Supplementary Methods

Enrichment Analysis

Firstly, some definitions: in general, we are interested in protein interaction networks with N proteins (nodes) in the network and use indices $i = 1, 2, \ldots, N$ to represent each node. Then, $e_{ij} = e_{ji} = 1$ if nodes i and j are connected by an edge and $e_{ij} = e_{ji} = 0$ if nodes i and j are not connected. Furthermore, each node is annotated as belonging to one of K mutually exclusive categories (in our case, subcellular location), and we use indices $\alpha = 1, 2, \ldots, K$ to represent each category. If protein i is in category α, $c_{i\alpha} = 1$. If not, $c_{i\alpha} = 0$. Finally, each node has a degree k_i, which is the number of other nodes to which it is connected by an edge.

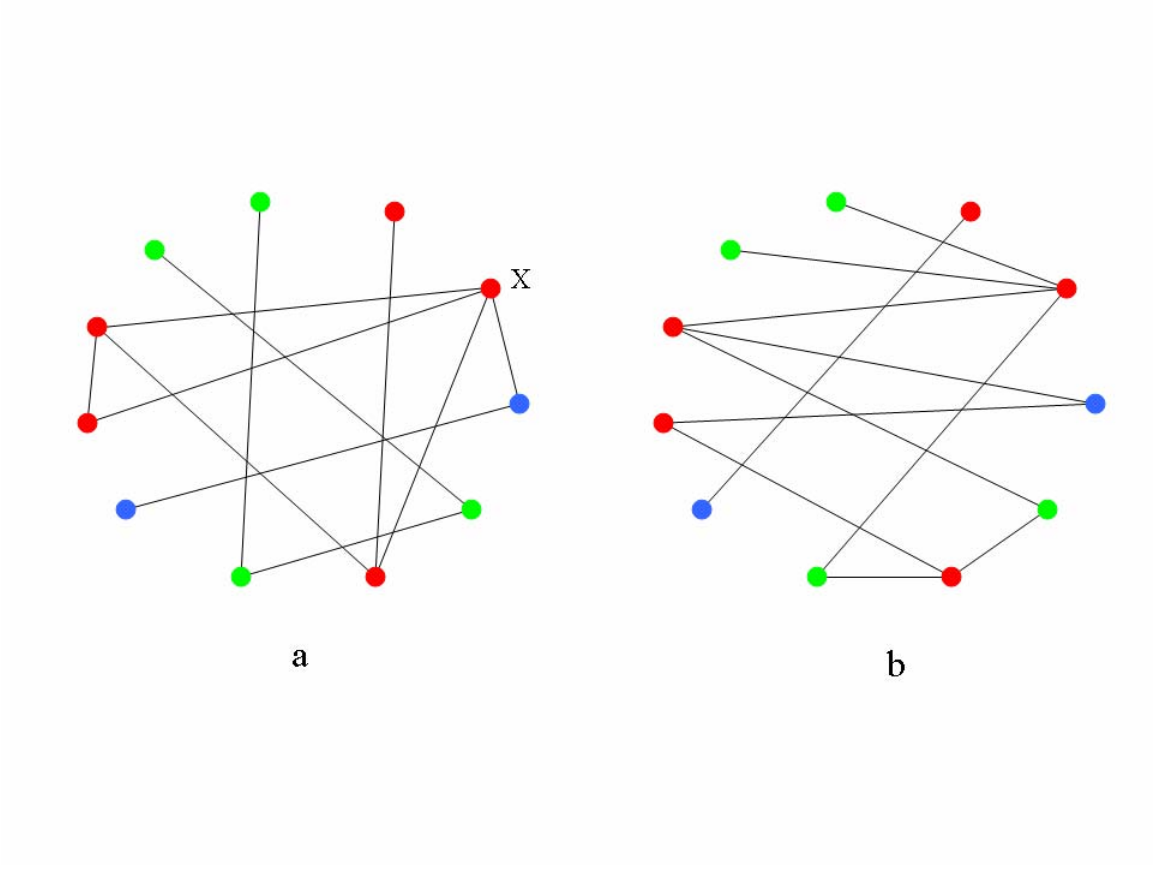

Figure SM1. a) "enriched" network. b) random network.

Consider the small network in Figure 1a, with $N = 11$ nodes. Colors are used to represent the annotation of each node to one of $K = 3$ categories. Upon inspection the figure gives the impression that nodes with the same color are more connected than would be expected at random (as we expect for proteins within the same subcellular compartment) and the goal of the present analysis is to test for this quality rigorously. For each pair of categories α and β, the number of edges between nodes with categories α and β is denoted by $n_{\alpha\beta}$ (*obs*). For example, in Figure 1a, the number of edges between red nodes is n_{redred} (obs) = 6. We would like to compare this number with the expected number of edges between proteins with categories α and β in an ensemble of random networks. But what type of random network? The classic, exactly solvable Erdös-Rényi random networks (Erdös and Rényi, 1959) connect nodes i and j in a network of N nodes with a uniform probability p. If such a random network were constructed for the nodes in Figure 1a, it would be possible that the highly connected node of degree 4, marked X, is not connected to any other node. A more suitable set of random networks for the purposes of the present analysis are those networks for which:

a) the degree k of each node is preserved
b) the category α of each node is preserved
c) the total number of nodes is preserved
d) the total number of edges is preserved

A random network fulfilling these criteria is illustrated in Figure 1b. In this case, $n_{redred} = 2$. These networks are a special case of random graphs with specified degree distributions, which have recently been used to investigate properties of the World Wide Web and social interaction networks. (e.g. Newman et al. (2002), or Newman (2003) for an excellent review).

The number of edges between proteins with categories α and β is given by

$$n_{\alpha\beta}(obs) = \sum_{j} \sum_{i<j} (c_{i\alpha} c_{j\beta} \; OR \; c_{i\beta} c_{j\alpha}) e_{ij}$$

where the term "OR" indicates that if both proteins are in the same compartment the term within the parentheses is 1.

Let us denote the total number of edges in a network as E. Now, in the random networks described above, provided that k_i and $k_j << 2E$, the probability that nodes i and j are connected by an edge is

$$\bar{e}_{ij} = \frac{k_i k_j}{2E + k_i k_j}$$

(Bader, J., Personal communication; note that the condition above is *not* fulfilled by the illustrative example in the figure, but is true for the networks of the present study.) Therefore, for the ensemble of these random networks, the mean value of $n_{\alpha\beta}$ is, from the first two equations

$$\bar{n}_{\alpha\beta} = \sum_j \sum_{i<j} \frac{\left(c_{i\alpha} c_{j\beta} \; OR \; c_{i\beta} c_{j\alpha}\right) k_i k_j}{\left(2E + k_i k_j\right)}$$

For each pair of categories α and β, enrichment is present in the observed network for $n_{\alpha\beta}$ (obs) $\geq \bar{n}_{\alpha\beta}$, and depletion for $n_{\alpha\beta}$ (obs) $< \bar{n}_{\alpha\beta}$.

As described by Gandhi et al. (2006) a Poisson distribution is then suitable to calculate statistical significance:

$$P(n_{\alpha\beta}) = \begin{cases} \sum_{j=0}^{n_{ab}} \bar{n}_{\alpha\beta}^{\,j} \exp(-\bar{n}_{\alpha\beta}) / j!, & n_{\alpha\beta} < \bar{n}_{\alpha\beta} \; (depletion) \\ \sum_{j=n_{ab}}^{\infty} \bar{n}_{\alpha\beta}^{\,j} \exp(-\bar{n}_{\alpha\beta}) / j!, & n_{\alpha\beta} \geq \bar{n}_{\alpha\beta} \; (enrichment) \end{cases}$$

Finally, a conservative Bonferroni multiple testing correction is applied, as $P(\text{multi}) = 1 - (1 - P)^m$, where P is the single test P value and m is the number of tests. For enrichment, m is *a priori* the number of $\alpha\beta$ pairs for which $n_{\alpha\beta}$ (obs) > 0. Similarly, for depletion, m is the number of all $\alpha\beta$ pairs for which $\bar{n}_{\alpha\beta} > 0$ (which is the total number of $\alpha\beta$ pairs).

References

Erdös P and Rényi A (1959). On random graphs. Publ. Math. **6**: 290-297.

Newman ME, Watts DJ and Strogatz SH (2002). Random graph models of social networks. Proc Natl Acad Sci. USA. **99**, Suppl 1: 2566-2572.

Newman MEJ (2003). The structure and function of complex networks. SIAM Rev. **45**, 167-256.