Protein complex membership estimation using *apComplex*

Denise Scholtens

apComplex contains functions for estimating protein complex membership using data from affinity-purification/mass-spectrometry (AP-MS) experiments. Users must specify the believed sensitivity and specificity of the AP-MS technology and may incorporate external similarity data for the proteins under investigation. The statistical details of the algorithm are reported in Scholtens and Gentleman (2004) and the biological implications are discussed in Scholtens, Vidal, and Gentleman (2005).

```
> library(apComplex)
>
```

AP-MS Data

AP-MS technology is designed to detect complex comembership among proteins. A set of proteins are used as baits, and in each purification, the bait protein finds the set of hit proteins with which it shares membership in at least one complex. Suppose proteins P_1, P_2, P_4, and P_6 compose one complex and proteins P_3, P_4 and P_5 compose a separate complex. If proteins P_1, P_2, and P_3 are used as baits, then with perfectly sensitive and specific AP-MS technology, the following data should be observed.

		Hits					
		P_1	P_2	P_3	P_4	P_5	P_6
Baits	P_1	1	1	0	1	0	1
	P_2	1	1	0	1	0	1
	P_3	0	0	1	1	1	0

The rows of the matrix are baits, the columns are hits, an entry of 1 in the ith row and jth column indicates that bait protein i finds protein j as a hit, and an entry of 0 in the ith row and jth column indicates that bait protein i does not find protein j as a hit. The diagonal entries are 1 since a protein is always a complex comember with itself. Note that bait proteins can be found as hits by other bait proteins. Also note that some proteins are never used as baits.

A graph of the data is useful for understanding which comembership relationships are tested in AP-MS experiments and which are not. In the graph in Figure 1, nodes represent proteins and directed edges from baits to hits represent complex comembership. Bait proteins are yellow and hit-only proteins (i.e. proteins that are found as hits but never used as baits) are white. Directed edges always originate at yellow bait proteins and point to the set of hits

detected by that bait. The red reciprocated edge connecting P_1 and P_2 represents a bait-bait relationship that is tested twice, once in each purification. The gray unreciprocated edges represent bait-hit-only edges that are only tested once. Missing edges between baits and other baits or hit-only proteins represent comemberships that are tested, but not observed. Edges between hit-only proteins are always missing, not because the comemberships are known not to exist, but because they are never tested.

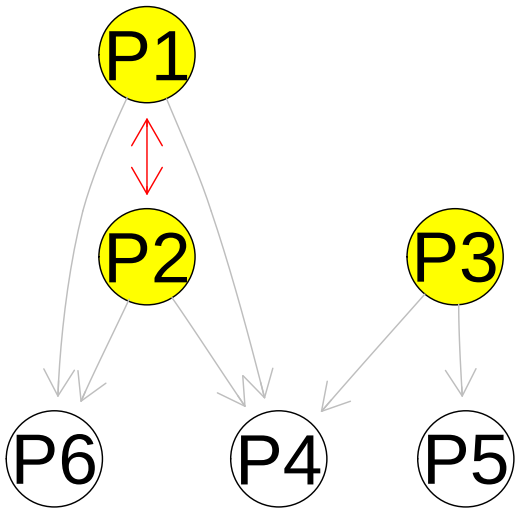

Figure 1: True complex comemberships that would be detected with perfectly sensitive and specific AP-MS technology using P_1, P_2, and P_3 as baits.

In reality, AP-MS technology is neither perfectly sensitive nor specific, resulting in false positive (FP) and false negative (FN) observations of the complex comemberships under investigation. Suppose in this experiment, we make a FN observation between P_2 and P_4, i.e. P_4 is not found as a hit when we use P_2 as a bait. Also suppose we make two FP observations: 1) when we use P_3 as a bait, we find an extraneous hit-only protein P_7, and 2) when performing a purification using P_8 as a bait, we find P_3 as a hit. Data for this example are contained in the matrix `apEX`. In this matrix, rows again represent baits and columns represent hits.

```
> data(apEX)
> apEX
```

```
   P1 P2 P3 P8 P4 P5 P6 P7
P1  1  1  0  0  1  0  1  0
P2  1  1  0  0  0  0  1  0
P3  0  0  1  0  1  1  0  1
P8  0  0  1  1  0  0  0  0
```

The graph of the data in Figure 2 demonstrates the observations recorded in `apEX`. Note the missing edge from P_2 to P_4 and the new edge from P_3 to P_7. Also note the blue unreciprocated

edge between P_3 and P_8 – this is a complex comembership that was tested twice when P_3 and P_8 were used as baits, but only detected once in the purification using P_8 as a bait.

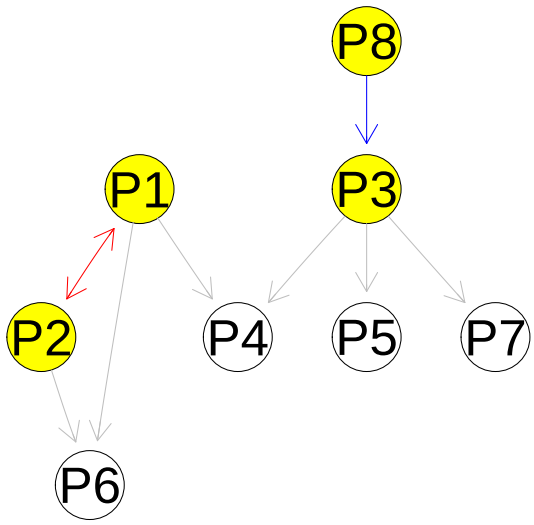

Figure 2: Hypothetical data from an AP-MS experiment with a FN observation between P_2 and P_6 and FP observations between P_3 and P_7 and P_8 and P_3.

apComplex starts with observed complex *comembership* data from AP-MS and estimates complex *membership*.

Algorithm

The edges in an AP-MS graph represent complex comembership. If all proteins were used as baits, then maximal complete subgraphs (or cliques) in the AP-MS graph would contain entire collections of proteins that compose a complex. The maximal complete subgraphs could then be used to form a protein complex membership graph (PCMG): a bipartite graph in which one set of nodes represents proteins, the other set represents complexes, and an edge from a protein node to a complex node represents membership of the protein in that complex. Bipartite graphs can also be represented using an affiliation matrix in which the rows represent proteins, the columns represent complexes, and an entry of 1 in the jth row and ith column represents membership of protein i in complex j. *apComplex* is essentially a maximal complete subgraph finding algorithm that is adapted for the bait/hit status of proteins, as well as imperfect observation of edges.

The first step in estimating complex membership is to find the *maximal BH-complete subgraphs* in the observed AP-MS data. A *BH-complete subgraph* is defined to be a collection of baits and hits for which all bait-bait edges and all bait-hit-only edges exist; a *maximal BH-complete subgraph* is a BH-complete subgraph that is not contained in any other BH-complete subgraph. In the event of unreciprocated observations between pairs of baits, the

edges are estimated to exist when the sensitivity of the AP-MS technology is less than the specificity. Under a logistic regression model where the parameters represent sensitivity and specificity, this treatment of unreciprocated bait-bait edges maximizes the likelihood L for the data (Scholtens and Gentleman, 2004). In our example, the observed data contains four maximal BH-complete subgraphs, shown in Figure 3. The function `bhmaxSubgraph` will detect these maximal BH-complete subgraphs and report them using an affiliation matrix. Figure 4 shows the corresponding bipartite PCMG for the initial complex estimates.

```
> PCMG0 <- bhmaxSubgraph(apEX)
> PCMG0

$maxCliques
$maxCliques[[1]]
[1] "P4" "P3" "P5" "P7"

$maxCliques[[2]]
[1] "P4" "P1" "P6"

$maxCliques[[3]]
[1] "P2" "P1" "P6"

$maxCliques[[4]]
[1] "P8" "P3"
```

The initial maximal BH-complete subgraph estimate of the PCMG does not allow missing edges between bait and hit-only proteins; since AP-MS technology is not perfectly sensitive, it is reasonable to expect a number of missing edges in the subgraph for each complex estimate. *apComplex* accommodates this by employing an objective function to evaluate the complex estimates. For a complex c_k, let $C(c_k)$ represent the product of 1) the binomial probability for the number of observed edges in c_k given the number of tested edges, and 2) a two-sided p-value from Fisher's exact test for the distribution of missing incoming edges for complex estimate c_k. Then let C equal the product of $C(c_k)$ over all complexes $c_1, ..., c_K$. The objective function P is the product of L and C, or $P = L \times C$. L is maximized with the initial maximal BH-complete subgraphs – the algorithm in *apComplex* looks to increase C in favor of small decreases in L.

After the initial PCMG estimate is made using `bhmaxSubgraph`, `mergeComplexes` proposes pairwise unions of individual complex estimates. If P increases when the complexes are treated as one, then the combination is accepted. If more than one union increases P, then the union with the largest increase is accepted. Algebraic details of the acceptance criteria are available in Scholtens and Gentleman (2004).

```
> PCMG1 <- mergeComplexes(PCMG0,apEX,sensitivity=.7,specificity=.75)

[1] "calculating initial penalty terms"
[1] "looking at complex combinations"

> PCMG1
```

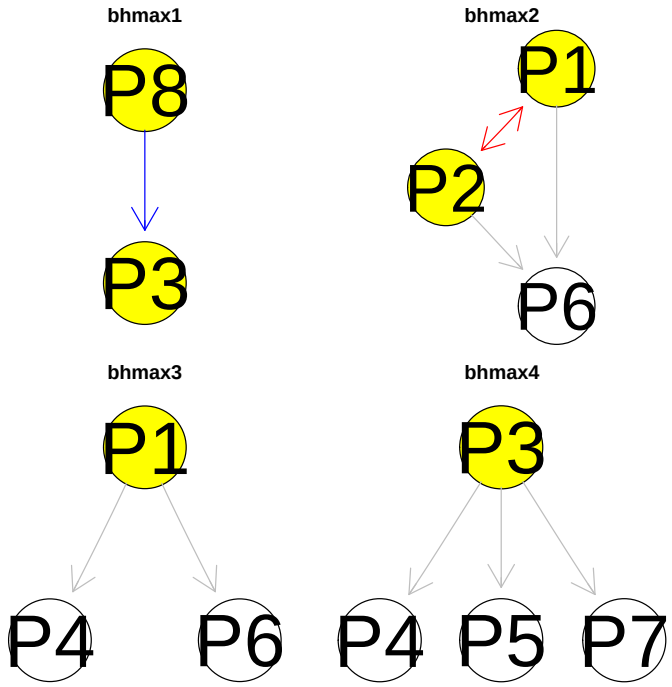

Figure 3: Four maximal BH-complete subgraphs in the observed data.

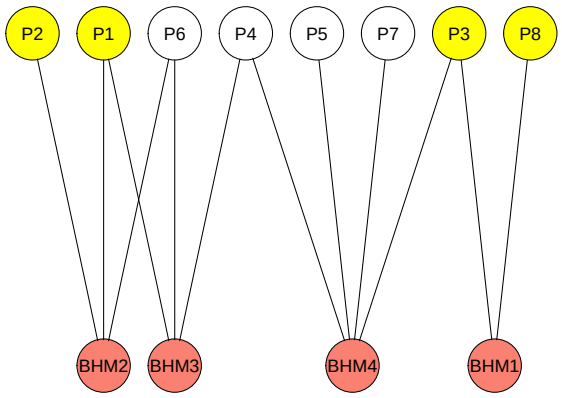

Figure 4: Bipartite PCMG for the initial complex estimates determined by locating maximal BH-complete subgraphs in the graph of observed AP-MS data.

```
$Complex1
[1] "P4" "P3" "P5" "P7"

$Complex2
[1] "P4" "P1" "P6"

$Complex3
[1] "P2" "P1" "P6"

$Complex4
[1] "P8" "P3"

>
```

In this case, `bhmax2` and `bhmax3` were combined into `Complex2`. (`bhmax1` and `bhmax4` remained as originally estimated and are now named `Complex1` and `Complex3`, respectively). The one missing edge out of the six tested in `Complex2` is consistent with the sensitivity of the technology and the distribution of missing edges (in this case only one) is sufficiently random in the subgraph for `Complex2`. Figure 5 contains the subgraphs for the new complex estimates and Figure 6 shows the corresponding bipartite PCMG.

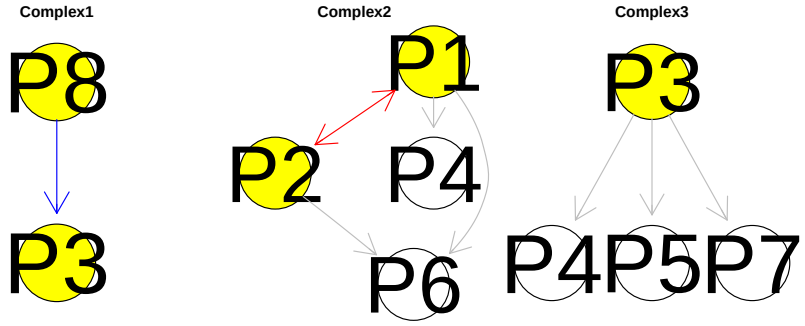

Figure 5: Subgraphs for new complex estimates after using `mergeComplexes`.

The function `findComplexes` can be used to run both steps together.

```
> PCMG2 <- findComplexes(apEX,sensitivity=.7,specificity=.75)

[1] "Finding Initial Maximal BH-complete Subgraphs"
[1] "Combining Complex Estimates"
[1] "calculating initial penalty terms"
[1] "looking at complex combinations"

> PCMG2

$Complex1
[1] "P2" "P1" "P6"
```

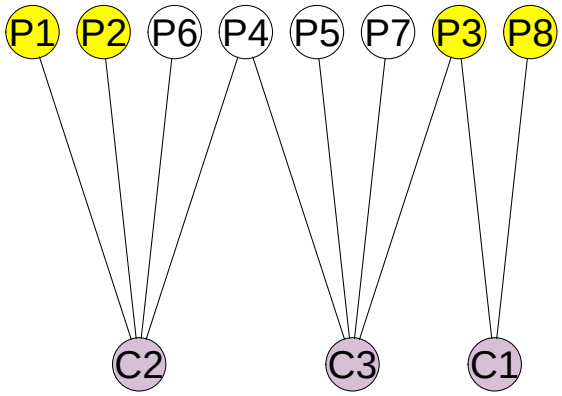

Figure 6: Bipartite graph for new complex estimates after using `mergeComplexes`.

```
$Complex2
[1] "P8" "P3"

$Complex3
[1] "P4" "P3" "P5" "P7"

$Complex4
[1] "P4" "P1" "P6"

>
```

Our algorithm makes three types of complex estimates: multi-bait-mult-edge (MBME) complexes that contain multiple baits and multiple edges, single-bait-multi-hit (SBMH) complexes that contain a single bait and a collection of hit-only proteins, and unreciprocated bait-bait (UnRBB) complexes that only contain two bait proteins connected by one unreciprocated edge. MBME complexes are the most reliable outputs since they contain the most tested data. SBMH complexes are useful for proposing future experiments since the topology among the hit-only proteins is unknown. UnRBB complexes may result from FP observations since the edges are tested twice, observed once, and not confirmed by other subgraph edges. On the other hand, the unreciprocated edge may also result from a FN observation between the two baits. The PCMG affiliation matrix resulting from `mergeComplexes` or `findComplexes` can be sorted into the MBME, SBMH, and UnRBB components using the function `sortComplexes`.

```
> sortComplexes(PCMG2,adjMat=apEX)
```

```
$MBME
$MBME$MBME1
[1] "P2" "P1" "P6"


$SBMH
$SBMH$SBMH1
[1] "P4" "P3" "P5" "P7"

$SBMH$SBMH2
[1] "P4" "P1" "P6"


$UnRBB
$UnRBB$UnRBB1
[1] "P8" "P3"
```

Recall that the true complexes in this example consist of P_1, P_2, P_4, and P_6 and P_3, P_4, and P_5. `MBME1` accurately estimates P_1, P_2, P_4, and P_6 as composing one complex. `SBMH1` predicts P_3, P_4, and P_5 as members of one complex, but also includes the FP observation for P_7. The limited data for this complex makes it impossible to distinguish the FP from the true positive (TP) observations. Further purifications using P_4, P_5, and P_7 as baits would likely resolve this difficulty. `UnRBB1` is the result of the FP observation between P_8 and P_3. While it is reported as a complex estimate, further experimental testing would likely confirm that this is not a true complex.

If the user desires, a matrix of similarity data can be included as the `simMat` argument in `mergeComplexes` and `findComplexes` in an extended logistic regression model. The similarity measure can be used to lend credence to the existence of an edge, even if it is not detected using AP-MS. Users must specify the parameter `Beta` which weights contribution of the similarity measure to the model. See Scholtens and Gentleman (2004) for details.

Publicly available data

Two publicly available data sets are included in *apComplex*. `TAP` is an adjacency matrix of the AP-MS data (called 'TAP') reported by Gavin, et al. (2002). There were 3420 comemberships reported using 455 baits and 909 hit-only proteins. `TAPgraph` contains a graph of class `graphNEL` of the TAP data. The TAP data were originally compiled into 232 yTAP complexes, available in Supplementary Table 1 of Gavin, et al. (2002) at http://www.nature.com and at http://yeast.cellzome.com. These yTAP complex estimates, along with the annotations given by Gavin, et al. are available in `yTAP`.

```
> data(TAP)
> dim(TAP)

[1]  455 1364
```

```
> data(TAPgraph)
> which(TAP["Abd1",]==1)

Rpb2 Spt5
 926 1049

> adj(TAPgraph,"Abd1")

$Abd1
[1] "Rpb2" "Spt5"

> data(yTAP)
>
```

HMSPCI is an adjacency matrix of the AP-MS data (called 'HMS-PCI') reported by Ho, et al. (2002). There were 3687 comemberships reported using 493 baits and 1085 hit-only proteins. HMSPCIgraph contains a graph of class graphNEL of the HMS-PCI data.

```
> data(HMSPCI)
> dim(HMSPCI)

[1]  493 1578

> data(HMSPCIgraph)
> which(HMSPCI["YAL015C",]==1)

YDL029W YJR068W YMR146C YPR110C YDR214W YEL030W YEL060C YMR012W YMR058W YNL037C
     62     280     367     488     704     759     765    1274    1279    1333

> adj(HMSPCIgraph,"YAL015C")

$YAL015C
 [1] "YDL029W" "YJR068W" "YMR146C" "YPR110C" "YDR214W" "YEL030W" "YEL060C"
 [8] "YMR012W" "YMR058W" "YNL037C"
```

These data were analyzed using *apComplex*, and the results are described in Scholtens, Vidal, and Gentleman (submitted). The complex estimates are available for both data sets - MBMEcTAP, SBMHcTAP, and UnRBBcTAP for the TAP data, and MBMEcHMSPCI, SBMHcHMSPCI, and UnRBBcHMSPCI for the HMS-PCI data.

One example of the detail with which the *apComplex* algorithm can estimate complex membership involves the PP2A proteins Tpd3, Cdc55, Rts1, Pph21, and Pph22. These five proteins compose four heterotrimers (Jiang and Broach, 1999). We accurately predict these trimers as distinct complexes. and furthermore note the exclusive association of Zds1 and Zds2 with the Cdc55/Pph22 trimer. Confirmation of this prediction in the lab may help clarify the cellular function of this particular trimer and the reason for its joint activity with Zds1 and Zds2.

```
> data(MBMEcTAP)
> which(MBMEcTAP[,37]==1)

Cdc55 Pph22  Tpd3  Zds2
   30   145   253   301

> which(MBMEcTAP[,38]==1)

Cdc55 Pph22  Tpd3  Zds1
   30   145   253   300

> which(MBMEcTAP[,39]==1)

Cdc55  Tpd3 Pph21
   30   253   454

> which(MBMEcTAP[,195]==1)

Pph22  Rts1  Tpd3
  145   197   253

> which(MBMEcTAP[,233]==1)

 Rts1  Tpd3 Pph21
  197   253   454

>
```

In summary, *apComplex* can be used to predict complex membership using data from AP-MS experiments. An accurate catalog of complex membership is a fundamental requirement for understanding functional modules in the cell. Integration of *apComplex* analyses with other high-throughput data, including binary physical interactions assayed by yeast two-hybrid technology, gene expression data, and binding domain data are promising avenues for further systems biology research.

References

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

Ho, Y., et al. (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry, *Nature*. **415**, 180-183.

Jiang, Y., Broach, J. (1999) Tor proteins and protein phosphatase 2A reciprocally regulate Tap42 in controlling cell growth in yeast, *EMBO Journal*, **18**, 2782-2792.

Scholtens, D., Gentleman, R. (2004) Making sense of high-throughput protein-protein interaction data, *Statistical Aplications in Genetics and Molecular Biology*, **3**, Article 39.

Scholtens, D., Vidal, M., Gentleman, R. Local modeling of global interacome networks, *Submitted*.