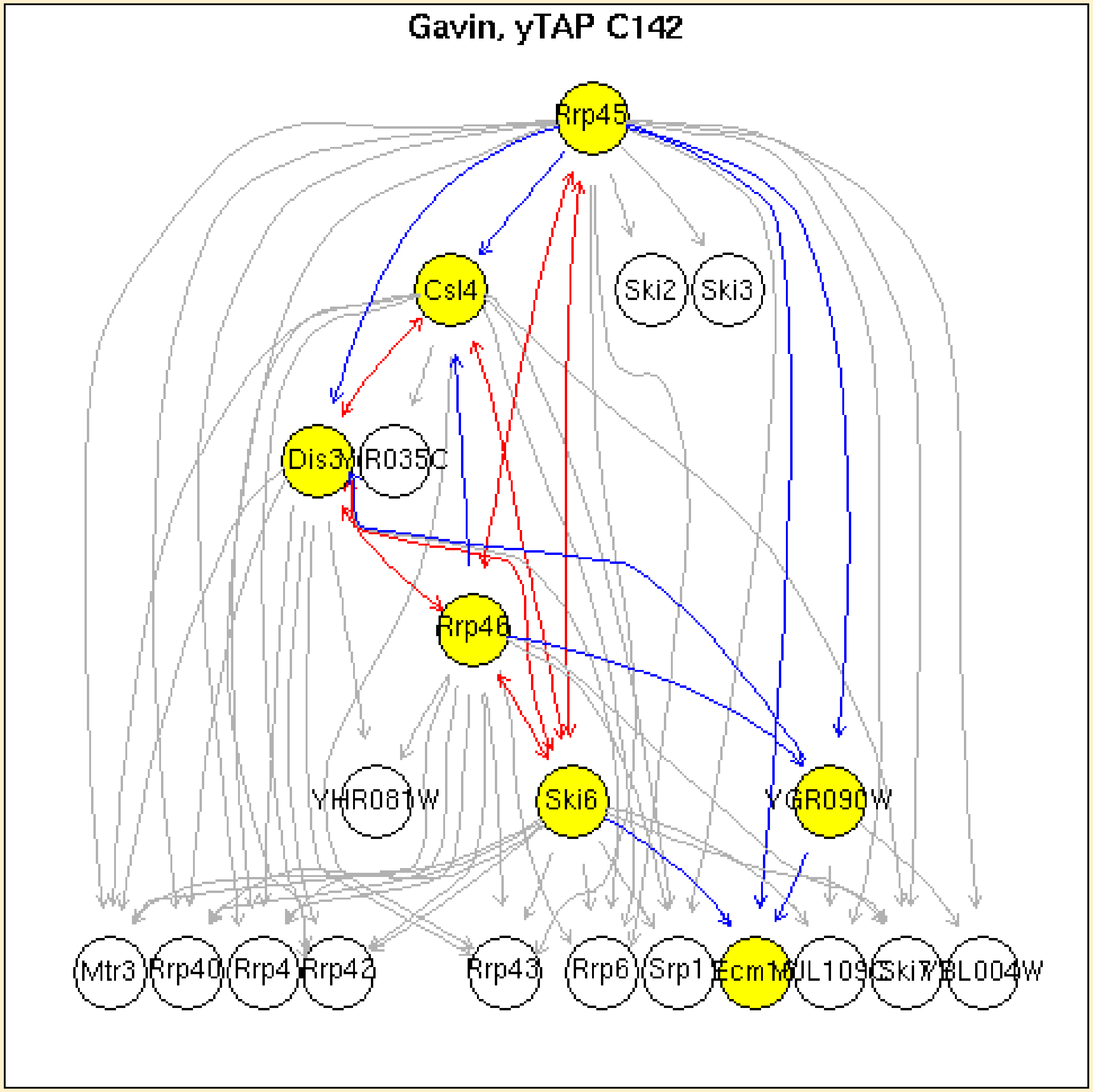


Exosome

Exosome:

Rrp4
Rrp41 (Ski6)
Rrp42
Rrp43
Rrp44 (Dis3)
Rrp45
Rrp46
Mtr3
Rrp40
Csl4
Rrp6 (only in nuclear exosome)

Allmang, et al (1999). *Genes Devel.*

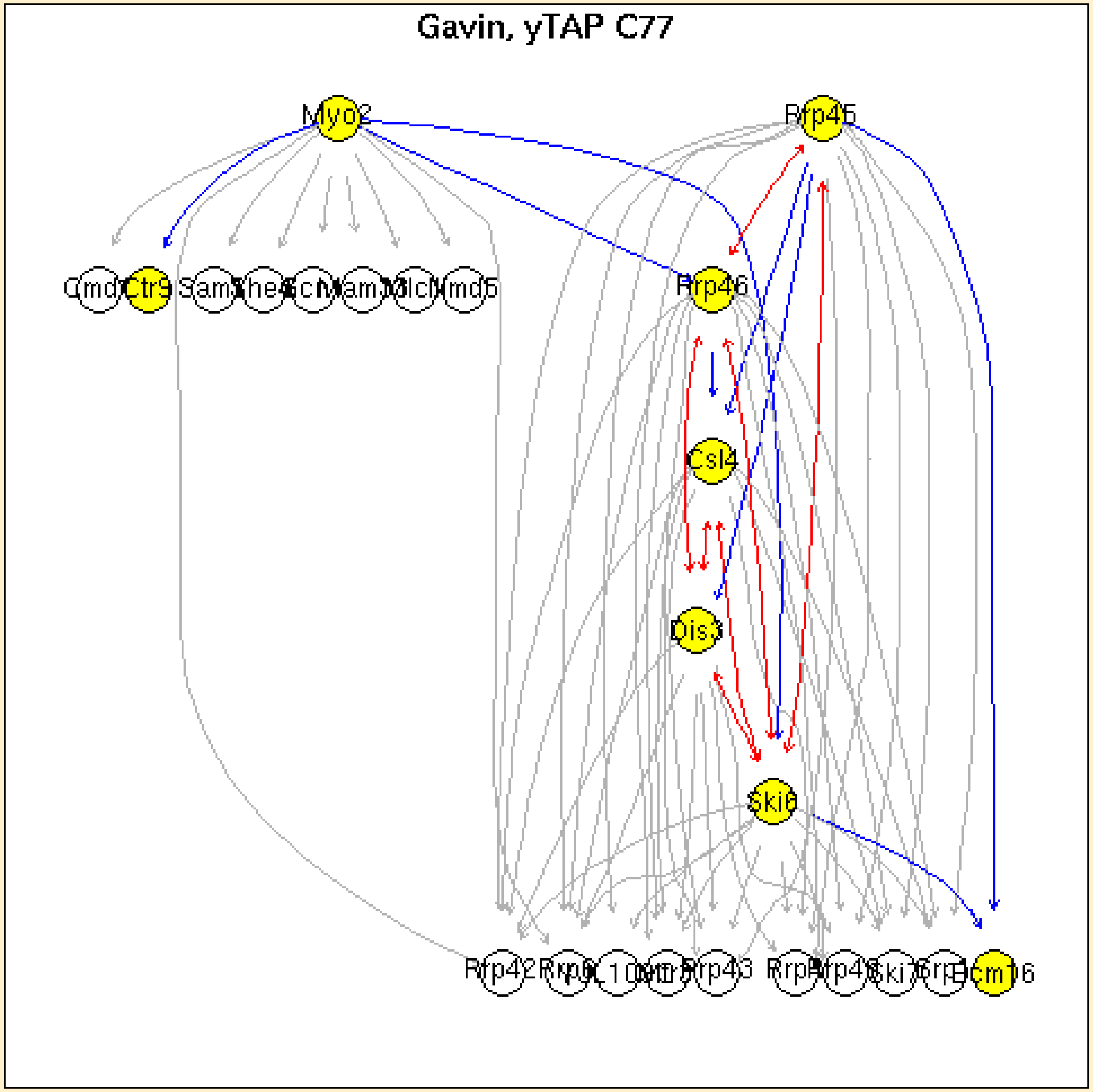


Exosome

Exosome:

Rrp4
Rrp41 (Ski6)
Rrp42
Rrp43
Rrp44 (Dis3)
Rrp45
Rrp46
Mtr3
Rrp40
Csl4
Rrp6 (only in nuclear exosome)

Allmang, et al (1999). *Genes Devel.*

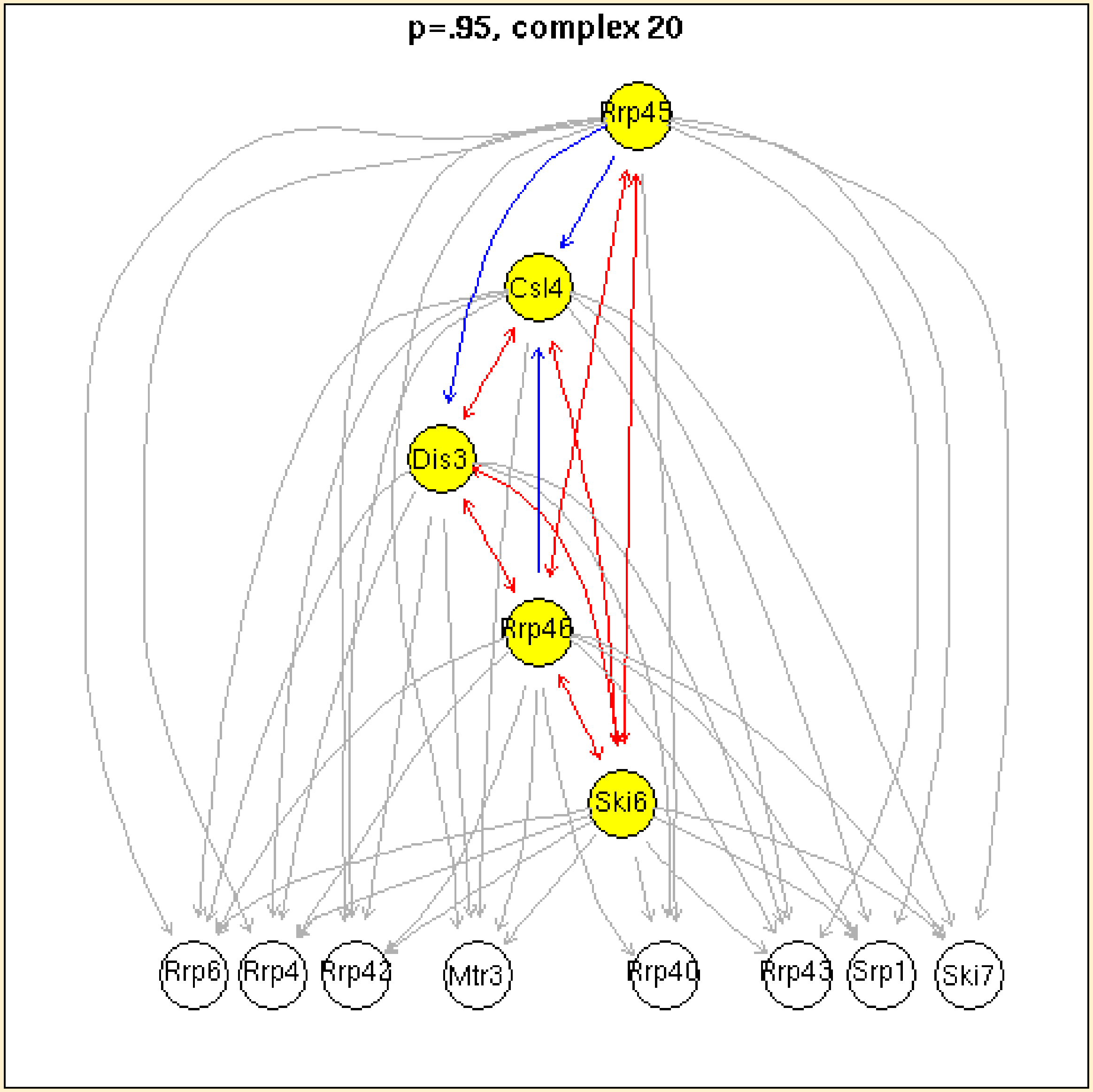


Exosome

Exosome:

Rrp4
Rrp41 (Ski6)
Rrp42
Rrp43
Rrp44 (Dis3)
Rrp45
Rrp46
Mtr3
Rrp40
Csl4
Rrp6 (only in nuclear exosome)

Allmang, et al (1999). *Genes Devel.*



PP2A

Heterotrimeric complex consisting of:

Tpd3

- regulatory A subunit

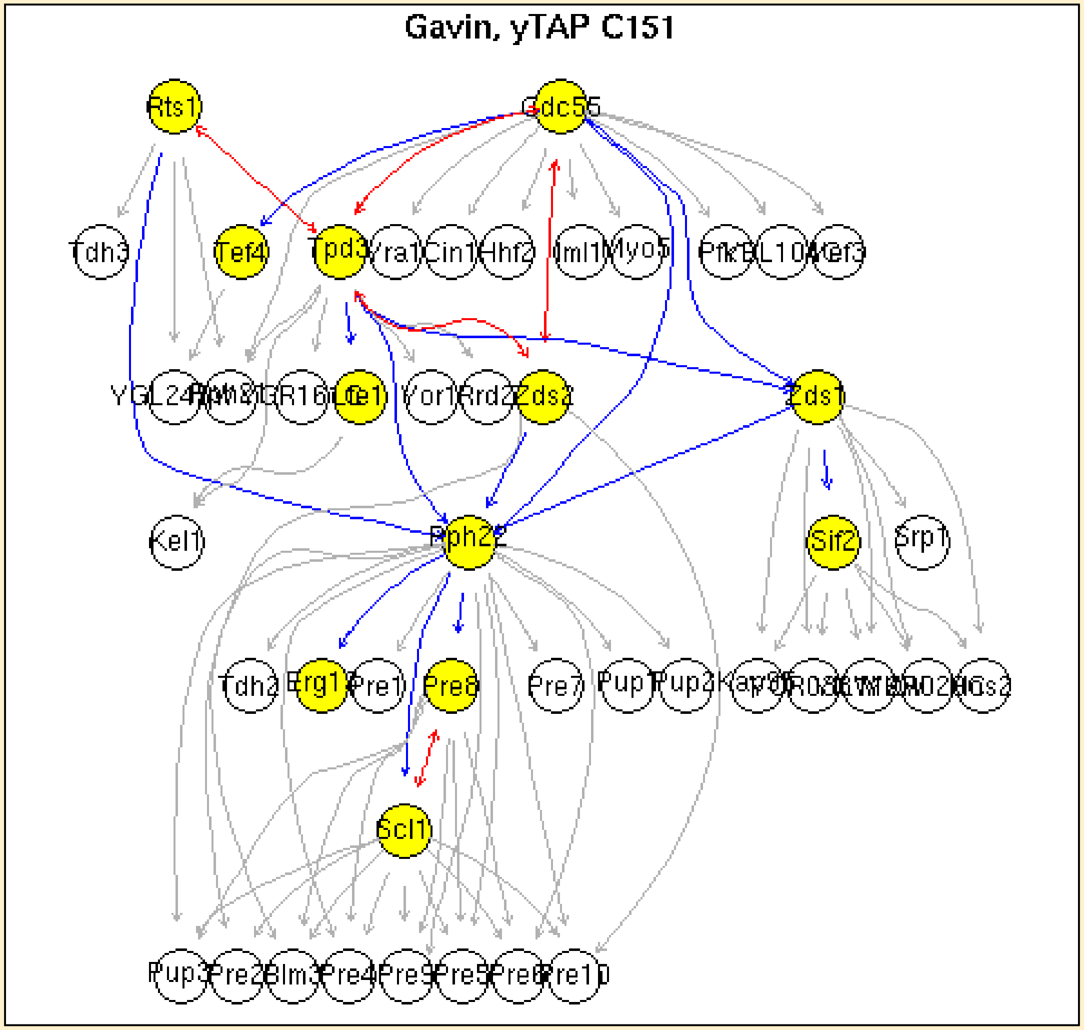
Cdc55 ***or*** **Rts1**

- regulatory B subunits

Pph21 ***or*** **Pph22**

- catalytic subunits

Jiang and Broach (1999). *EMBO.*



PP2A

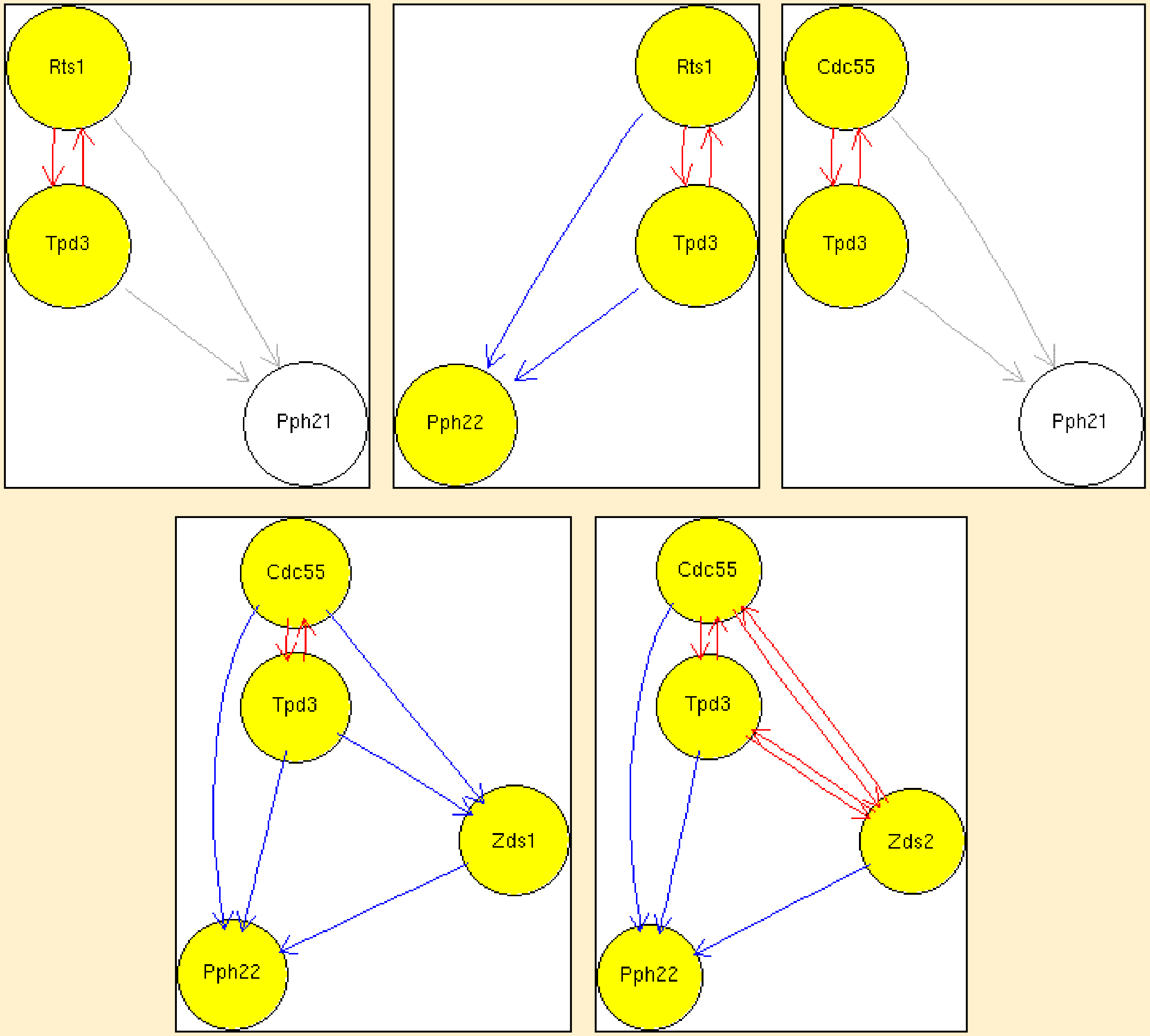
Heterotrimeric complex consisting of:

Tpd3
- regulatory A subunit

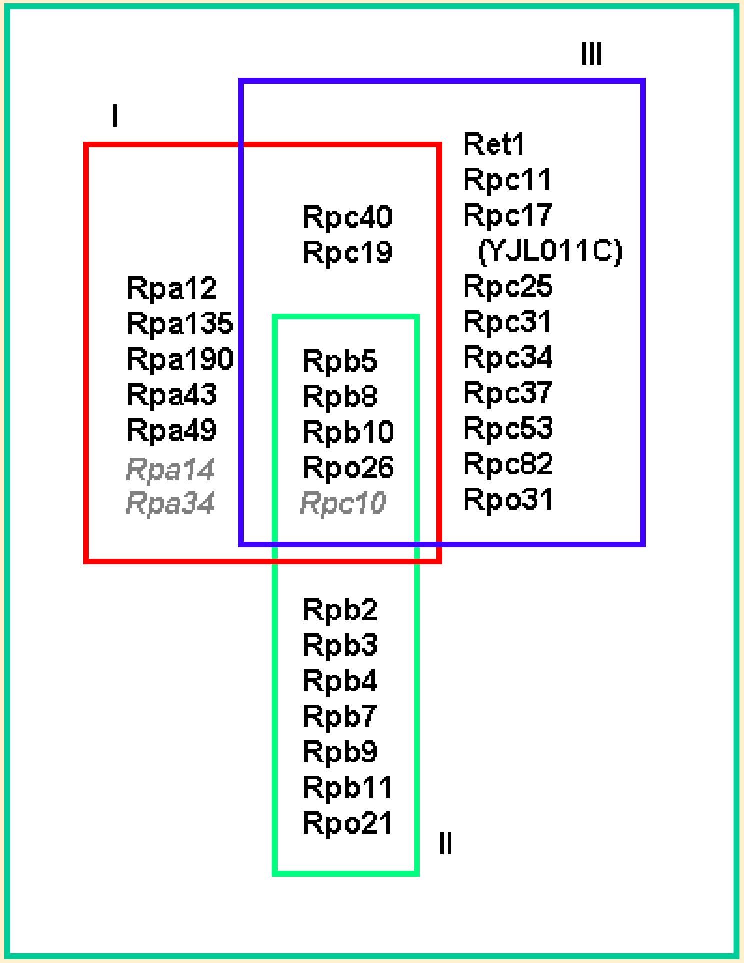
Rts1 *or* Cdc55
- regulatory B subunits

Pph21 *or* Pph22
- catalytic subunits

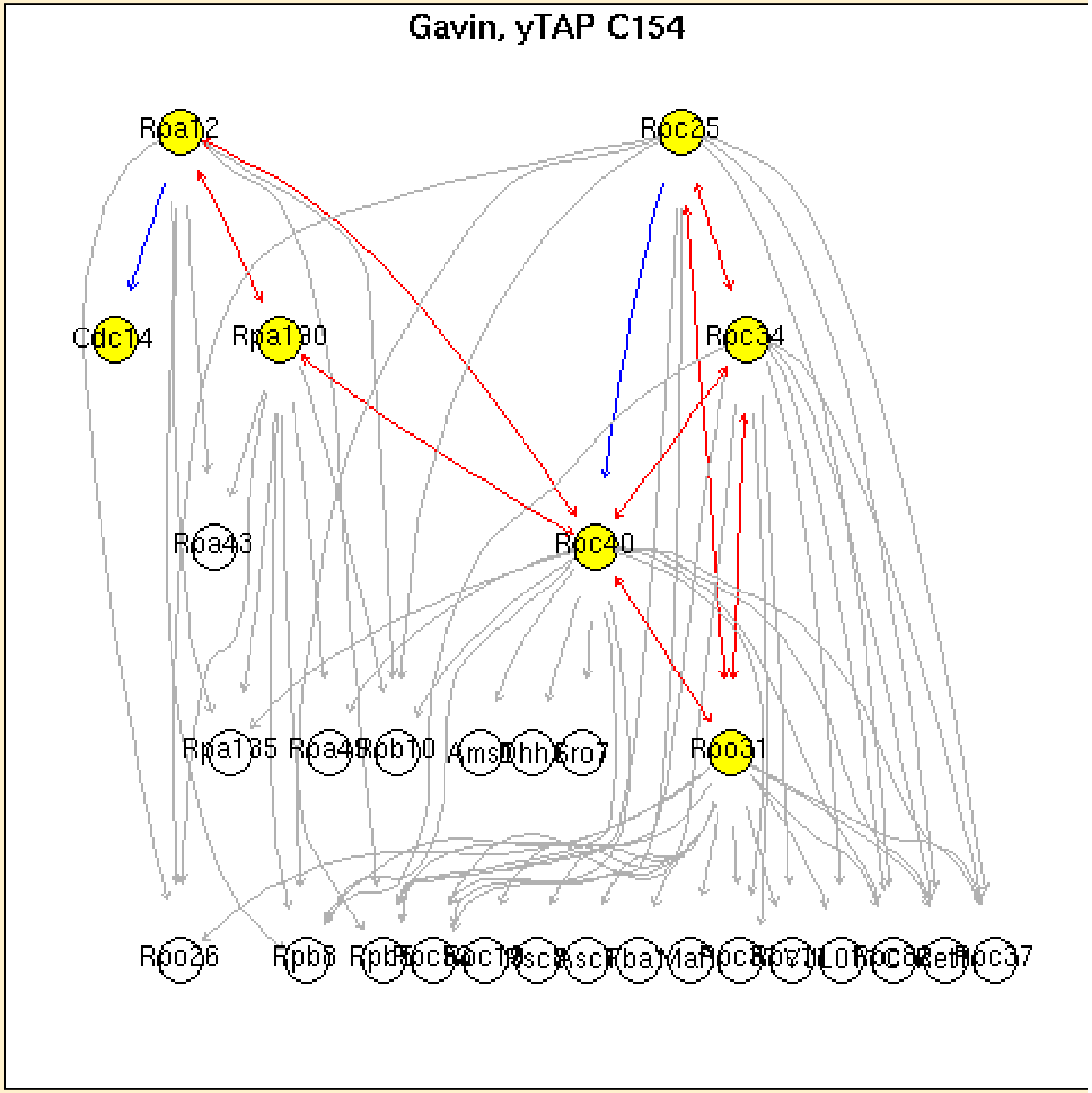
Jiang and Broach (1999). *EMBO.*



RNA Polymerases I, II and III

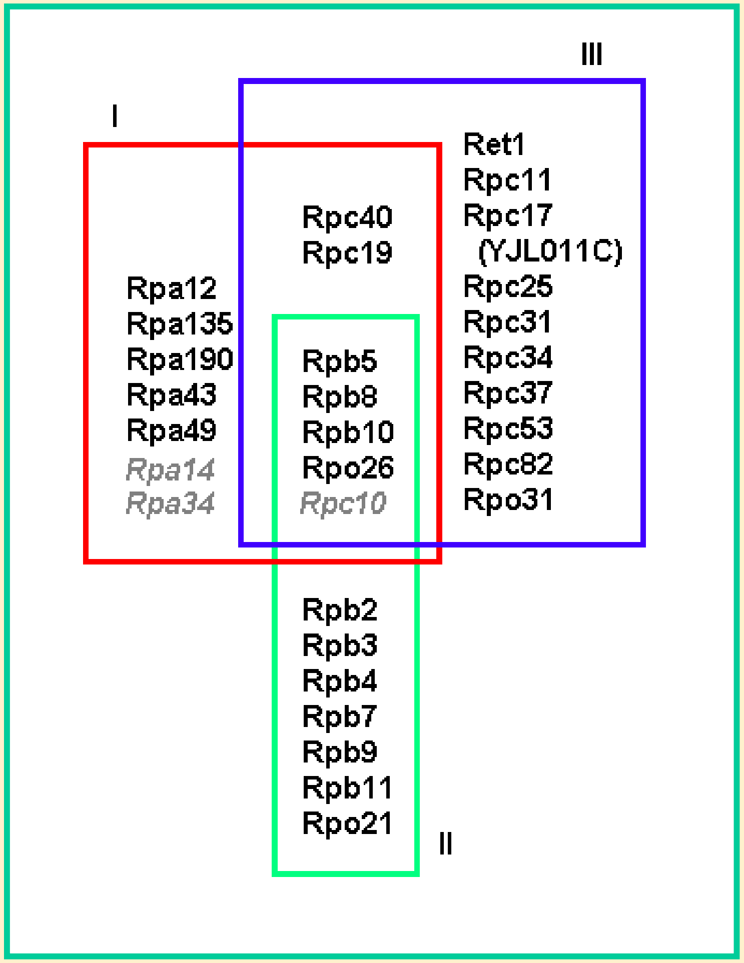


Archambault and Friesen (1993). *Microbiol Rev.*
Myer and Young (1998). *J Biol Chem.*
Smid, et al (1995). *J Biol Chem.*
Ferri, et al (2000). *Mol Cell Biol.*



Gavin, yTAP C154

RNA Polymerases I, II and III



I

Rpa12
Rpa135
Rpa190
Rpa43
Rpa49
Rpa14
Rpa34

III

Rpc40
Rpc19

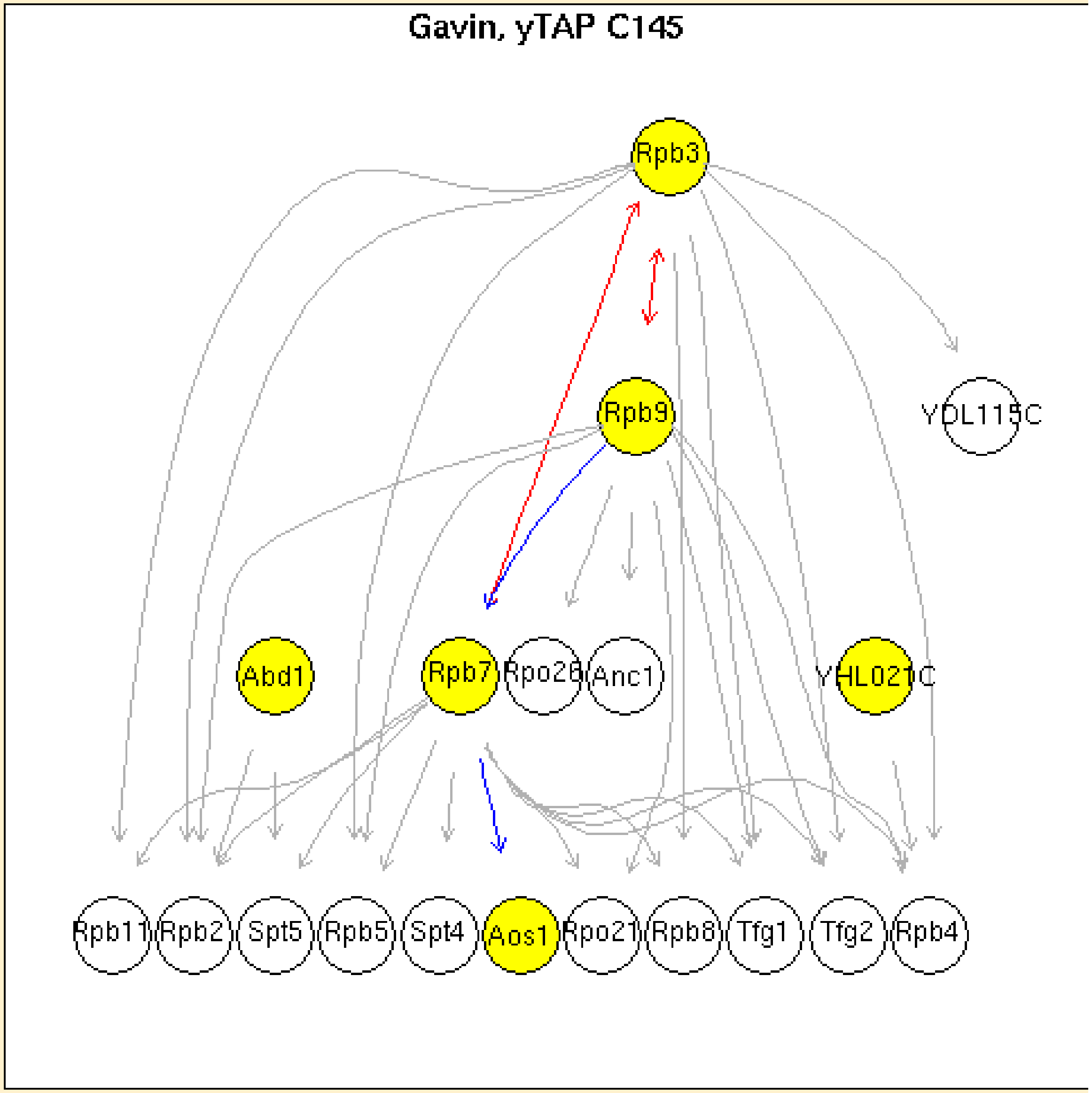
Ret1
Rpc11
Rpc17
(YJL011C)
Rpc25
Rpc31
Rpc34
Rpc37
Rpc53
Rpc82
Rpo31

Rpb5
Rpb8
Rpb10
Rpo26
Rpc10

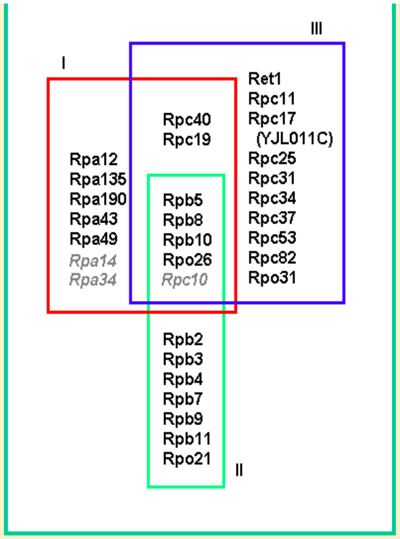
Rpb2
Rpb3
Rpb4
Rpb7
Rpb9
Rpb11
Rpo21

II

Gavin, yTAP C145



RNA Polymerases I and III



I

III

Rpa12
Rpa135
Rpa190
Rpa43
Rpa49
Rpa14
Rpa34

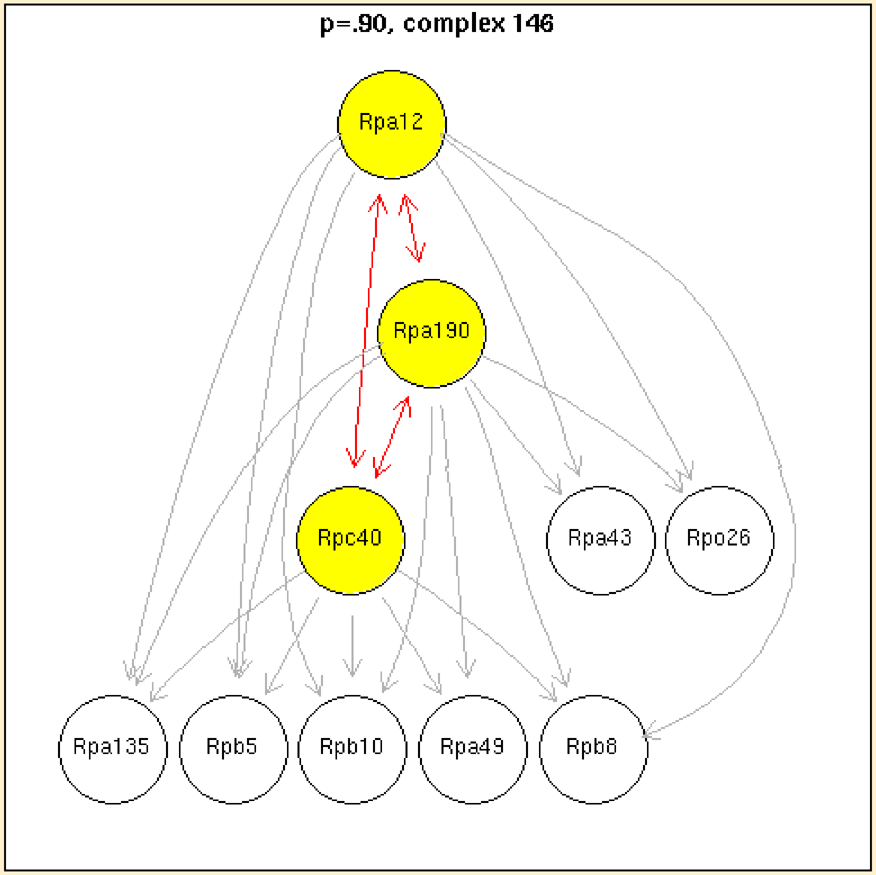
Rpc40
Rpc19

Rpb5
Rpb8
Rpb10
Rpo26
Rpc10

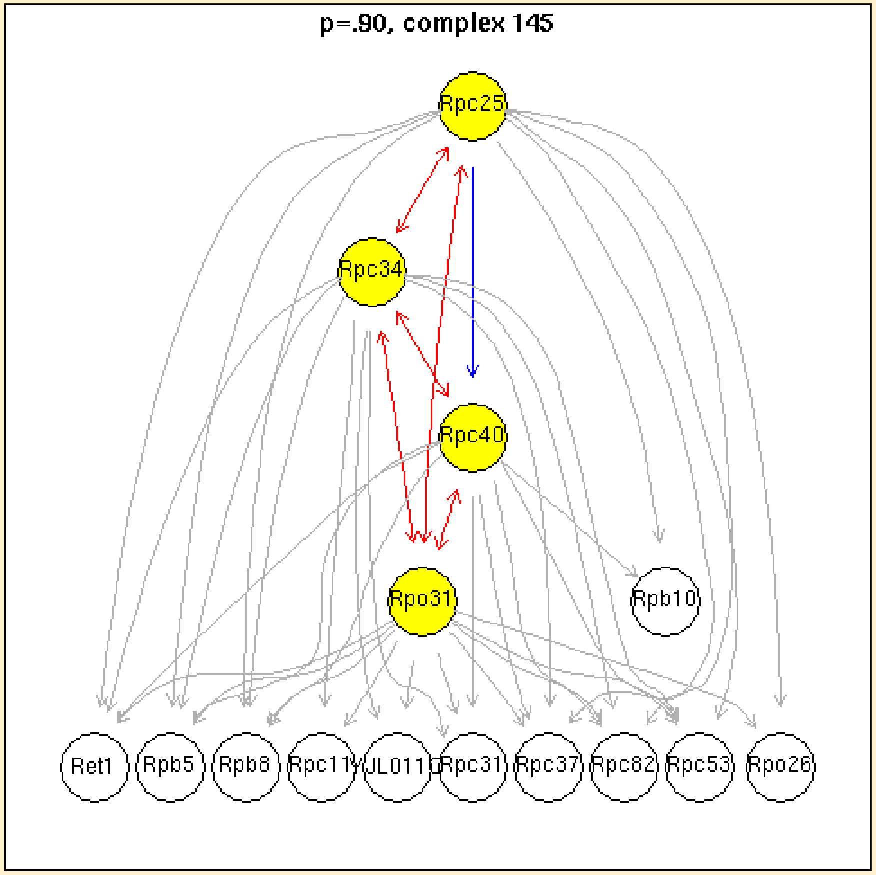
Ret1
Rpc11
Rpc17
(YJL011C)
Rpc25
Rpc31
Rpc34
Rpc37
Rpc53
Rpc82
Rpo31

Rpb2
Rpb3
Rpb4
Rpb7
Rpb9
Rpb11
Rpo21

II

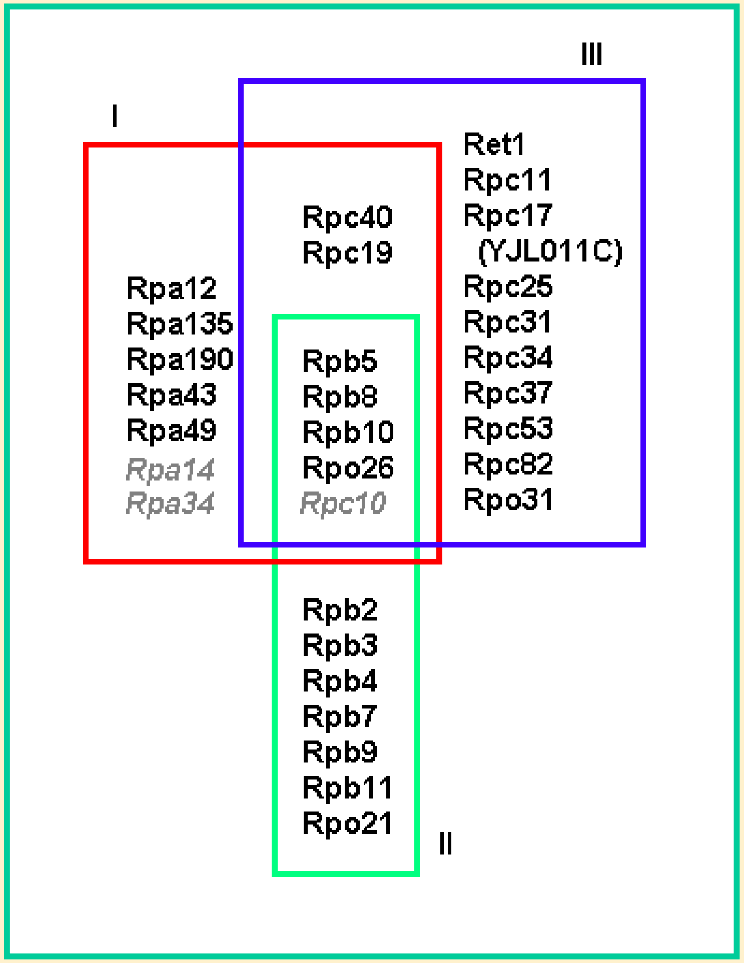


p=.90, complex 146



p=.90, complex 145

RNA Polymerase II



III

I

Rpa12
Rpa135
Rpa190
Rpa43
Rpa49
Rpa14
Rpa34

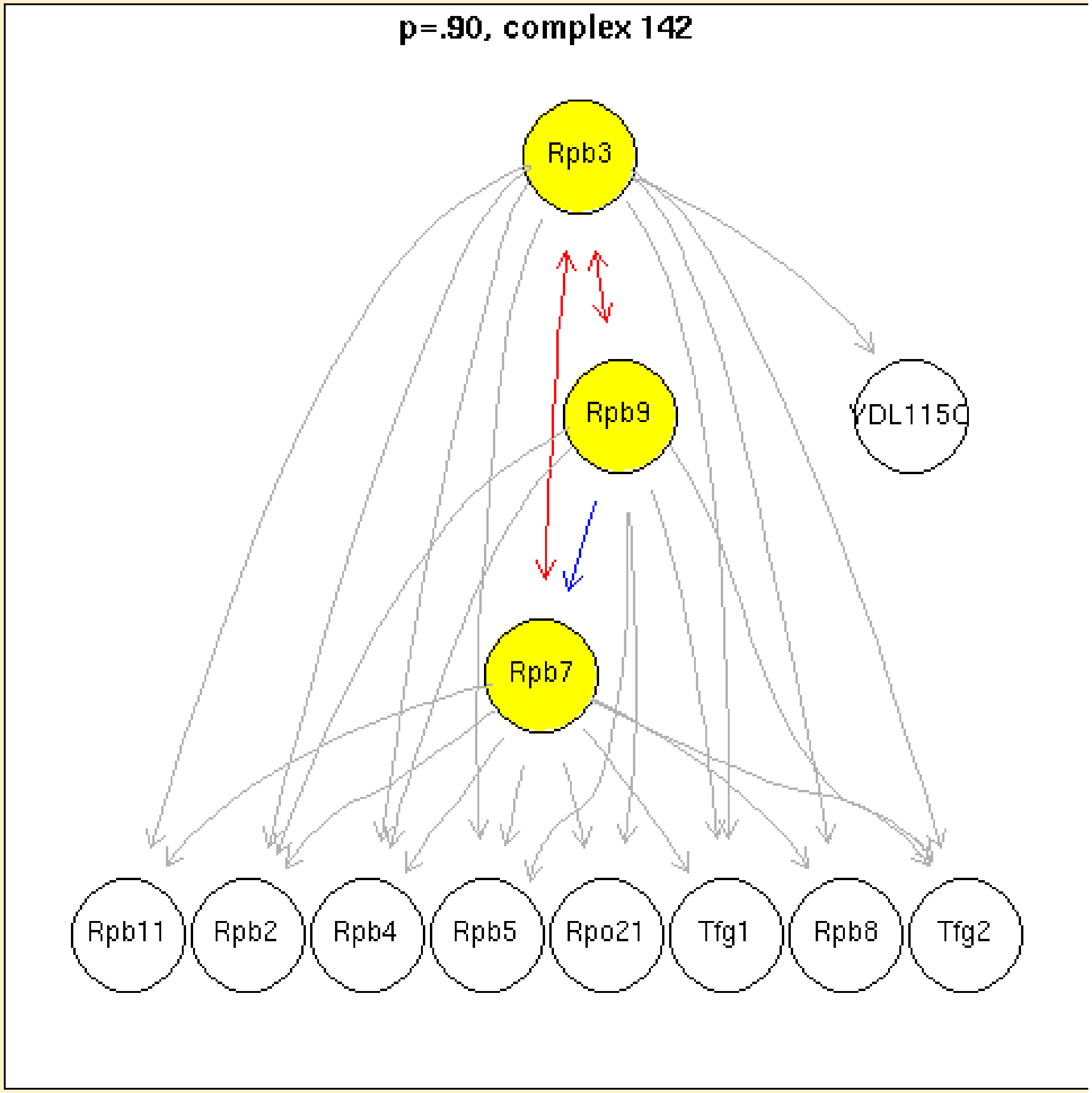
Rpc40
Rpc19

Rpb5
Rpb8
Rpb10
Rpo26
Rpc10

Ret1
Rpc11
Rpc17
(YJL011C)
Rpc25
Rpc31
Rpc34
Rpc37
Rpc53
Rpc82
Rpo31

Rpb2
Rpb3
Rpb4
Rpb7
Rpb9
Rpb11
Rpo21

II



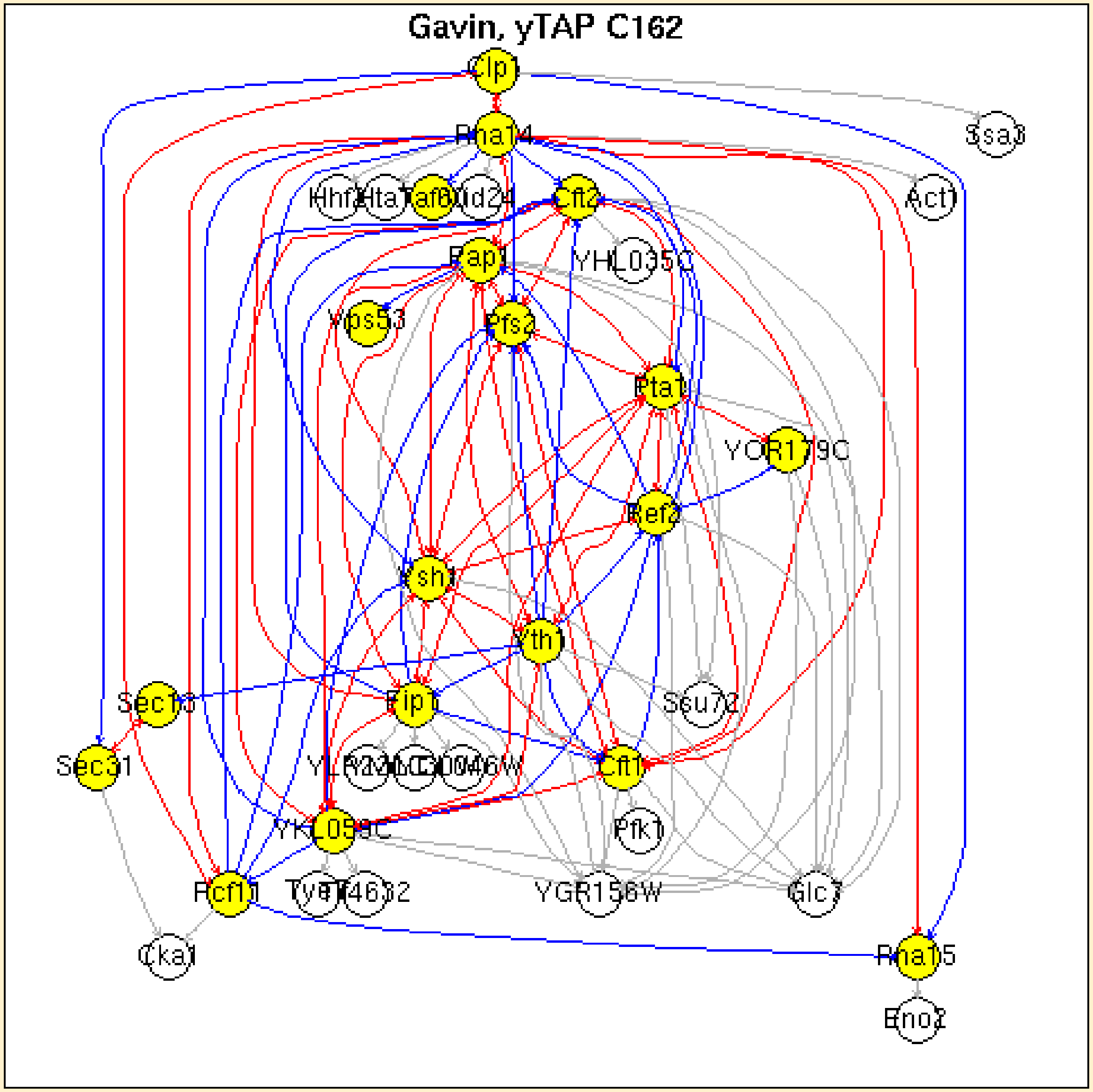
p=.90, complex 142

mRNA cleavage and polyadenylation

CF I:	PF I:
Rna14	**Cft1**
Rna15	**Cft2**
Pcf11	**Ysh1 (Brr5)**
ClpI	**Pta1**
Hrp1	**Fip1**
	Pfs2
	Yth1
	YKL059C (Mpe1)
	YGR156W (Pti1)
	Pap1
	Pfs1

-Hrp1 is CFIB – a separate component that shuttles between the nucleus and cytoplasm
-CF II is Cft1, Cft2, Ysh1, Pta1
-Yeast requires the cooperativity of CFI & PFI
-Pfs2 and Rna14 exhibit an in vitro interaction

Gross and Moore (2001). *PNAS.*
Zhao, et al (1997). *J Biol Chem.*
Skaar and Greenleaf (2002) *Mol Cell.*
Vo, et al (2001). *Mol Cell Biol.*



mRNA cleavage and polyadenylation

CF I:

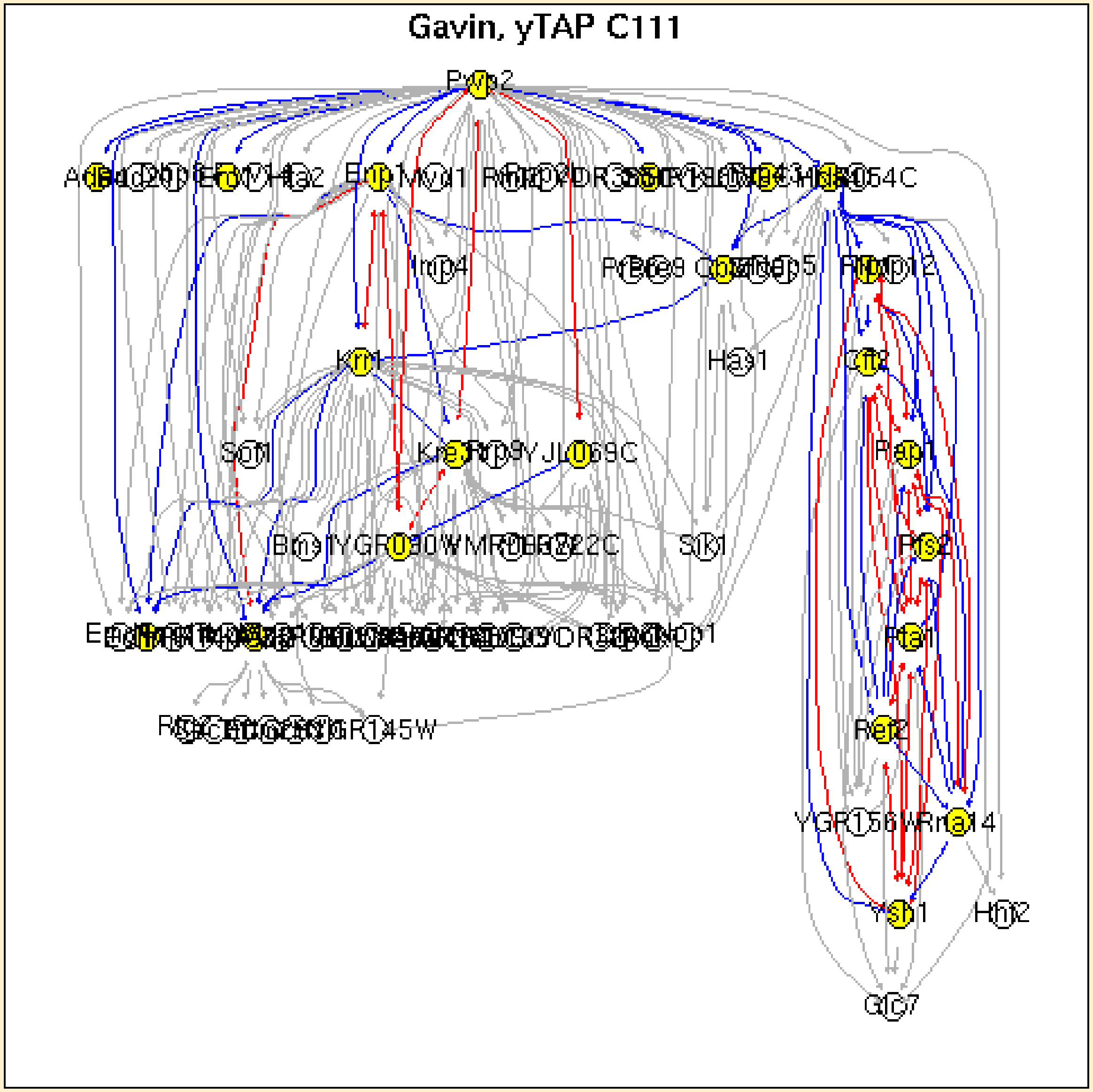
Rna14
Rna15
Pcf11
ClpI
Hrp1

PF I:

Cft1
Cft2
Ysh1 (Brr5)
Pta1
Fip1
Pfs2
Yth1
YKL059C (Mpe1)
YGR156W (Pti1)
Pap1
Pfs1

-Hrp1 is CFIB – a separate component that shuttles between the nucleus and cytoplasm
-CF II is Cft1, Cft2, Ysh1, Pta1
-Yeast requires the cooperativity of CFI & PFI
-Pfs2 and Rna14 exhibit an in vitro interaction

Gross and Moore (2001). *PNAS.*
Zhao, et al (1997). *J Biol Chem.*
Skaar and Greenleaf (2002) *Mol Cell.*
Vo, et al (2001). *Mol Cell Biol.*



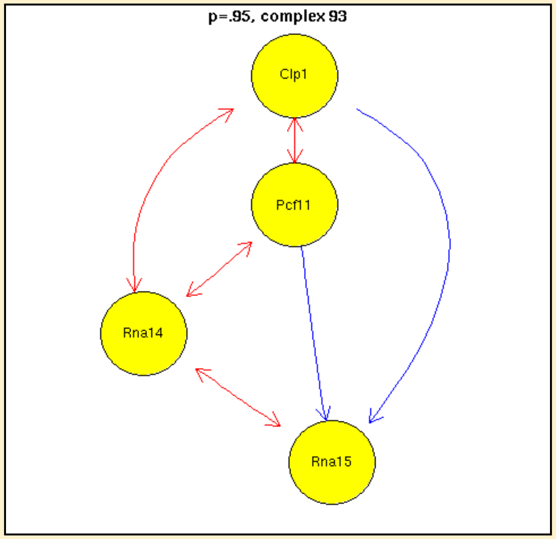
mRNA cleavage and polyadenylation

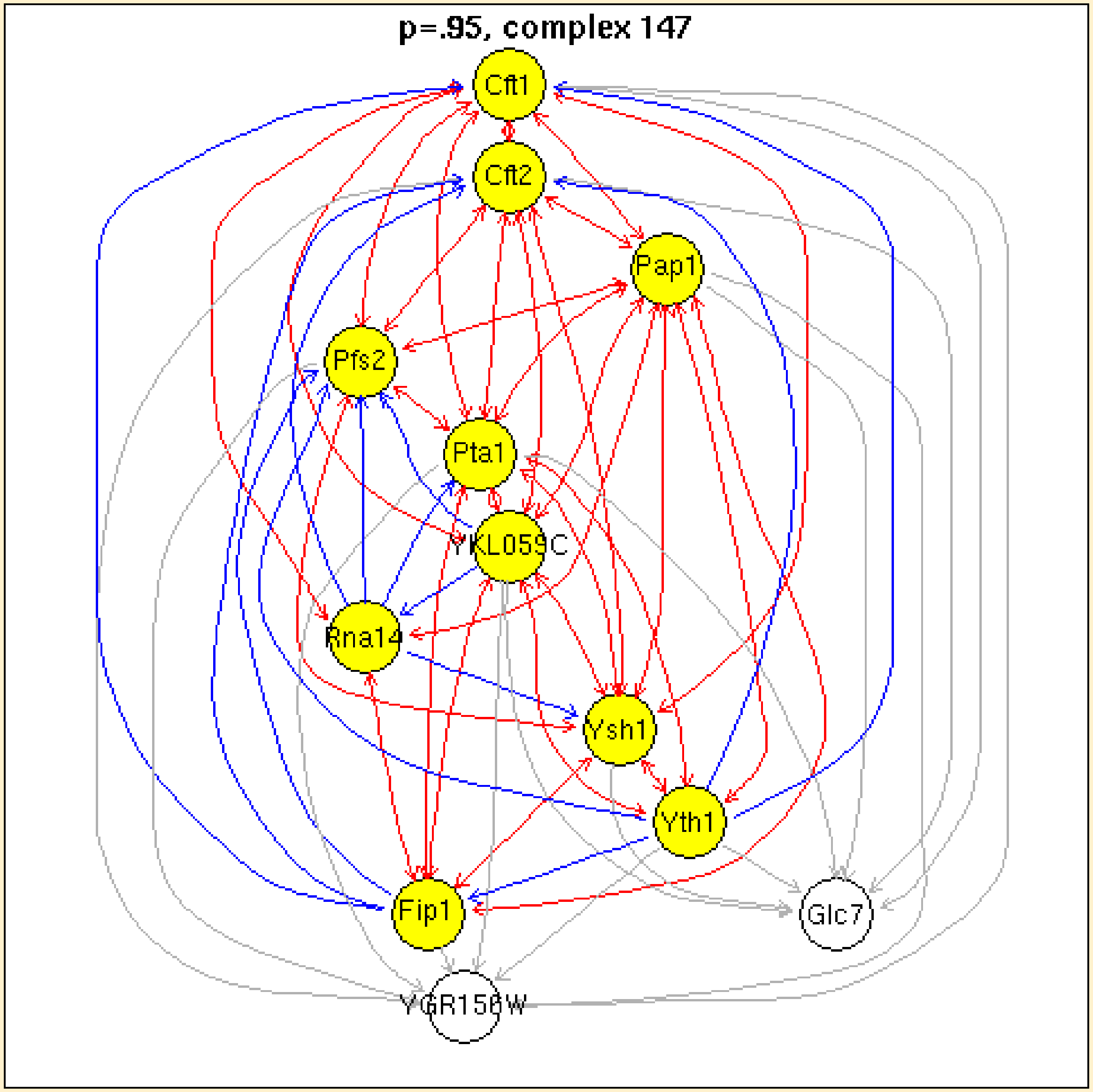
CF I:

Rna14
Rna15
Pcf11
ClpI
Hrp1

PF I:

Cft1
Cft2
Ysh1 (Brr5)
Pta1
Fip1
Pfs2
Yth1
YKL059C (Mpe1)
YGR156W (Pti1)
Pap1
Pfs1



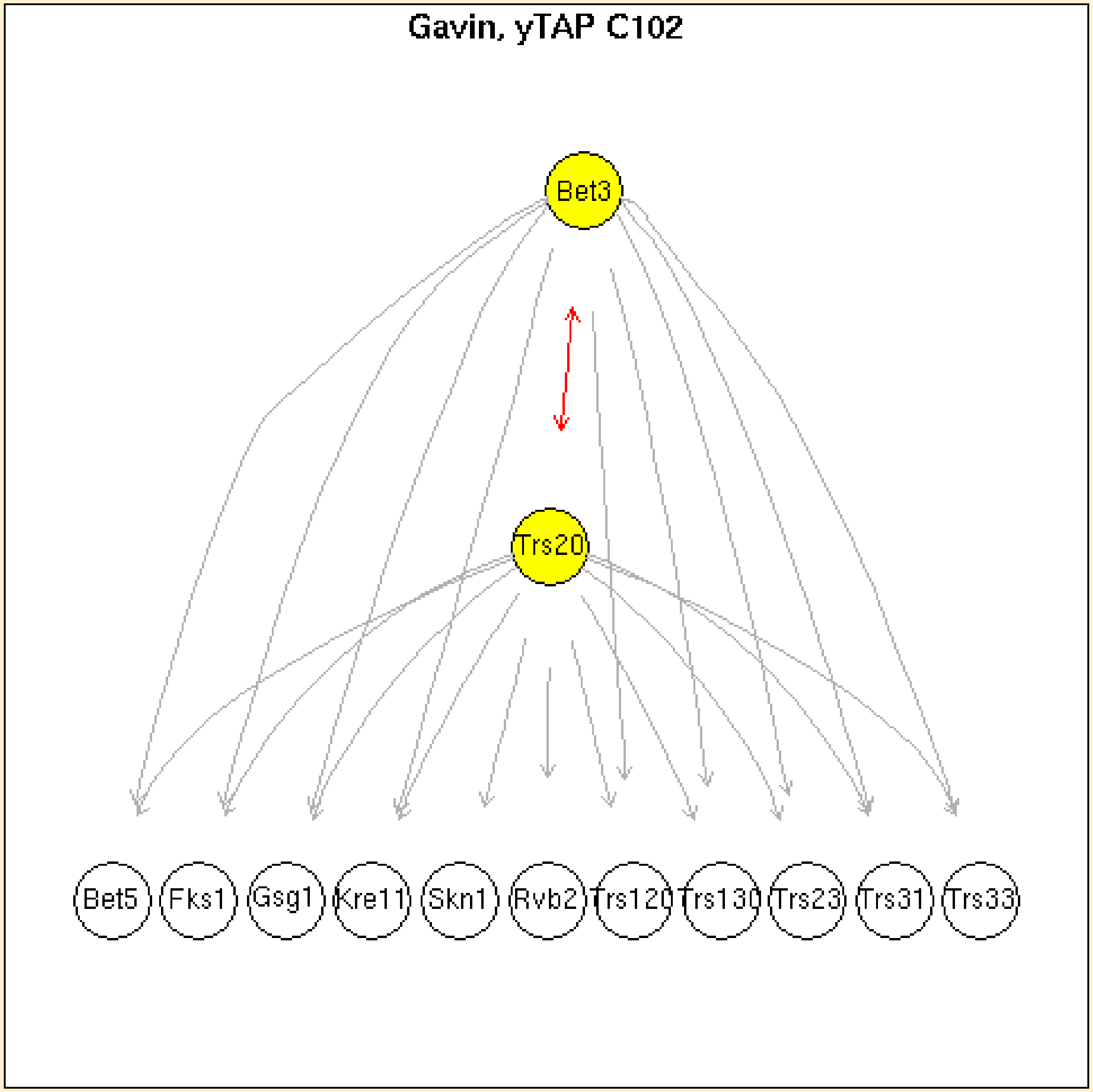


TRAPP

TRAPP:

- Bet3
- Trs20
- Bet5
- Trs23
- Trs33
- Trs31
- Trs65 (Kre11)
- Trs85 (Gsg1)
- Trs120
- Trs130

Sacher, et al (2000). *EJCB.*

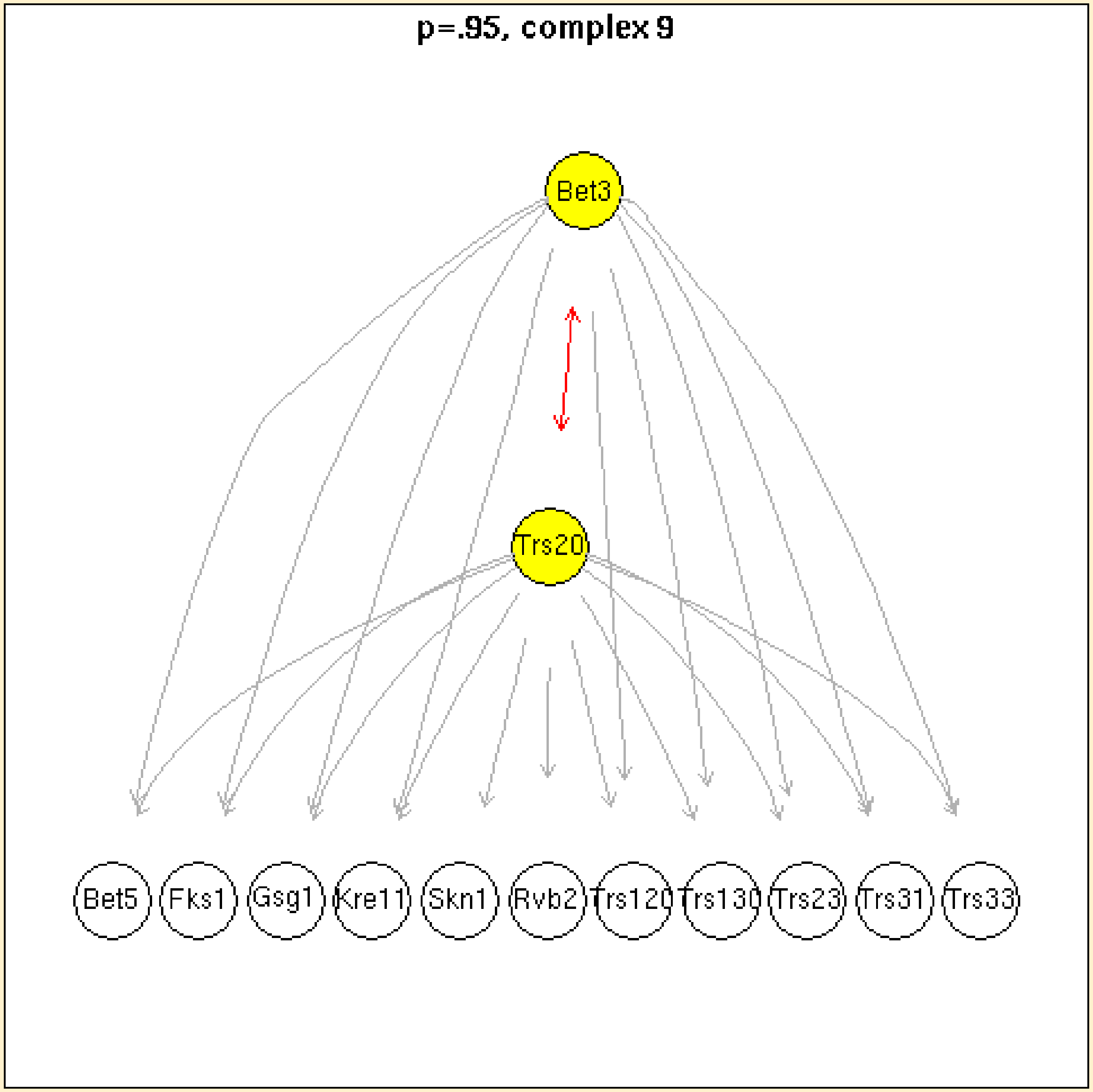


TRAPP

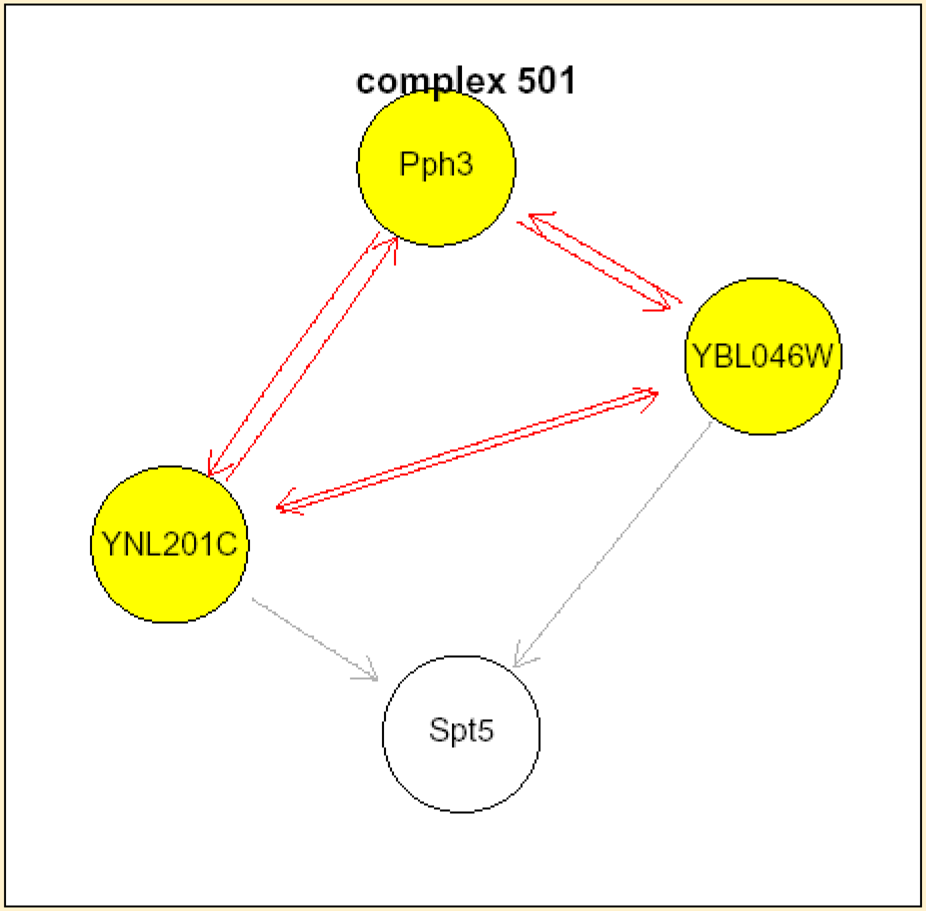
TRAPP:

- Bet3
- Trs20
- Bet5
- Trs23
- Trs33
- Trs31
- Trs65 (Kre11)
- Trs85 (Gsg1)
- Trs120
- Trs130

Sacher, et al (2000). *EJCB.*

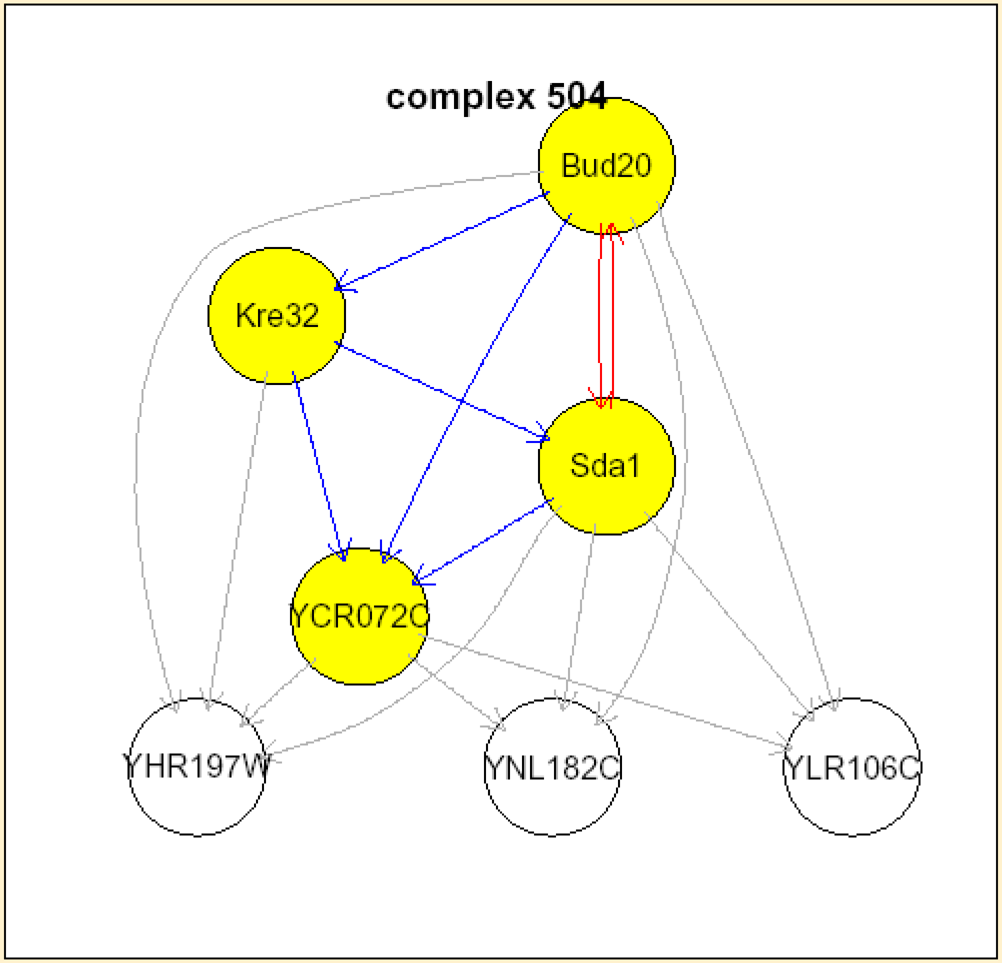


New complexes to Test?



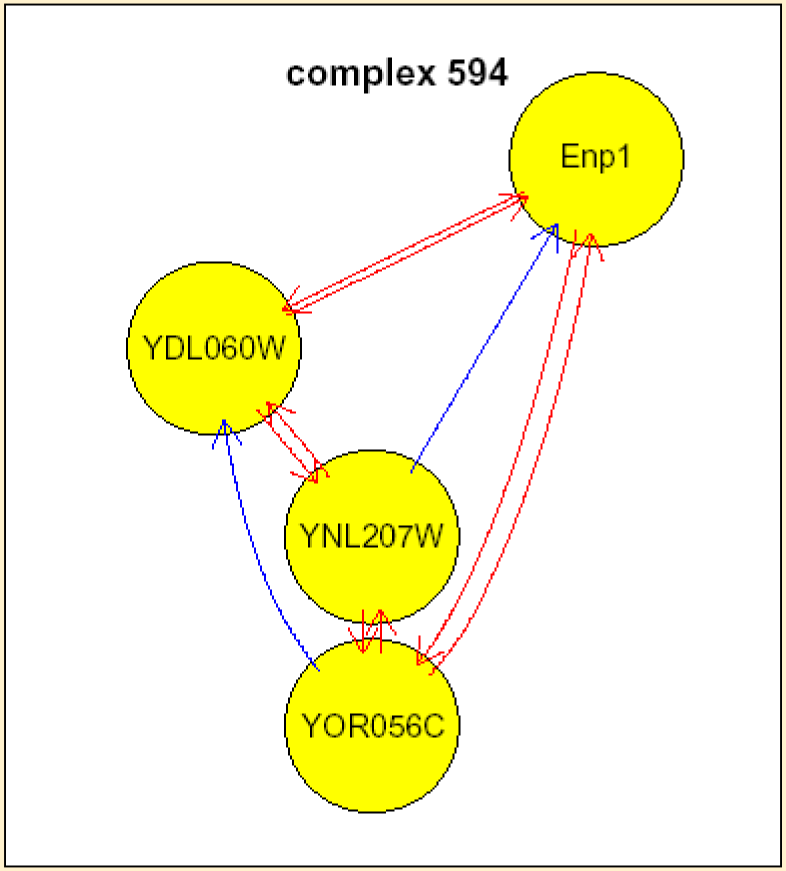
Only complex in our analysis involving these four, except for some SBMH complexes. Currently unreported in the literature.

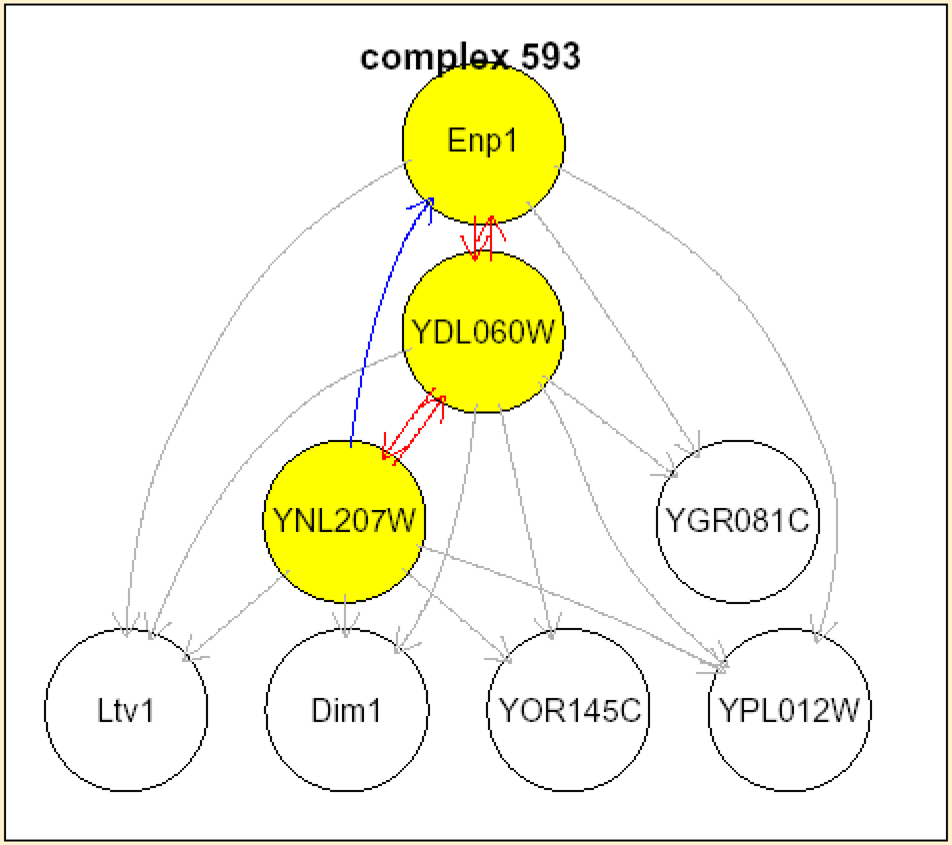
New complexes to Test?



YCR072C and Kre32 have no annotation in GO or PubMed.

New complexes to Test?





These are both undocumented in the literature – note that
Enp1, YDL060W (Tsr1), and YNL207W (Rio2) are in both complexes.

Conclusions

- Distinction between the structures of the graphs representing both the estimation goal and the available data afforded a simple complex membership estimation algorithm allowing multiple complex membership by individual proteins.
- These complex membership estimates allow a more detailed view of complexes than other analyses.

What’s Next?

- New Experiments
  - Test previously unidentified complexes
  - Mutate a gene and see what happens to its complex composition?
- Coordination with Other Data
  - Y2H data to determine physical connectivity of the proteins in a complex
  - Cell-cycle gene expression data to determine which complexes function in a cell cycle-dependent manner, and to determine the expression profile of multi-complex proteins
  - Sequence data to determine binding sites

Thanks to

- Marc Vidal, DFCI
  - Very helpful discussions about the biology
- Jeff Gentry, DFCI
  - Graph plotting software: **`Rgraphviz`**
- Jianhua Zhang, DFCI
  - Annotation package: **`yeast`**
- Vince Carey, Channing Lab
  - Helpful discussion and insights
- Members of Gentleman/Carey Lab