Correspondence

Protein interaction networks beyond artifacts

Sergei Maslov[a,*], Kim Sneppen[b]

First published online 23 September 2002

In their comment Aloy and Russell [1] point at an important property of protein interaction networks generated by high-throughput two-hybrid experiments [2,3]: for poorly understood reasons bait-hybrids of proteins are more prolific binders than their prey counterparts. We were well aware of this fact in our analysis of the large-scale structure of yeast protein interaction network [4]. Precisely for that reason in [4] we compared the protein interaction network in yeast with its random counterpart in which numbers of bait and prey partners of each node are individually conserved. By doing so we accounted for a trivial effect reducing the number of direct connections between hub proteins, namely, that two bait-hybrids are unable to directly connect to each other simply because each two-hybrid interaction has to involve one bait- and one prey-hybrid. Aloy and Russell further claim that the correlation properties of the yeast protein interaction network, reported by us in [4], are an artifact of us using the full dataset of [3], which is known to contain numerous false positives, and not excluding from our analysis the prolific bait-hybrids with ≥ 30 partners.

A straightforward way to validate the correlation pattern found in [4] is to repeat our analysis for several better curated datasets of yeast protein interactions. Here we present the correlation profile [4] of the yeast protein interaction network measured first in the dataset from the high-throughput two-hybrid screen by another group [2] (Fig. 1A), and then in the curated core dataset of ref. [3] (Fig. 1B), which contains only those pairs of interacting proteins, that were independently detected three or more times in the course of the experiment. The combination of these two datasets constitute the majority of entries about yeast protein interactions in most databases, such as e.g. [5,6], and are commonly believed to be reliable. Correlation profiles measured using these datasets (Fig. 1A,B) have the same qualitative features as the one we reported in fig. 2A of [4] for the full dataset of ref. [3]. The only difference is that the characteristic connectivity of hub–proteins, above which direct links between them are suppressed, is reduced from about 30 in fig. 2A of [4] to about 10 in Fig. 1A,B. This confirms that qualitative features of the correlation profile of a network are very robust with respect to false positives and false negatives. Indeed, as previously undetected edges are added to the network (or falsely detected edges are removed from it) the average connectivity of its nodes changes. As a result all large-scale correlation patterns visible in the correlation profile may shift their positions and intensity, but are likely to persist up to a very high level of false positives or false negatives.

Interactions among hubs in Fig. 1B are visibly suppressed for proteins with the sum of bait- and prey-connectivities ≥ 10. The set of 10 hub–proteins responsible for this suppression lacks a strong bait-hybrid bias, which caused the concern of Aloy and Russell about our original study. Indeed, in addition to AGP17 (29 interaction partners as prey and only 9 as bait), discarded as an exception in [1], the set of hubs now includes other highly connected prey-hybrids such as YDL100C (10 partners as prey and 2 partners as bait), and STD1 (9 partners as prey and 1 partner as bait).

Another reason why we do not believe that highly connected hubs are an artifact of the two-hybrid experimental technique is the reproducible observation of the same interaction partners of a given bait-hybrid hub. If, as suggested in the above comment, the mechanism giving rise to such hubs is the low-frequency activation of the reporter gene in the ab-

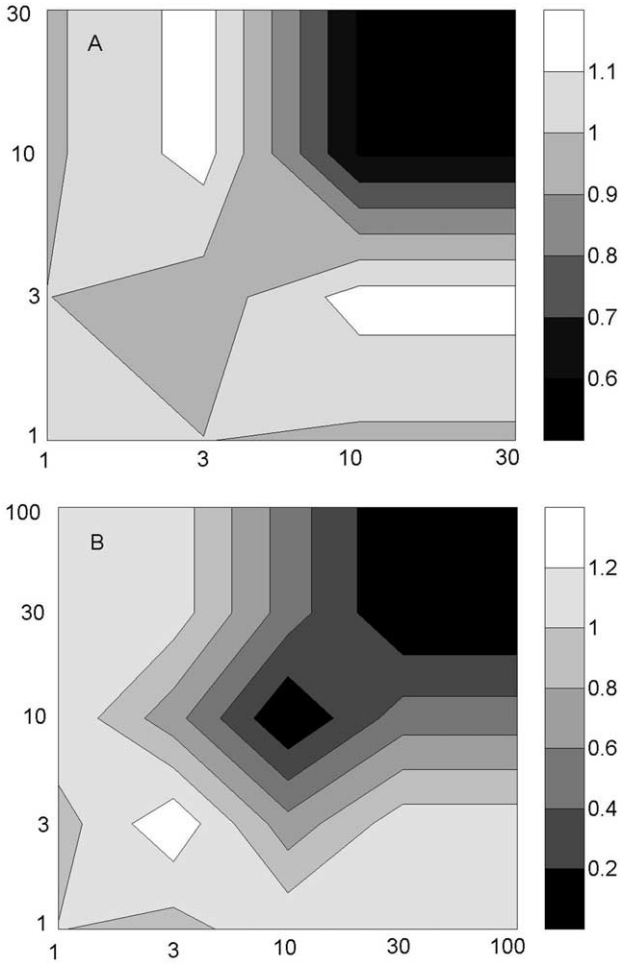

Fig. 1. Correlation profiles of yeast protein interaction network given by the ratio $P(K_0,K_1)/P_r(K_0,K_1)$. Here $P(K_0,K_1)$ is the probability that a pair of proteins with total numbers of interaction partners given by K_0 and K_1 correspondingly, directly interact with each other (A) in the dataset of ref. [2] (957 interactions involving 1004 proteins); (B) in the core dataset of ref. [3] (841 interactions involving 797 proteins). It is normalized by the $P_r(K_0,K_1)$ – the same probability measured in a randomized version of the corresponding network, prepared as described in [4]. Both panels A and B as well as fig. 2A of [4] indicate the suppression of direct connections between highly connected proteins, visible as a dark region in the upper right corner of each plot.

PII: S0014-5793(02)03428-2

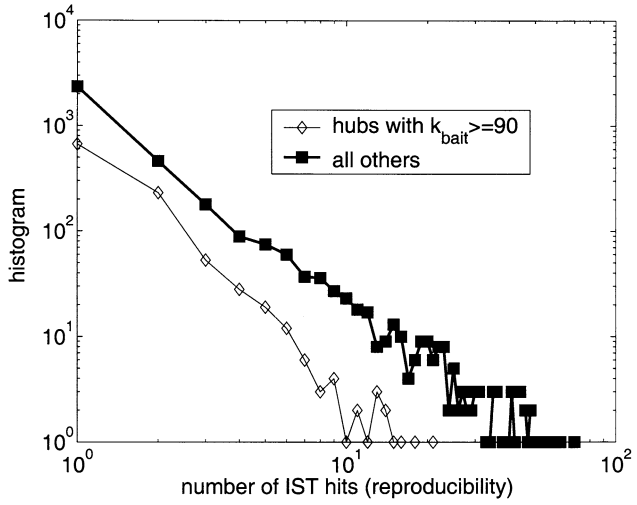

Fig. 2. Histograms of the number of IST hits (reproducibility) of interactions involving eight bait-hybrid hubs with more than 90 partners each (open diamonds), and all other bait-hybrid proteins (filled squares). Note that, apart from an overall number of interactions, these two histograms look very similar to each other confirming that the same interaction mechanisms operate in both cases.

sence of the C-terminal activation domain of the prey-hybrid, then most of the prey-hybrid partners of a given bait-hybrid hub would be observed only once in the course of the two-hybrid experiment. The data set of ref. [3] contains information about how many times a given pair of interacting proteins was detected in the course of experiment (the so-called 'IST hit number'). It is on the basis of this measure of reproducibility Ito et al. filtered out their core set formed by all interacting pairs observed three or more times (IST hit number ≥ 3) in the course of the experiment. As shown in Fig. 2 the distribution of IST hit numbers for interactions involving bait-hybrid hubs with more than 90 interaction partners each is no significantly different from that involving the rest of bait-hybrids, most of which have very low connectivity. This strongly indicates that the same mechanism generates the observed interactions irrespective of the connectivity of a bait-hybrid involved.

Last but not least, the fact that connections between hubs were found to be suppressed also in the yeast transcription regulatory network [4] indicates that, perhaps, such suppression is a robust and universal feature of all bio-molecular networks and not an artefact of a particular experimental technique.

In summary, we advocate our basic statistical approach as a robust way of detecting correlations in any network. When applied to the network of protein interactions derived from high-throughput two-hybrid experiments [2,3], it properly takes into account systematic effects associated with the bait–prey asymmetry, and is robust with respect to a potentially large amount of false positives and false negatives.

References

[1] Aloy, P., Russell, R.B. (2002) FEBS, this issue.
[2] Uetz, P. et al. (2000) Nature 403, 623.
[3] Ito, T. et al. (2001) Proc. Natl. Acad. Sci. USA 98, 4569.
[4] Maslov, S. and Sneppen, K. (2002) Science 296, 910.
[5] Yeast Protein Database (Incyte Genomics, Palo Alto, CA), described in: Costanzo, M.C., et al., Nucleic Acids Res. 29, 75 (2001).
[6] The BIND database (http://www.bind.ca) described in: Bader, G.D., Hogue, C.W. (2000) Bioinformatics 16, 465–477.

*Corresponding author. Fax: (1)-631-344 2918.
E-mail address: maslov@bnl.gov (S. Maslov).

[a] *Department of Physics, Brookhaven National Laboratory, Upton, NY 11973, USA*
[b] *Department of Physics, Norwegian University of Science and Technology, N-7491 Trondheim, Norway*

PII: S0014-5793(02)03428-2