Modeling the Two-Hybrid Detector: Experimental Bias on Protein Interaction Networks

Karin B. Stibius*† and Kim Sneppen*
*Niels Bohr Institute, Copenhagen, Denmark; and †Risø National Laboratory, Roskilde, Denmark

ABSTRACT This work was done to investigate the two-hybrid experiment for finding protein-protein interactions to explain the asymmetry found in the experimental data, and to help screen the data for high confidence interactions. By looking at the bait-prey experimental setup the resulting protein interaction network can be examined as a directed network (bait → prey). We have investigated two possible scenarios for the asymmetry in the directed network by developing a biochemical model for the protein-DNA and protein-protein bindings inside the living yeast. One scenario assumes a background activity of bait proteins acting even without the prey, the other scenario explores the asymmetry in the chemistry associated with the bait being automatically located in the right position on the DNA. We conclude that the latter model gives the best description of the observed asymmetry.

INTRODUCTION

Protein-protein interactions are central for both signaling and structures inside living cells. These interactions can be studied in different ways. One example is to examine complexes formed around a tagged protein with mass spectrometry (1–3); another is to use the two-hybrid method (4–10). In this in vivo method each potential protein interaction is tested by linking one partner to the DNA binding subunit of a transcription factor (bait), and linking the other protein to the subunit that recruits/activates the RNA polymerase (prey). Thereby the activity of the transcribed gene provides information about the strength of the bait-prey interaction. However, as pointed out by the literature (11,12), there are systematic biases in the bait-prey setup. In this article we discuss some of these biases, and present a frame to validate their respective impact on the obtained protein networks.

Constructing a model for the two-hybrid detector could be of great importance if it is to be used to examine protein-protein interactions on a large scale. A model of the system can help to improve the experimental setup and help to screen the data for the most reliable interactions. The work presented here is meant to give an idea of some features that are of importance for the two-hybrid experiment.

RESULTS AND DISCUSSION

Experimental data from large-scale two-hybrid experiments (6,8) was examined and as seen in Fig. 1 there is a systematic difference in connectivity between bait and prey. This difference comes from the bait-prey experimental setup, and the resulting network should thus be examined as a directed network (bait → prey). Further, the asymmetry of the data can be quantified in terms of a systematic tendency with proteins acting as bait having larger connectivity than the same proteins acting as prey.

This asymmetry was investigated by two models; see Fig. 2 and the Method section. The left panel, model I, in Fig. 2 assumes a background activity of bait proteins acting even without the prey (random firing), and the right panel, model II, explores the asymmetry in the chemistry associated with bait being automatically located in the right position on the DNA (sequestering).

Our model I corresponds to the standard explanation for bait-prey bias due to some baits acting as autoactivators—an explanation that at least is right to the extent that a number of proteins are known to activate the reporter gene independent of which prey they are tested with. In large-scale screens these proteins are removed from the data. But other proteins that in themselves weakly activate the reporter might be detected as interaction partners to proteins with which they only interact weakly.

In our investigation we deal with a total of five networks: a), the real network of protein interactions in the cell, henceforth called the ‘‘real network’’; b), the two-hybrid experiment gives us the ‘‘observed network’’; c), we create a ‘‘simulated network’’, and on this simulated network we perform the two different simulated two-hybrid experiments just mentioned, d), model I (random firing); and e), model II (sequestering).

The situation is like reconstructing collision events in nuclear or high energy physics on basis of the incomplete data obtained from the detectors. Thus we want to analyze a situation where

$$original\,network \rightarrow observed\,network, \qquad (1)$$

for both the real/experimental system, and our simulated/model image.

The results from the two models (see Method section and the Supplementary Material) with different values of the

Submitted September 27, 2006, and accepted for publication May 30, 2007.

Address reprint requests to Karin B. Stibius, E-mail: karin_stibius@hotmail.com.

Editor: Ruth Nussinov.

0006-3495/07/10/2562/05 $2.00

doi: 10.1529/biophysj.106.098236

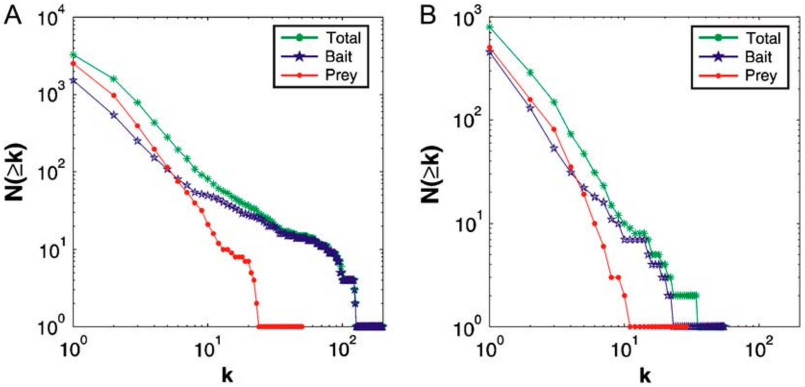

FIGURE 1 (*A* and *B*) Shown is the total connectivity distribution in Ito's (3) two-hybrid data, and the same decomposed into bait and prey parts. Panel *A* is the full data set and panel *B* is the high confidence core data.

parameters used can be seen in Figs. 3 and 4. In Fig. 3 we have investigated the models where we for simplicity have assumed all proteins in the cell to have the same total concentration. For model I (*A*) we do not see a clear difference between bait and prey, whereas for model II (*B*) we see clearly different trends for the two. To see whether this effect would still be present if the protein concentrations of bait and prey were systematically higher than that of other individual proteins, Fig. 4 shows the effects of assuming the total concentrations of the bait and prey to be 10 times that of other proteins. This is relevant because bait and prey are typically expressed on multicopy plasmids. Fig. 4 demonstrate that the systematic differences between bait and prey persists at this more realistic setup, although the effects are less pronounced than for the parameters used in Fig. 3.

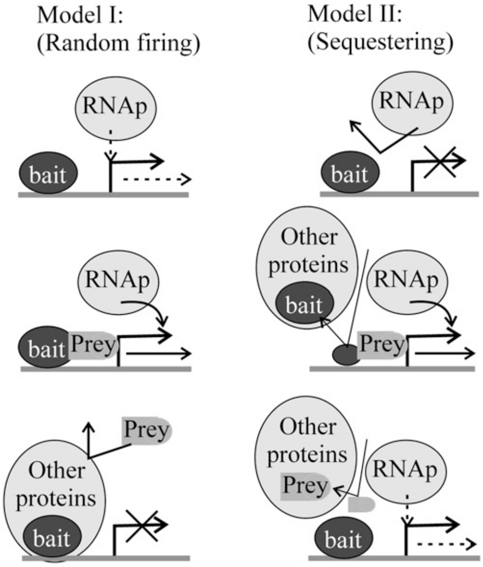

FIGURE 2 The two alternative models for detecting activity of a given bait-prey set. The left panel, model I, shows a schematic representation of the random firing model where bait alone on the DNA operator site gives rise to some transcription (*top*). The bait-prey complex gives full transcription (*middle*) and bait bound to other proteins block for any transcription (*bottom*). The right panel, Model II, shows the sequestering model. A bait alone gives no transcription (*top*). A small free bait concentration does not affect production, because the free bait available will be more localized around the DNA since the binding to the DNA operator site is very strong (*middle*). However, a small free prey concentration will affect production because of the small amount of prey being localized at the right place, since the binding to bait is not strong enough to affect the concentration of prey around the DNA (*bottom*).

To test the robustness of our results against the simplified assumption of having equal concentrations of proteins in the cell we also performed simulations where protein concentrations were drawn from a more realistic distribution. We have chosen to use a distribution similar to the distribution experimentally found by Ghaemmaghami et al. (13), where the protein concentrations are log-normal distributed. To fit our model we gave the protein concentrations log-normal distributed between 0.01 and 5 with a mean value of ~1. Overall, such variations tend to decrease the difference between bait and prey connectivity. However, for larger values of the detection threshold, *T*, the findings from above are nicely reproduced. With varying protein concentrations, the random firing model I always fails to obtain noticeable difference between bait and prey connectivity. In contrast the sequestering model II predicts substantial bait-prey asymmetry for any distribution of protein concentrations, provided a sufficiently large threshold is used. In the Supplementary Material we show figures that substantiate the robustness of our conclusions with respect to initial protein concentrations and choice of threshold values.

Of the two hypotheses for the asymmetry we conclude that model II provides the best explanation for the observed features. For example model II is completely consistent with the fact that more proteins act as prey than as bait. We also find that the high connectivities are mostly seen for proteins functioning as bait, an effect not nearly as pronounced in model I. The fact that proteins with prey connectivity $k_{\mathrm{prey}} = 0$ has surprisingly high values of bait connectivity k_{bait}, is also better explained by the sequestering model II. One explanation of this effect could be that if the connectivity of a protein is very large, the free concentration of the protein will be small; see Fig. 2. When a bait protein has a small concentration, we can imagine that because of the very strong

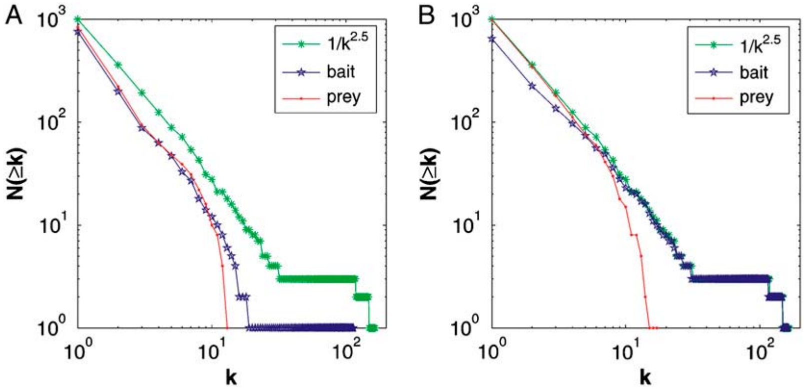

FIGURE 3 (*A*) Shown is the result of the random firing model, with threshold 0.1, assuming an underlying $1/k^{2.5}$ connectivity distribution. (*B*) Shown is the sequestering model with the same parameters. In both cases we investigate the case where bait and prey are at the same concentrations as other proteins in the cell.

binding to the DNA operator site it will be located close to the DNA at all times. For the prey protein, however, this effect does not exist. Here the binding to DNA is only a result of the binding to the bait protein, and this is expected to be a weaker interaction. Proteins with a small free concentration may therefore be seen in the experiment when acting as bait, whereas it will be very difficult for them to be seen binding as prey.

Our approach also opens for analysis of to what extent various network motifs (14) may survive given the bait-prey asymmetry. In particular we find that for triangles of three interacting proteins it is particularly difficult to survive this asymmetry, and thus a triangle in itself should be taken as an indication of a more reliable/stronger interaction.

To the extent that model II describes the data, our analysis suggests that one should believe prey data for prey proteins with low connectivity, and bait data for proteins acting as bait with high connectivity. This conclusion would be softened if model I is also contributing. That is, if baits acting as autoactivators contribute to some additional interactions and thereby make some weak interactions detectable for bait proteins that have autoactivating that in itself was below the detection threshold. In any case prey data missed many links associated to high connectivity proteins.

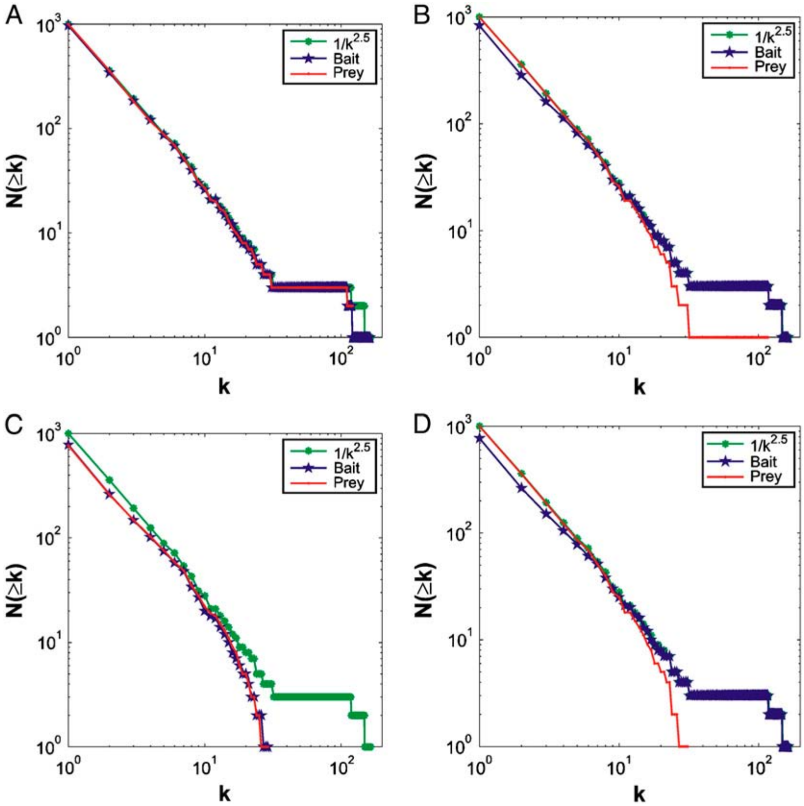

FIGURE 4 The left panel (*A* and *C*) shows the result of the random firing model, assuming an underlying $1/k^{2.5}$ distribution. The right panel (*B* and *D*) is sequestering with the same parameters. In all figures we investigate the case where bait and prey are at 10 times the concentrations of the other proteins in the cell. The upper panel (*A* and *B*) is for threshold 0.1, and the lower panel (*C* and *D*) is for threshold 0.2.

CONCLUSION

We have suggested a possible mechanism for the previously reported (11,15) difference between behavior of proteins when functioning as bait and when functioning as prey. This asymmetry indicates some basic difference in the biochemistry of the ''bait position'' on the DNA and the ''prey position'' in the cell nucleus or cytoplasm. In these terms we indeed found that the sequestering model II explains more of the asymmetry features seen in the two-hybrid data than the random firing model I. Thus the sequestering model explains the nearly exponential distribution of the prey connectivities, that more proteins act as prey than as bait, as well as the effect that some proteins with no bindings as prey have a large number of connections as bait.

METHOD

For further details see Stibius (16).

To examine the models we create a ''simulated network'' (17) consisting of N nodes by assuming that their connectivity, k, is power distributed. That is, we assign each node a connectivity drawn from a probability distribution

$$f(k) \propto k^{-\gamma}. \tag{2}$$

In practice this is done through the tabulated cumulative distribution function, $F(k) = \sum_{c=1}^{k} 1/c^{\gamma} / \sum_{c=1}^{c=N} 1/c^{\gamma}$. For each node, i, one chooses a random number $\eta \in [0;1]$, and selects its connectivity $k_i = k$ such that $F(k-1) \leq \eta \leq F(k)$.

After each node, i, is given a certain connectivity, k_i, the nodes are sorted by descending connectivities and subsequently connected into a network as described by Trusina et al. (17). Finally the network is randomized by link swapping as described by Maslov and Sneppen (15) to generate a truly random connected network with connectivity distribution represented by Eq. 2.

In the models the interest lies in the probability of binding the bait-prey complex to the DNA operator site, because this complex alone is able to activate transcription. The probability of having an operator site, O, with any molecule, X, bound to it can be calculated by Eq. 4, where $K_D = [X]_{free}[O]_{free}/[XO]$ is the binding constant of the molecule X to the operator site O on the DNA. Thus [XO] is found by

$$[XO] = \frac{[X]_{free}[O]_{total}}{[X]_{free} + K_D}, \tag{3}$$

and the probability for the operator site to be occupied by X is:

$$P_{binding,X} = \frac{[XO]}{[O]_{total}} = \frac{[X]_{free}}{[X]_{free} + K_D}. \tag{4}$$

In the following we shall consider two simple approaches for determining this probability.

Model I

In this, the random firing model, we assume that the only molecules that bind are the ones with a bait protein, i.e., free bait, bait-prey complexes, and bait in complex with other molecules, Y. This means that the total concentration of X will be:

$$[X]_{total} = [bait]_{free} + [bait - prey] + [bait - Y] = [bait]_{total}. \tag{5}$$

The number of DNA operator sites, O, is considered to be much smaller than the number of molecules, X, in the cell. Therefore we consider the free X concentration to be equal to the total X concentration in the cell nucleus. The probability, $P_{bait\text{-}prey}$, of seeing the bait-prey complex bound to O is then found by multiplying the fraction of bait prey to the total bait by the probability given by Eq. 4 that a bait molecule is bound to the operator site:

$$P_{bait\text{-}prey} = \frac{[bait - prey]}{[bait]_{total}} \times \frac{[bait]_{total}}{[bait]_{total} + K_D} = \frac{[bait - prey]}{[bait]_{total} + K_D}; \tag{6}$$

see also Shea and Ackers (18). Equation 6 does not give a possibility for any asymmetry between bait and prey. However an asymmetry may arise if a bait bound to O could activate the transcription with a probability below the threshold value.

In the two-hybrid experiment some baits always activate transcription, and these proteins were not used in the final experiment. Thus in model I we make the hypothesis that the bait proteins will have some binding to the RNA-polymerase and thereby activate transcription, but only at a level insufficient for the survival of the cell. This implies that the threshold for seeing a particular bait will depend on the level of activation that bait protein has. In terms of our model this means that the promoter activity associated to the bait-prey complex $P_{bait\text{-}prey}$ is supplemented with an additional activity associated to the bait itself.

We have simulated the extra bait firing by giving each bait a random value, $bait - firing$ ($r_b \in [0, 0.1]$), which is selected from a flat distribution. In the figures in the main text we typically chose threshold $T = 0.1$, just above the maximal random firing, whereas we in the supplement investigate larger T values. The total activity associated with a given bait-prey complex is then calculated from:

$$P_{bait\text{-}prey} = \frac{[bait - prey]}{[bait]_{total} + K_D} + r_b, \tag{7}$$

where we again stress that r_b is a bait property, and thus independent of the particular prey. Thus baits that are assigned a larger value of r_b are relatively easy to detect in complexes.

Model II

In this, the sequestering model, we assume that baits bound to other molecules than prey are unable to bind to the DNA operator site. It could, e.g., be that, Y, bound to the bait is a membrane protein, and the complex therefore is located at the membrane. Thus we will have the molecules that are able to bind to the DNA operator site to be:

$$[x]_{total} = [bait]_{free} + [bait - prey]. \tag{8}$$

From Eqs. 4 and 8 we find the probability of seeing the bait-prey complex:

$$P_{bait\text{-}prey} = \frac{[bait - prey]}{[bait]_{free} + [bait - prey] + K_D}. \tag{9}$$

Calculations

To calculate these probabilities for the two models we need an estimate of the free protein concentrations. These can be found by Eq. 11

$$\begin{aligned}[p_i]_{total} &= [p_i]_{free} + \sum_{j\neq i}^{N} [p_i p_j] + 2[p_i p_i] \\ &= [p_i]_{free} + [p_i]_{free} \sum_{j\neq i}^{N} \frac{[p_j]_{free}}{K_{ij}} + \frac{2[p_i]_{free}[p_i]_{free}}{K_{ii}}, \end{aligned} \tag{10}$$

giving

$$[p_i]_{\text{free}} = \frac{[p_i]_{\text{total}}}{1 + \sum_{j \neq i}^{N} \frac{[p_j]_{\text{free}}}{K_{ij}} + \frac{2[p_i]_{\text{free}}}{K_{ii}}}, \qquad (11)$$

where K_{ij} is the binding constant between the proteins p_i and p_j, and $[p_i p_j]$ is the concentration of the complex formed by binding p_i and p_j.

In the models we assign binding constants, K_{ij}, between all proteins in the network. Two proteins, i and j, that have a connection in the network are given the same binding constant, $K_{ij} = K_{\text{binding}}$. In the model this is given the value, $K_{\text{binding}} = e^{-5} \approx 10^{-2}$ (the smaller the value the stronger the binding). This binding strength should be seen in the perspective that typical concentrations of individual proteins are of the order of one, thus representing fairly strong interactions.

Proteins, p_i and p_k, with no connection in the network are given a binding constant, $K_{ik} = K_{\text{no binding}} = 10^8$, which is so large that we effectively disregard all bindings not present in the assumed network (no false positives are possible).

Finally we select the binding constant $K_D = 10^{-5}$ reflecting a very strong binding of the GAL4 binding region to the operator site. Other values of the binding constants have been investigated (see Supplementary Material) without altering the conclusions given in the main article.

In the first iteration we have used a value of $[p_i]_{\text{free}} = 0.1 \times [p_i]_{\text{total}}$ for all free protein concentrations, but the final free protein concentration is independent of which value is used to begin the numerical simulation. The iteration was continued until the value of the free concentration did not vary more than 10^{-10}.

Now the model networks can be formed by calculating the probability of two proteins binding for each combination of proteins thereby creating a $N \times N$ matrix of probabilities. This matrix can be converted to a network by applying a threshold, T, where node i as bait is connected to node j as prey if $P_{ij} \geq T$, where we have investigated values of T from 0.05 to 0.9; see also Supplementary Material.

SUPPLEMENTARY MATERIAL

To view all of the supplemental files associated with this article, visit www.biophysj.org.

This work was supported by a scholarship from Novo-Nordisk and by the Danish National Research Foundation through the Center for Models of Life.

REFERENCES

1. Gavin, A. C., M. Bosche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. M. Rick, A. M. Michon, C. M. Cruciat, M. Remor, C. Hofert, et al. 2002. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature.* 415:141–147.
2. Kumar, A., and M. Snyder. 2002. Protein complexes take the bait. *Nature.* 415:123–124.
3. Ho, Y., A. Gruhler, A. Heilbut, G. D. Bader, L. Moore, S. L. Adams, A. Millar, P. Taylor, K. Bennett, K. Boutilier, L. Y. Yang, C. Wolting, et al. 2002. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature.* 415:180–183.
4. Fields, S., and O. Song. 1989. A novel genetic system to detect protein-protein interactions. *Nature.* 340:245–246.
5. Chien, C., P. Bartel, R. Sternglanz, and S. Fields. 1991. The two-hybrid system: a method to identify and clone genes for proteins that interact with a protein of interest. *Proc. Natl. Acad. Sci. USA.* 88:9578–9582.
6. Ito, T., T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki. 2001. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. USA.* 98:4569–4574.
7. Ito, T., K. Tashiro, S. Muta, R. Ozawa, T. Chiba, M. Nishizawa, K. Yamamoto, S. Kuhara, and Y. Sakaki. 2000. Toward a protein-protein interaction map of the budding yeast: a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc. Natl. Acad. Sci. USA.* 97:1143–1147.
8. Giot, L., J. S. Bader, C. Brouwer, A. Chaudhuri, B. Kuang, Y. Li, Y. L. Hao, C. E. Ooi, B. Godwin, E. Vitols, G. Vijayadamodar, P. Pochart, et al. 2003. A protein interaction map of Drosophila melanogaster. *Science.* 302:1727–1736.
9. Uetz, P., and M. Pankratz. 2004. Protein interaction maps on the fly. *Nature.* 22:43–44.
10. Stelzl, U., U. Worm, M. Lalowski, C. Haenig, F. H. Brembeck, H. Goehler, M. Stroedicke, M. Zenkner, A. Schoenherr, S. Koeppen, J. Timm, S. Mintzlaff, et al. 2005. A human protein-protein interaction network: a resource for annotating the proteome. *Cell.* 122:957–968.
11. Aloy, P., and R. Russell. 2002. Potential artefacts in protein-interaction networks. *FEBS Lett.* 530:253–254.
12. Mrowka, R., A. Patzak, and H. Herzel. 2001. Is there a bias in proteome research? *Genome Res.* 11:1971–1973.
13. Ghaemmaghami, S., W. Huh, K. Bower, R. W. Howson, A. Belle, N. Dephour, E. K. O'Shea, and J. S. Weissman. 2003. Global analysis of protein expression in yeast. *Nature.* 425:737–741.
14. Shen-Orr, S. S., R. Milo, S. Mangan, and U. Alon. 2002. Network motifs in the transcriptional regulation of *Escherichia coli. Nat. Genet.* 31:64–68.
15. Maslov, S., and K. Sneppen. 2002. Specificity and stability in topology of protein networks. *Science.* 296:910–913.
16. Stibius, K. B. 2004. Analysis and modelling of protein interaction networks—a study of the two-hybrid experiment. Master's thesis. Niels Bohr Institute, University of Copenhagen, Copenhagen, Denmark. http://cmol.nbi.dk/thesis/Karin.pdf. [Online].
17. Trusina, A., S. Maslov, P. Minnhagen, and K. Sneppen. 2004. Hierarchy and anti-hierarchy in real and scale free networks. *Phys Rev Lett.* 92:178702.
18. Shea, M. A., and G. K. Ackers. 1985. The OR control system of bacteriophage lambda—a physical-chemical model for gene regulation. *J. Mol. Biol.* 181:211–230.