same for all FCM applications.

Currently, in a collaboration of several groups involved in high-throughput FCM together with instrument manufacturers and members of the flow cytometry standards initiative a **flowCore** package and a number of additional FCM utility packages are developed. The aim is to merge both **prada** and **rflowcyt** into one core package which is copmpliant with the data exchange standards that are currently developed in the community. Visualization as well as quality control will than be part of the utility packages that depend on the data structures defined in the **flowCore** package.

Bibliography

L. C. Seamer, C. B. Bagwell, L. Barden *et al.* Proposed new data file standard for flow cytometry, version fcs 3.0. *Cytometry*, 28(2):118–122, Jun 1997.

Nolwenn Le Meur
Computational Biology
Fred Hutchinson Cancer Research Center
Seattle, WA, USA
nlemeur@fhcrc.org

Florian Hahne
Molecular Genome Analysis
German Cancer Research Center
Heidelberg, Germany
f.hahne@dkfz.de

Protein Complex Membership Estimation using apComplex

by Denise Scholtens

Graphs of protein-protein interactions, so called 'interactomes', are rapidly surfacing in the systems biology literature. In these graphs, nodes represent cellular proteins and edges represent interactions between them. Global interactome analyses are often undertaken to explore topological features such as network diameter, clustering coefficients, and node degree distribution. Local interactome modeling, particularly at the protein complex level, is also important for identifying distinct functional components of the cell and studying their interactivity (Hartwell et al., 1999). The **apComplex** package contains functions to locally estimate protein complex membership as described in Scholtens and Gentleman (2004) and Scholtens et al. (2005).

Two technologies are generally used to query protein-protein relationships. Affinity purification-mass spectrometry (AP-MS) technologies detect protein complex co-membership. In these experiments a set of proteins are used as baits, and in separate purifications, each bait identifies all hits with which it shares protein complex membership. AP-MS baits and their hits may physically bind to each other, or they may be joined together in a complex through an intermediary protein or set of proteins. If a bait protein is a member of more than one complex, all of its hits may not necessarily themselves be complex co-members. These biological realities become essential components of complex membership estimation. Publicly available AP-MS data sets for *Saccharomyces cerevisiae* include those reported by Gavin et al. (2002, 2006), Ho et al. (2002), and Krogan et al. (2004, 2006).

Yeast-two-hybrid (Y2H) technology is another bait-hit system that measures direct physical interactions. The distinction between AP-MS and Y2H data is subtle, but crucial. Two proteins that are part of the same complex may not physically interact with each other. Thus an interaction detected by AP-MS may not be detected by Y2H. On the other hand, two proteins that do physically interact by definition form a complex so any interaction detected by Y2H should also be detected by AP-MS. Under the same experimental conditions, Y2H technology should in fact consist of a subset of the interactions detected by AP-MS technology, the subset consisting of complex comembers that are physically bound to each other. Ito et al. (2001) and Uetz et al. (2000) both offer publicly available Y2H data sets for *Saccharomyces cerevisiae*.

apComplex deals strictly with data resulting from AP-MS experiments. The joint analysis of Y2H and AP-MS data is an interesting and important problem and is in fact an obvious next step after complex membership estimation, but is not currently dealt with in **apComplex**.

Protein Complex Membership and *Co*-Membership

apComplex estimates protein complex membership given a set of AP-MS co-membership data. The distinction between the complex membership and co-membership ties back to affiliation relationships in social networks analyses (Wasserman and Faust, 1999). As a simple example to be discussed throughout this article, suppose proteins P_1, P_2, P_4, and P_6 compose complex C_1 and proteins P_3, P_4, and P_5 compose complex C_2. Then their affiliation matrix, A, is as follows.

$$A = \begin{array}{c|cc} & C_1 & C_2 \\ \hline P_1 & 1 & 0 \\ P_2 & 1 & 0 \\ P_3 & 0 & 1 \\ P_4 & 1 & 1 \\ P_5 & 0 & 1 \\ P_6 & 1 & 0 \end{array}$$

A can also be represented as a bipartite graph in which one set of nodes represent proteins, another set represents complexes, and edges from proteins to complexes denote complex membership.

Instead of A, AP-MS technology assays $Y = A \otimes A'$ where $\otimes$ represents matrix multiplication under the Boolean algebra $0 + 0 = 0 \times 0 = 1 \times 0 = 0 \times 1 = 0$ and $1 + 0 = 0 + 1 = 1 + 1 = 1 \times 1 = 1$. Entries of 1 in Y represent co-membership of two proteins in a complex. In this simple example, we have Y as follows.

$$Y = \begin{array}{c|cccccc} & P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \\ \hline P_1 & 1 & 1 & 0 & 1 & 0 & 1 \\ P_2 & 1 & 1 & 0 & 1 & 0 & 1 \\ P_3 & 0 & 0 & 1 & 1 & 1 & 0 \\ P_4 & 1 & 1 & 1 & 1 & 1 & 1 \\ P_5 & 0 & 0 & 1 & 1 & 1 & 0 \\ P_6 & 1 & 1 & 0 & 1 & 0 & 1 \end{array}$$

apComplex estimates A using assays of Y.

Observed Data

The entire matrix Y is not tested in AP-MS experiments. For this to happen, all cellular proteins would need to be used as baits. Even in small model organisms such as *Saccharomyces cerevisiae* with approximately 6000 cellular proteins, genome-wide testing is logistically prohibitive. Instead, only a subset of the rows of Y are tested (letting rows represent baits and columns represent hits). In a graph of Y, this is best described by neighborhood sampling of all edges extending from baits to all other proteins. Recognition of this sampling scheme is crucial as it draws a very important distinction between two types of edges that are absent from a graph of AP-MS data. Figure 1 shows ideal results from a neighborhood sampling of our simple interactome using P_1, P_2, and P_3 as baits. The edge between P_1 and P_3 is tested and observed to be absent, but the edge between P_4 and P_6 is absent because it is never tested. Scholtens and Gentleman (2004) discuss why inference should only be based on the tested edges, leaving the untested edges to be estimated or tested in further experiments.

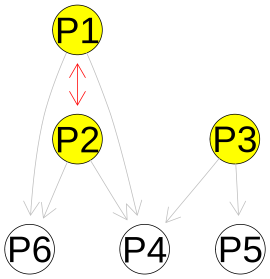

Figure 1: Co-memberships detected using neighborhood sampling scheme with P_1, P_2, and P_3 as baits.

In addition to being incomplete AP-MS data are also imperfect, including both false positive (FP) and false negative (FN) observations. Figure 2 demonstrates hypothetical FP observations from P_8 to P_3 and P_3 to P_7 and a FN observation from P_2 to P_4.

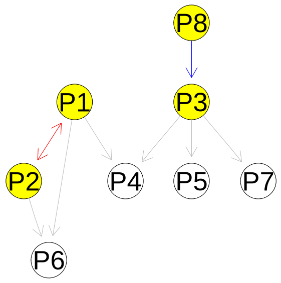

Figure 2: Observed data including FPs and FNs.

Penalized Likelihood Approach

Since 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 estimate A (see Scholtens and Gentleman, 2004 and Scholtens et al., 2005 for a more thorough discussion).

Since all proteins are not used as baits, **apComplex** instead searches for *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 other words, all edges observed in the neighborhood sampling scheme must exist for a subgraph to be *BH-complete*. The function bhmaxSubgraph can be used to find maximal BH-complete subgraphs in the observed AP-MS data.

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).

For our example, bhmaxSubgraph reports four maximal BH-complete subgraphs as shown below. Figure 3 depicts the bipartite graph of the results contained in BP1.

```
> 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
> BP1 <- bhmaxSubgraph(apEX)
   bhmax1 bhmax2 bhmax3 bhmax4
P1      0      1      1      0
P2      0      1      0      0
P3      1      0      0      1
P8      1      0      0      0
P4      0      0      1      1
P5      0      0      0      1
P6      0      1      1      0
P7      0      0      0      1
```

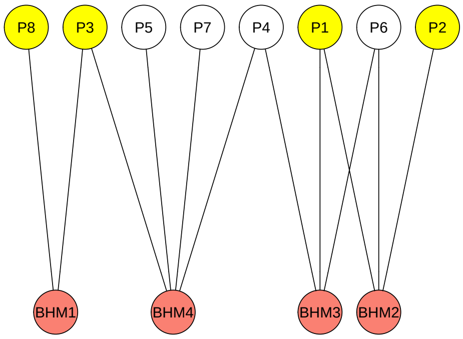

Figure 3: Bipartite graph for the initial estimate of A determined by locating maximal BH-complete subgraphs in the graph of observed AP -MS data.

The initial maximal BH-complete subgraph estimate of A 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. mergeComplexes accommodates this by employing a penalty term with the likelihood. 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 the supposed sensitivity of the technology, 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 penalized likelihood 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 mergeComplexes looks to increase C in favor of small decreases in L.

After the initial estimate of A 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. This is a greedy algorithm and mergeComplexes can be sensitive to the order in which the columns in the input matrix are specified. Users may want to order the columns putting the initial complex estimates with more bait proteins first since these contain proportionately more tested data.

```
> BP2 <- mergeComplexes(BP1, apEX,
sensitivity = 0.7, specificity = 0.75)
> BP2
   Complex1 Complex2 Complex3
P1        0        1        0
P2        0        1        0
P3        1        0        1
P8        1        0        0
P4        0        1        1
P5        0        0        1
P6        0        1        0
P7        0        0        1
```

Figure 4 shows the corresponding bipartite graph for the estimated A.

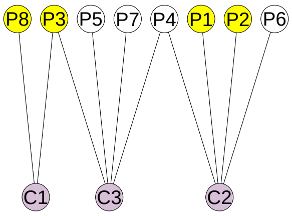

Figure 4: Bipartite graph for new complex estimates after using mergeComplexes.

The function findComplexes can be used to run both steps together. It will automatically reorder the input to mergeComplexes as suggested previously. Note that the user must specify the sensitivity and specificity of the AP-MS technology. Specificity should be considered in light of the dimension of the data under consideration. In our small example, a fairly high FP rate (i.e. low specificity) creates a reasonable number of suspected FP interactions. In large dimensional data sets, the number of true negative interactions is quite large. In these cases, very low FP probabilities (e.g. 0.001 or specificity=0.999) are usually appropriate.

Algorithm Output

apComplex 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 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 complex estimates resulting from mergeComplexes or findComplexes can be sorted into the MBME, SBMH, and UnRBB components using the function sortComplexes.

Results for three 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. 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.

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.

Krogan is an adjacency matrix of the AP-MS data reported by Krogan et al. (2004). There were 1132 comemberships reported using 153 baits and 332 hit-only proteins.

These data were analyzed using **apComplex**, and the results for the TAP and HMS-PCI data sets are described in Scholtens et al. (2005). Complex estimates are available for all three data sets - MBMEcTAP, SBMHcTAP, and UnRBBcTAP for the TAP data, MBMEcHMSPCI, SBMHcHMSPCI, and UnRBBcHMSPCI for the HMS-PCI data, and MBMEcKrogan for the Krogan 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). Using the TAP data, **apComplex** accurately predicts these trimers as distinct complexes and furthermore notes 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.

Related Packages

Several other packages based on the **apComplex** algorithm are currently being developed. The **ScISI** package contains an *in silico* interactome including **apComplex** estimates of publicly available AP-MS data. Given an interactome, **simulatorAPMS** can be used to simulate the neighborhood sampling scheme and both stochastic and systematic errors characteristic of AP-MS experiments for testing the performance of complex estimation algorithms, among other things. To complement the AP-MS data analysis, **y2hStat** contains algorithms for Y2H data analysis. This effort to better model Y2H observations will facilitate improved joint modeling of **apComplex** outputs and Y2H data.

Summary

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 Y2H physical interactions, gene expression data, and binding domain data are promising avenues for further systems biology research.

Bibliography

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

A. C. Gavin *et al.* Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 440:631–636, 2006.

L. Hartwell, J. Hopfield, S. Leibler *et al.* From molecular to modular cell biology. *Nature*, 402:C47, 1999.

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

T. Ito, T. Chiba, R. Ozawa, *et al.* A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Nat. Acad. Sci. U.S.A.*, 98:4569–4574, 2001.

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

N. Krogan *et al.* High-definition macromolecular composition of yeast RNA-processing complexes. *Molecular Cell*, 13(2):225–239, 2004.

N. Krogan *et al.* Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature*, 440:637–643, 2006.

D. Scholtens and R. Gentleman. Making sense of high-throughput protein-protein interaction data. *Statistical Applications in Genetics and Molecular Biology*, 3(1):Article 39, 2004.

D. Scholtens, M. Vidal, and R. Gentleman. Local modeling of global interactome networks. *Bioinformatics*, 21:3548–3557, 2005.

P. Uetz, L. Giot, G. Cagney, *et al.* A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature*, 403:623–627, 2000.

S. Wasserman and K. Faust. *Social Network Analysis*. Cambridge University Press, New York, 1999.

Denise Scholtens
Northwestern University Medical School
Chicago, IL, USA
dscholtens@northwestern.edu

SNP Metadata Access and Use with Bioconductor

by Vince Carey

Introduction

“Single nucleotide polymorphisms (or SNPs) ... are DNA sequence variations that occur when a single nucleotide in genomic sequence is altered”[1]. Conventionally, a given variation must be present in at least one percent of the population in order for the variant to be regarded as a SNP.

There are many uses of data on SNPs in bioinformatics. Two recent contributions which lay out aspects of the concept of “genetical genomics” are Li and Burmeister (2005) and Cheung et al. (2005). In this short contribution I review some functionality provided by Bioconductor for investigating analyses related to the Cheung *et al.* paper.

The *RSNPper* package

The SNPper[2] web service of the Children’s Hospital (Boston) Informatics Program provides interactive access to a curated database of metadata on SNPs. Details of the system are provided in Riva and Kohane (2005). In addition to the browser-based interface, SNPper has an XML-RPC query resolution system. The *RSNPper* package provides an interface to this XML-RPC-based service. The objective of *RSNPper* is to provide a convenient high-level interface to the SNPper database contents, by providing a small number of high-level query functions with simple calling sequence, and by translating XML responses to convenient R-language objects for further use.

Getting gene-level information

A `geneInfo` function takes a string argument with a HUGO gene symbol and returns an object of class `SNPperGeneMeta`:

```
> cpm = geneInfo("CPNE1")
> cpm
SNPper Gene metadata:
There are  8 entries.
Basic information:
  GENEID  NAME CHROM STRAND  PRODUCT NSNPS
1  12431 CPNE1 chr20      - copine I   160
  TX.START   TX.END CODSEQ.START CODSEQ.END
1 33677382 33705245     33677577   33684259
  LOCUSLINK   OMIM   UNIGENE SWISSPROT
1      8904 604205 Hs.166887    Q9NTZ6
    MRNAACC   PROTACC REFSEQACC
1 NM_003915 NP_003906      NULL
SNPper info:
     SOURCE           VERSION
[1,] "*RPCSERV-NAME*" "$Revision: 1.38 $"
```

[1] http://www.ornl.gov/sci/techresources/Human_Genome/faq/snps.shtml
[2] snpper.chip.org