Pacific Symposium on Biocomputing 14:391-402 (2009)

HIGH THROUGHPUT INTERACTION DATA REVEALS DEGREE CONSERVATION OF HUB PROTEINS

A. FOX*[,1], D. TAYLOR[1], and D. K. SLONIM[1,2]

[1] *Department of Computer Science, Tufts University, 161 College Avenue, Medford, MA 02155, USA*
[2] *Department of Pathology, Tufts University School of Medicine, 145 Harrison Ave., Boston, MA 02111, USA*
* *E-mail: afox01@cs.tufts.edu*

Research in model organisms relies on unspoken assumptions about the conservation of protein-protein interactions across species, yet several analyses suggest such conservation is limited. Fortunately, for many purposes the crucial issue is not global conservation of interactions, but preferential conservation of functionally important ones. An observed bias towards essentiality in highly-connected proteins implies the functional importance of such "hubs". We therefore define the notion of *degree-conservation* and demonstrate that hubs are preferentially degree-conserved. We show that a protein is more likely to be a hub if it has a high-degree ortholog, and that once a protein becomes a hub, it tends to remain so. We also identify a positive correlation between the average degree of a protein and the conservation of its interaction partners, and we find that the conservation of individual hub interactions is surprisingly high. Our work has important implications for prediction of protein function, computational inference of PPIs, and interpretation of data from model organisms.

1. Introduction

The power of comparative genomic research has grown steadily with the availability of genomic sequence and annotation for increasing numbers of organisms. A variety of techniques for solving such diverse problems as motif discovery,[1] gene expression analysis,[2] regulatory network inference[3] and interactome discovery[4] rely on the assumption that functionally important protein-DNA and protein-protein interactions (PPIs) are preferentially conserved across species. Yet much remains to be done to fully understand the conservation of protein interaction modules and functions.[5] Although studies of 'interologs' (orthologous pairs of proteins whose interactions are conserved)[6] report preferential conservation of interactions between highly

Pacific Symposium on Biocomputing 14:391-402 (2009)

similar orthologous protein pairs,[7] Mika and Rost[8] point out that inference of PPIs based on sequence homology across species continues to be quite inaccurate, even at very high levels of sequence identity.

For the interpretation of data from model organisms, such low interolog conservation might still be acceptable, provided that the most *functionally important* interactions for which orthologs exist in both species are conserved at an acceptably high rate. Yet even this remains to be shown. Recent work[9] shows that functional modules are no better conserved between yeast and fly than would be expected by chance. Furthermore, previous work in studying transcriptional regulation shows that despite the preferential sequence conservation of functionally important proteins, functional regulatory modules are not especially well conserved.[10]

Here, we focus not on the conservation of interactions among all proteins, but on the most highly-connected proteins, the "hubs" of a species' interaction network. We select these hub proteins precisely because their high connectivity may serve as an indication of functional importance. Indeed, proteins with many interactions are more likely to be essential,[11] and among these, hubs connected to other hubs may be more essential still.[12] If hub proteins are more likely functionally important, perhaps their roles as hubs are preferentially conserved, even when many of their individual interactions are not. We therefore investigate whether the property of being a hub protein is conserved across organisms, and we determine the level of PPI conservation between hubs and their neighbors. We show that interactions with high degree proteins are preferentially conserved and that even when specific interactions with hubs are lost or gained, the *high degree* of a hub protein is nonetheless conserved. Thus, there is greater hope that the the functional importance of these hubs is conserved as well.

We point out that there is ample reason to expect *a priori* that the high degree of hub proteins would be preferentially conserved during evolution.[13] This is based on the argument that once a protein has evolved functional interactions with many other proteins, any dramatic change would affect all of its many partners, and thus is likely to be evolutionarily disadvantageous. However, given that recent work[9,10] casts doubt upon modular functional conservation, this hypothesis requires proof.

There are also reasons to be skeptical of the quality and quantity of the available PPI data. Estimates of the false-positive rates for high throughput interaction data sets range from 45% to 90%,[14,15] and coverage estimates for the human and yeast interactomes are around 10% and 50%, respectively.[14,16]

Thus, any attempts to identify high degree proteins will be affected by noise, coverage, and testing bias.[16] Methods for estimating the true degree of interactions[17] are available, but are not directly applicable to the data sets studied here, because in many cases it is not clear which interactions were tested. Nonetheless, with only modest data filtering, we obtain clear evidence here of several important trends.

Our results provide additional evidence that identifying biologically relevant high-degree proteins in model organisms should shed light on human response. Our work also has important implications for protein function prediction, comparative genomics, and network inference and analysis.

2. Background and Definitions

2.1. *Data Sources*

All PPIs in *Homo sapiens*, *Drosophila melanogaster*, *Caenorhabditis elegans*, and *Saccharomyces cerevisiae* were downloaded from the BioGRID,[18] IntAct[19] and MINT[20] online databases in July, 2008 and combined to form a single large data set.

To map orthologous proteins between species, we used the InParanoid database[21] because it discriminates true orthologs and 'in' paralogs that differentiated after the relevant speciation event from 'out' paralogs that arose before speciation.

2.2. *Filtering Out PPI Noise*

To improve the reliability of the data,[17] we built a PPI graph for each assay method, with directed edges (b, p) indicating the roles of each protein as (b)ait or (p)rey where appropriate. We define the *baitrank* for a protein p in species s for assay type a as the fraction of proteins in species s having (non-zero) out-degree less than p's out-degree for assay a. The *preyrank* is defined analogously on the in-degree. If a given out-degree or in-degree is zero then the corresponding baitrank or preyrank is undefined.

We concluded that assay type a was *inconsistent* for protein p if both p's baitrank and preyrank were defined and $|baitrank(p) - preyrank(p)| > 0.1$. Under these conditions we removed data from that assay type for p. After this filtering, degrees were estimated using the naive method[17] from the remaining data. Overall, this approach removed 36% of the data, and led to stronger results. Full datasets and descriptive statistics are available at *http://bcb.cs.tufts.edu/hubsPSB09*.

2.3. *Ranking Protein Degrees*

Degrees of proteins in different species cannot be directly compared because each species has a different degree distribution. To address this issue, we ranked the proteins within each species by their degree in the filtered dataset for that species, and worked only with these ranks instead of the raw degree. We define r, the rank of protein p with degree d in species s, as the fraction of proteins in species s having non-zero degree less than d.

Individual proteins are labeled as hubs using a straightforward threshold on the rank. The hub threshold t is defined on the interval [0.5,1) since intuitively, it makes little sense to talk about a protein with rank less than 0.5 as a hub.

2.4. *Degree Conservation*

A protein is considered to have been *evolutionarily conserved* if it has one or more orthologous proteins in one or more of the other species under consideration. Most of the literature on conservation of hub proteins has focused on this type of sequence conservation.[22,23] At the other end of the spectrum, one can consider whether specific PPIs have been conserved, given pairs of proteins known to have orthologs in both species. (Such conserved interactions are frequently referred to as *interologs*.[6])

In this paper, we focus on the middle ground - looking at whether the interaction degree of a protein is conserved. This conservation can occur whether or not the specific interologs are maintained. We define a pair of orthologous proteins as *degree-conserved* at hub threshold t if both proteins are hubs at hub threshold t in their respective species. This definition can be naturally extended to more than two species.

3. Having an Orthologous Hub Increases Hub Likelihood

3.1. *Degree Conservation in Ortholog Pairs*

We first analyzed all pairs of orthologous proteins among the four species. We created a dataset D containing all proteins in the filtered data such that there are at least two species in which the proteins are orthologous. We were interested in determining whether proteins that have a hub ortholog have a significantly higher probability of being a hub protein.

To do this, we compare the observed posterior probability that a protein is a hub, given that it has an ortholog in another species that is a hub, to the prior probability of a protein being a hub in our data set. A simple way to define the prior probability of being a hub is just $(1 - t)$, where t is the hub threshold. We call this the *uninformed prior*, $U(t)$. However, since

Pacific Symposium on Biocomputing 14:391-402 (2009)

we are considering only proteins with orthologs in 1 of the 3 other species under consideration, and since it has been shown that proteins with distant orthologs have higher average degree,[24] the true prior may be higher. Thus, we compute the *informed prior*, $I(t)$. Let $hub_t(p)$ indicate that p is a hub at threshold t and let $orth(x, y)$ indicate that x and y are orthologs. Then $I(t) = Pr[hub_t(p) = 1]$, which represents the observed fraction of proteins in the data set that are hubs. We compute $I(t)$ by dividing the number of proteins in D with rank $\geq t$ by the total size of D.

We then define the posterior probability $P(t) = Pr[hub_t(p_i) = 1 \mid \exists\, p_j$ in another species, s.t. $orth(p_i, p_j) \wedge hub_t(p_j) = 1]$. To compute $P(t)$, we first create the set S_t containing all proteins that have an ortholog in another species that is a hub at threshold t. Then $P(t) = |\{p_i \in S_t \mid hub_t(p_i) = 1\}|/|S_t|$. Intuitively, $P(t)$ is the fraction of proteins in S_t that are hubs at threshold t. Figure 1a shows $P(t)$, $I(t)$ and $U(t)$.

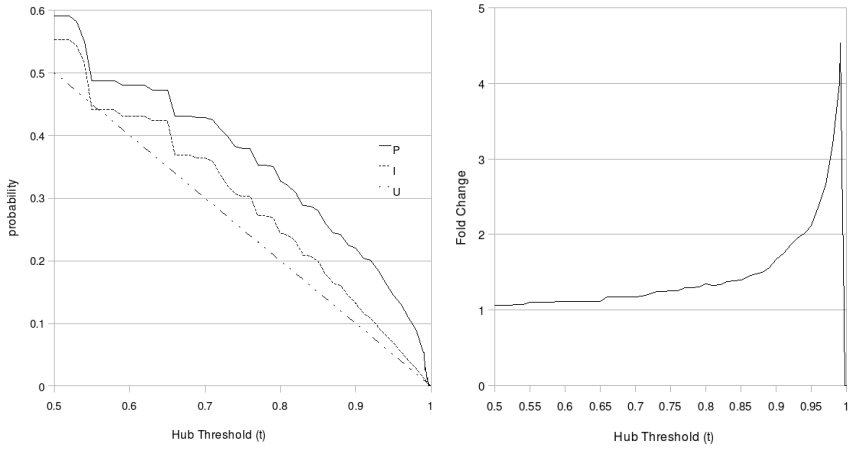


Fig. 1. Hub likelihood plots. a) Conditional probability of being a hub given that an ortholog is a hub, compared to the prior probability of being a hub. The solid line represents P(t), the posterior (conditional) hub probability; the dot-dash line U(t) is the uninformed prior probability, and dashed line I(t) is an informed prior based on the observed data. b) Fold Increase F in hub likelihood, given an orthologous hub protein, as a function of hub threshold t.

We compare the observed distribution P to the informed prior I using a Kolmogorov-Smirnov (KS) Test.[25] We used a 1-sided KS Test to determine whether $[P > I]$ on the intervals $t \in [x, 1)$ for each $x \in \{0.5, 0.55, \ldots, 0.95\}$. Tests were significant at $p < 10^{-6}$ on all intervals. These results support

our hypothesis that the probability of a protein being a hub is higher if its ortholog is a hub, implying that hub proteins are degree-conserved more often than expected by chance. In the same way, we determined that I is significantly greater than U, as expected for the reason described above.

To determine the increase in hub probability given the existence of an orthologous hub, we plot the fold-change in probabilities $F(t) = P(t)/I(t)$ in Figure 1b. From the graph of $F(t)$ we can see that the probability a given protein is a hub is greater if it has a hub ortholog and the magnitude of this effect increases with higher hub threshold. Indeed, $F(t)$ peaks at approximately $t = \gamma = 0.99$. When $t > \gamma$ the expected number of conserved hubs quickly approaches zero, due to the diminishing expectation of the number of *degree-conserved* pairs in the data set (the numerator of $P(t)$) for $t > \gamma$. For hub thresholds above $t \approx 0.93$ we see a two-fold increase in hub likelihood.

3.2. *Degree Conservation Across Four Species*

We next extend the pairwise ortholog analysis to look at ortholog groups across all four species. We create 4-component rank vectors $\mathbf{x_i} = [rank(y_i), rank(w_i), rank(f_i), rank(h_i)]$ where y_i, w_i, f_i and h_i are proteins that form an ortholog group across yeast, worm, fly and human. Let n_T be the total number of rank vectors, $n(t)$ be the number of rank vectors that exhibit degree-conservation at hub threshold t, and $O(t) = n(t)/n_T$ be the proportion of ortholog groups that exhibit hub conservation at hub threshold t. We also compute the expected rate of conservation if degree is not conserved by orthology and call this $E(t)$. Then $E(t) = \prod_{c \in M} Pr[r > t | r \in c]$, where M is the full matrix of rank 4-vectors generated from the orthology groups and c represents the column of ranks specific to a particular species. If $O(t)$ is significantly greater than $E(t)$ then we can conclude that degree is in fact conserved by orthology.

We compare the observed rate of 4-species degree-conservation $O(t)$ to the expected rate $E(t)$ using a series of KS tests. We tested each interval $t \in [a, b]$ where $a \in \{0.5, 0.51, \ldots, 0.95\}$ and $b \in \{0.55, 0.56, \ldots, 1.0\}$ and $a \leq (b - 0.05)$. The KS test results showed that $[O > E]$ for all t-intervals fully subsumed by the interval $t \in [0.66, 0.95]$ ($p < 10^{-6}$) and we conclude that high degree is conserved across the four considered species.

The lack of 4-species degree-conservation for $t > 0.95$ may be attributed to incomplete interaction data, low overlap of interactome coverage between species,[26] or a combination of these effects.

To gain an understanding of the functional classes of proteins that are degree-conserved, we performed functional enrichment analysis (DAVID[27])

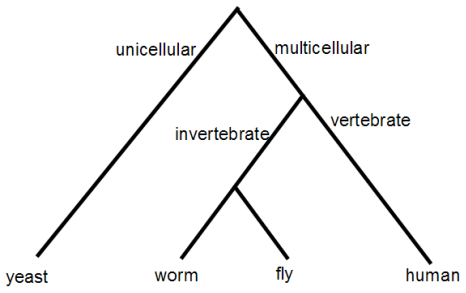
Pacific Symposium on Biocomputing 14:391-402 (2009)

of all yeast proteins that are degree-conserved ($t = 0.8$) in at least 1 of the other 3 species. Enriched functions included primary metabolism, protein synthesis, splicing, DNA repair and regulation of the cell-cycle (FDR < 0.05). Enrichment of basic cellular processes is expected given that these hubs are required to be degree-conserved between yeast and higher eukaryotes. In contrast, degree-conserved ($t = 0.8$) proteins between human and fly were enriched for processes such as cell signalling (the p53, Wnt, ErbB, Notch and TGF-β pathways), cell differentiation, cellular developmental processes, post-translational protein modification and the regulation of protein kinase activity (FDR < 0.05). Full details of functional enrichment results are available online at *http://bcb.cs.tufts.edu/hubsPSB09*.

4. Once a Hub, Always a Hub

Our hypothesis that hubs exhibit degree-conservation relies on the intuition that once a protein has evolved many connections, it is difficult for that protein to lose its importance because so many neighbors would be affected by the change. However, our analyses so far fail to account for orthologs that may have diverged *before* one of the proteins evolved its observed "hub" role. A simple evolutionary model that expects hubs to remain highly connected once they become so might account for a larger fraction of the data. We call this the "once a hub, always a hub" hypothesis.

To test this hypothesis, we describe a simple phylogenetic tree (Figure 2a) relating the four species in our data set. We then seek to determine how closely the data matches this phylogenetic model. For each ortholog group



Consistent YWFH	Inconsistent YWFH
0000	0011
0001	0101
0010	1001
0100	1010
1000	1011
0110	1100
0111	1101
1111	1110

Fig. 2. a) Simple phylogenetic model of the four species in our data set, based on data from the Tree of Life database (www.tol.org). b) Consistency of hub-status bitstrings (yeast,worm,fly,human) with the 'Once a hub, always a hub' hypothesis.

spanning all four species, we form a rank vector as in Section 3.2. We then create a 4-bit bitstring describing the hub status of its component protein

in each species. (If a given protein's rank is greater than the hub threshold t, then that species' component of the bitstring is 1; otherwise it is 0.) An ortholog group is considered *consistent* with the model (with assumption of parsimony) if its bitstring can be explained by the evolution of the hub property at at most one point in the tree (Figure 2b).

We then consider all possible ortholog groups over four species. Figure 3 shows how the fraction of ortholog groups consistent with the model grows as the hub threshold increases. For $t > 0.73$, the majority of the data can be explained by this straightforward phylogenetic model in which the hub property is under selective pressure to be retained once it has evolved.

One caveat is that for sufficiently high t, all the data will have an all-zero bitstring which is consistent with the model. Thus, potentially all the growth that we see in the fraction of consistent ortholog groups is due to that one category. However, we show that this is not the case by repeating the analysis but excluding the all-zero data (Figure 3). Full details of the consistency analysis are provided at *http://bcb.cs.tufts.edu/hubsPSB09*.

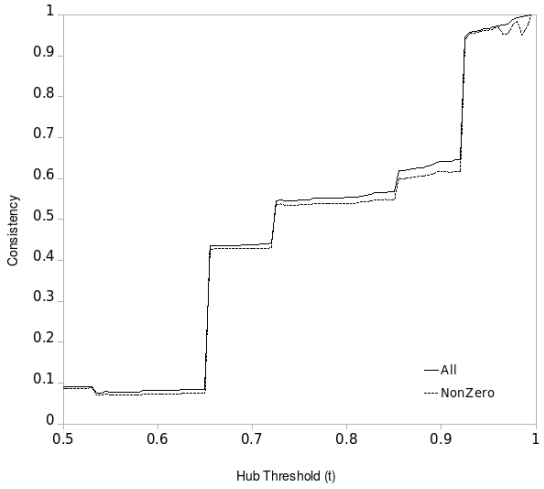


Fig. 3. Consistency with the "once a hub, always a hub" hypothesis, as a function of hub threshold. The y-axis shows the percentage of ortholog groups that are consistent with this hypothesis, given the phylogenetic model from Figure 2. The solid line represents all consistent ortholog groups; the dashed line excludes the all-zero vectors (proteins that are not hubs in any species) from the computation.

5. Hub Interactions Are Preferentially Conserved

Finally, we examine whether the retention of the hub property is due to the retention of individual PPIs, or simply to the general importance of high

degree proteins. We consider two models. In one, interologs are conserved sufficiently well - at least in high degree proteins - that the "module" structure of a hub and its neighbors is retained. This suggests that the function of the protein and its partners is likely to be conserved as well. The alternative hypothesis assumes that once a protein has evolved a large number of partners, it is under pressure to evolve more slowly.[24] Although many of the protein's interaction partners may be lost over time, the high sequence stability of hubs may allow new interactions to arise and be retained more easily. Note that the former hypothesis is somewhat at odds with a growing body of work suggesting that interologs are conserved at only a moderate level,[7,26] and functional modules are generally *not* conserved intact.[9]

To address this question, we analyzed the conservation of individual interactions as a function of protein interaction degree. To measure the conservation of interactions for a given protein p in species s_1, in a designated second species s_2, we rely on the following definitions. First, let G_1 be the protein-protein interaction graph for species s_1: $G_1 = \langle V_1, E_1 \rangle$, and let G_2 be similarly defined for species s_2. Furthermore, let the relation $orth$, defined on $V_1 \times V_2$, denote orthologous pairs of proteins between the two species. Let the potential number of conserved interactions of p, $Poten(p)$ be the number of protein interaction partners q of p in s_1 such that orthologs of both p and q exist in s_2. Also let the actual number of conserved interactions of p, $Conserv(p)$ be the number of protein interactions of p in s_1 that are conserved in s_2. Formally, for protein $p \in V_1$, we define $Poten(p) = |\{q \in V_1 | \langle p, q \rangle \in E_1 \wedge (\exists p', q' \in V_2 | orth(p, p') \wedge orth(q, q'))\}|$ and $Conserv(p) = |\{q \in V_1 | \langle p, q \rangle \in E_1 \wedge (\exists p', q' \in V_2 | orth(p, p') \wedge orth(q, q') \wedge \langle p', q' \rangle \in E_2)\}|$. Then we can compute the conservation rate $R(p) = Conserv(p)/Poten(p)$ which is the proportion of interactions conserved out of all interactions that could possibly have been conserved.

We measured the conservation rate $R(p)$ of each protein in each pair of species compared these to $rank(p)$. Results for the fly proteins are shown in Figure 4. The plot reveals that the proportion of a protein's interactions that are conserved increases as protein rank (or degree) increases. In other words, hubs show preferential conservation of individual protein interactions over non-hubs.

For other species pairs, the graphs are similar in shape, though the overall conservation of the highest degree proteins is often lower (65%-85%). The especially high conservation between fly and human interologs may in part reflect a bias in how protein interactions were chosen as test assays for these particular species. Even for the less similar species pairs,

however, we find that at high hub thresholds, more than half of interologs are conserved. This contrasts sharply with the more general analysis of Mika and Rost,[8] which shows interolog conservation hovering around 25% even at the highest levels of sequence similarity, using data that overlaps significantly with ours. We conclude that the interolog conservation we see with hub proteins may not simply be due to the fact that highly-connected proteins have higher sequence conservation, but that the interactions are truly preferentially conserved.

We also note that for a number of the species pairs, a significant fraction (up to 40%) of the interactions are *not* conserved even at high hub thresholds. This suggests that the degree-conservation of hubs may not rely solely on the conservation of individual pairwise interactions, but on the functional roles of the hub proteins themselves.

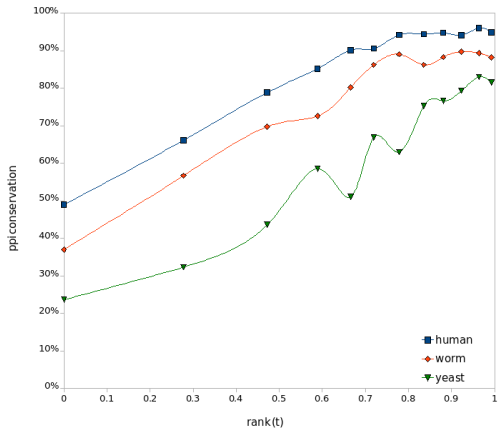


Fig. 4. Interolog conservation as a function of protein interaction degree rank. The data shown are for all fly proteins with orthologs in at least one of the other three species. Similar-rank proteins are binned and their conservation rates averaged so the trend can be seen clearly. Conservation of interologs is low for low-degree proteins, but increases as a function of degree-rank, and is high for proteins above a typical hub threshold of 0.8 or greater.

6. Discussion

We have demonstrated three important and novel facts about conservation of protein interactions and function. First, the likelihood of a protein being a hub increases if it has an ortholog that is a hub. Second, once a protein evolves to become a hub, it tends to remain so, even though specific interaction partners may be gained or lost. Finally, the rate of interolog conservation for hub proteins is much higher than for low-degree proteins.

These results imply that, whatever can be said about functional conservation of modules defined by various clustering methods (e.g., Wang and Zhang[9]), functional modules around network hubs are likely to persist across species. This result has crucial implications for the unspoken assumption underlying most research in model organisms: that the functional relationships between key proteins are conserved between humans and the model organisms being studied. Furthermore, the "once a hub, always a hub" property implies that it might be possible to infer the identity of hub proteins even in organisms where sufficient protein interaction data sets are not yet available. For example, if two orthologous proteins in worm and fly are both PPI hubs, it may be reasonable to assume that the evolution of the hub property occurred in some ancestor of both worm and fly; therefore, an orthologous protein in bees is likely to be a hub as well.

In addition, the stronger-than-expected conservation of individual interactions with hub proteins has important implications for the computational prediction of protein-protein interactions. Although the growth of experimentally-generated protein interaction data is substantial, the interactomes of many organisms are likely to remain largely incomplete for some time, so computational inference of interactions will continue to play an important role. Furthermore, as suggested by Matthews *et al.*,[4] computational predictions can be used to guide the selection of experimental assays, so higher predictive accuracy can reduce the cost of such screening, allowing the generation of more data.

Despite these results, there are many issues surrounding the data that require some caution in interpreting any results based on existing protein interaction networks. Most importantly, the data are both noisy and extraordinarily incomplete. Since different subsets of the interactomes of the four species have been investigated, it is possible that bias in species coverage may affect any conclusions based on the data currently available. In addition, some methods of identifying protein interactions (such as coimmunoprecipitation and affinity precipitation[9]) identify protein complexes rather than pairwise interactions. Many pairwise interactions are sometimes inferred for the participants in these complexes,[28] potentially creating high-degree nodes erroneously. Thus, it will be informative to revisit these conclusions as the quality and quantity of interactome data grow.

Acknowledgments

This research was supported by grant number 004911 to DKS from the National Library of Medicine. We thank the members of the Tufts BCB research group for helpful discussions and comments.

References

1. M. Blanchette and M. Tompa, *Genome Res.* **12:5**, p. 739 (2002).
2. M. Eisen, P. Spellman, P. Brown and D. Botstein, *PNAS* **95**, p. 14863 (1998).
3. J. Yu, V. Smith, P. Wang, A. Hartemink and E. Jarvis, *Bioinformatics* **20**, p. 3594 (2001).
4. L. Matthews, P. Vaglio, J. Reboul, H. Ge, B. Davis, J. Garrels, S. Vincent and M. Vidal, *Genome Res.* **11**, 2120 (2001).
5. J. Han, N. Bertin, T. Hao, D. Goldberg, G. Berriz, D. Dupuy, A. Walhout, M. Cusick, F. Roth and M. Vidal, *Nature* **88**, p. 6995 (2004).
6. A. Walhout, R. Sordella, X. Lu, J. Hartley, G. Temple, M. Brasch, N. Thierry-Mieg and M. Vidal, *Science* **287**, 116 (2000).
7. H. Yu, N. Luscombe, H. Lu, X. Zhu, Y. Xia and J. Han, *Genome Res* **14** (2004).
8. S. Mika and B. Rost, *PLoS Comput. Biol.* **2**, p. e79 (2006).
9. Z. Wang and J. Zhang, *PLoS Comp Biol* **3** (2007).
10. M. Babu, N. Luscombe, L. Aravind and M. Gerstein, *Curr Opin Cell Biol* **14**, 283 (2004).
11. H. Jeong, S. Mason, A. Barabasi and Z. Oltvai, *Nature* **411**, 41 (2001).
12. N. Batada, T. Reguly, A. Breitkreutz, L. Boucher, B. Breitkreutz, L. Hurst and M. Tyers, *PLoS Biol* **4**, p. e317 (2006).
13. S. Wuchty, A. Barabasi and M. Ferdig, *BMC Evol Biol* **6**, p. 1471 (2006).
14. G. Hart, A. Ramani and E. Marcotte, *Genome Biol.* **7**, p. 120 (2006).
15. P. D'haeseleer and G. Church, 216(Aug 2004).
16. T. Chiang, D. Scholtens, D. Sarkar *et al.*, *Genome Biol* **8**, p. 186 (2007).
17. D. Scholtens, T. Chiang, W. Huber and R. Gentleman, *Bioinformatics* **24**, 218 (2008).
18. C. Stark, B. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz and M. Tyers, *Nucleic Acids Res* **34**, D535(Jan 2006).
19. H. Hermjakob, L. Montecci-Palazzi, C. Lewington, S. Mudali, S. Kerrien *et al.*, *Nucleic Acids Res* **32**, D452 (2004).
20. A. Zanzoni, L. Montecci-Palazzi, M. Quondam, G. Ausiello, M. Helmer-Citterich *et al.*, *FEBS Letters* **513**, 135 (2002).
21. M. Remm, C. Storm and E. Sonnhammer, *JMB* **314**, p. 1041 (2001).
22. H. Fraser, A. Hirsh, L. Steinmetz, C. Scharfe and M. Felman, *Science* **296**, 750 (2002).
23. I. Jordan, I. Wolf and E. Koonin, *BMC Evol Biol* **3**, p. 1 (2003).
24. K. Brown and I. Jurisica, *Genome Biol* **8** (2007).
25. W. Press, S. Teukolsky, W. Vetterling and B. Flannery, *Numerical Recipes in C, the Art of Scientific Computing* (Cambridge Univeristy Press, 1992).
26. T. Gandhi, J. Zhong, S. Mathivanan *et al.*, *Nature Gen* **38** (2006).
27. G. D. Jr., B. Sherman, D. Hosack *et al.*, *Genome Biol* **4**, p. 3 (2003).
28. G. Bader and C. Hogue, *Nature Biotechnol* **20**, 991 (2002).