Quality scoring of protein-protein interaction data

Introduction to protein-protein interaction data

The proteins of a cell does not necessarily work as individual units. They are often found in functional modules made up by two or more interacting proteins. Protein-protein interactions can be measured experimentally in large-scale studies either as binary interactions, e.g. using yeast two-hybrid [1-3], PCA [4] or as multiple interactions in protein complexes using mass spectrometry [5-8].

Unfortunately, the error rates in the first large-scale studies were high, estimated to be as high as 60% for yeast two hybrid and 50% in the protein complex pull downs [9]. Thus, there was (and still is) a need for determining the reliability of each interaction to enable scientist to evaluate the large-scale data sets.

Scoring interactions from binary interaction methods (e.g. Y2H, PCA)

For the two different types of high-throughput data sets, scoring schemes have been developed that allow the reliability of individual, binary interactions to be compared across data sets. For the yeast two-hybrid experiments, the reliability of an interaction has been found to correlate well with the number of non-shared interaction partners for each interacting pair [10]. This can be summarized in the following raw quality score:

$$S(A,B)_{bin} = -\log_{10}\left((N_A+1)(N_B+1)\right)$$

where N_A and N_B are the numbers of non-shared interaction partners for an interaction between protein A and B, see Figure 1.

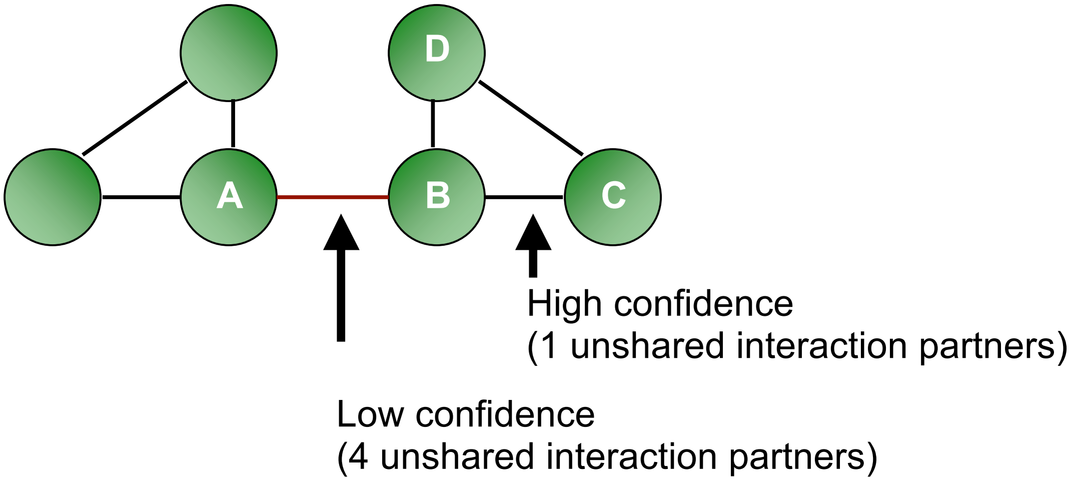

D
A
B
C
High confidence
(1 unshared interaction partners)
Low confidence
(4 unshared interaction partners)

Figure 1: The reliability of a binary interaction has been found to correlate with the number of non-shared interaction partners.

Scoring interactions inferred from MS methods (e.g. TAP, APMS, HMS-PCI)

In the case of complex pull-down experiments, the reliability of the inferred binary interactions has been found to correlate better with the number of times the proteins were co-purified vs. the number of pull-downs they are identified in. The following pull-down score is an adapted version of 'S2' found in the Supplemental methods of de Lichtenberg *et al.* 2005 [10].

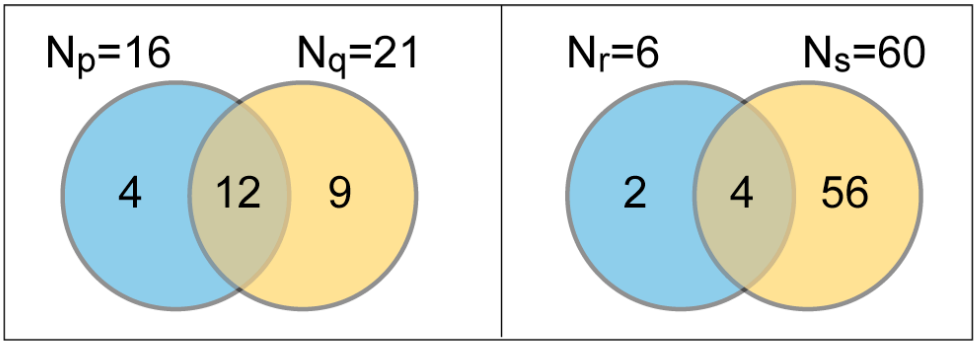
$$S(A,B)_{pul} = \log_{10}\left(\frac{(N_{A\cap B})(N_{A\cup B})}{(N_A+1)(N_B+1)}\right)$$

where:

- $N_{A\cap B}$ is the number of purifications containing both proteins, i.e. the intersection of experiment sets that find them
- $N_{A\cup B}$ is the total number of purifications that find either A or B, i.e. the union of experiments that find them
- N_A is the number of purifications containing A
- N_B is the number of purifications containing B

Note that $N_{A\cap B} \geq 1$, $N_{A\cup B} \geq 1$, $N_A \geq 1$ and $N_B \geq 1$ as we only consider protein pairs that have been detected at least once.

Examples:



The Venn diagrams show two examples for a medium-scale APMS study where 350 pull downs were performed and included the proteins p, q, r and s. The numbers in the diagram represent the number of experiments that purified either or both proteins, e.g. N_p, N_q, $N_{p\cap q}$, and $N_{p\cup q}$. The pull-down score for the p-q and r-s interactions would be calculated as shown,

$$S(p,q)_{pul} = \log_{10}\left(\frac{(12)(25)}{(16+1)(21+1)}\right) = -0.096$$

$$S(r,s)_{pul} = \log_{10}\left(\frac{(4)(62)}{(6+1)(60+1)}\right) = -0.236$$

The pull-down scores indicate that we have more confidence in the interaction between p and q than between r and s. This is primarily because protein s is pulled-down somewhat non-specifically in a large fraction if experiments.

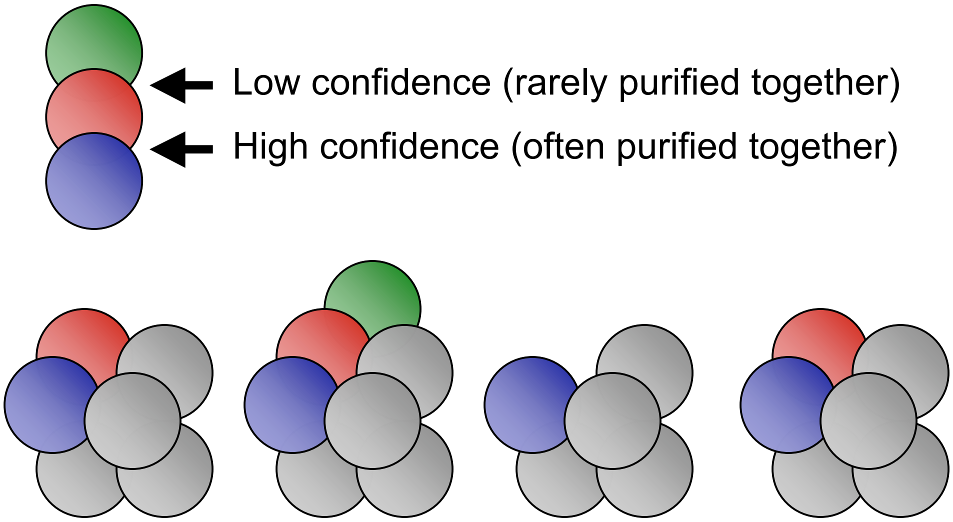

Low confidence (rarely purified together)
High confidence (often purified together)

Figure 2: For pull-down experiments, interactions between proteins that often co-purify together are more reliable than those that rarely co-purify together.

More complicated interaction scoring approaches have been used. For example, a socio-affinity measure that respects aspects of the experimental design was proposed by Gavin *et al.* 2006 [7].

References:

1. Uetz, P., et al., *A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae.* Nature, 2000. **403**(6770): p. 623-7.
2. Ito, T., et al., *A comprehensive two-hybrid analysis to explore the yeast protein interactome.* Proc Natl Acad Sci U S A, 2001. **98**(8): p. 4569-74.
3. Yu, H., et al., *High-quality binary protein interaction map of the yeast interactome network.* Science, 2008. **322**(5898): p. 104-10.

4. Tarassov, K., et al., *An in vivo map of the yeast protein interactome.* Science, 2008. **320**(5882): p. 1465-70.
5. Gavin, A.C., et al., *Functional organization of the yeast proteome by systematic analysis of protein complexes.* Nature, 2002. **415**(6868): p. 141-7.
6. Ho, Y., et al., *Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry.* Nature, 2002. **415**(6868): p. 180-3.
7. Gavin, A.C., et al., *Proteome survey reveals modularity of the yeast cell machinery.* Nature, 2006. **440**(7084): p. 631-6.
8. Krogan, N.J., et al., *Global landscape of protein complexes in the yeast Saccharomyces cerevisiae.* Nature, 2006. **440**(7084): p. 637-43.
9. von Mering, C., et al., *Comparative assessment of large-scale data sets of protein-protein interactions.* Nature, 2002. **417**(6887): p. 399-403.
10. de Lichtenberg, U., et al., *Dynamic complex formation during the yeast cell cycle.* Science, 2005. **307**(5710): p. 724-7.