nature biotechnology

Global protein function prediction from protein–protein interaction networks

Alexei Vazquez[1,2], Alessandro Flammini[2], Amos Maritan[2,3] & Alessandro Vespignani[4]

Determining protein function is one of the most challenging problems of the post-genomic era. The availability of entire genome sequences and of high-throughput capabilities to determine gene coexpression patterns has shifted the research focus from the study of single proteins or small complexes to that of the entire proteome[1]. In this context, the search for reliable methods for assigning protein function is of primary importance. There are various approaches available for deducing the function of proteins of unknown function using information derived from sequence similarity or clustering patterns of co-regulated genes[2,3], phylogenetic profiles[4], protein-protein interactions (refs. 5–8 and Samanta, M.P. and Liang, S., unpublished data), and protein complexes[9,10]. Here we propose the assignment of proteins to functional classes on the basis of their network of physical interactions as determined by minimizing the number of protein interactions among different functional categories. Function assignment is proteome-wide and is determined by the global connectivity pattern of the protein network. The approach results in multiple functional assignments, a consequence of the existence of multiple equivalent solutions. We apply the method to analyze the yeast *Saccharomyces cerevisiae* protein-protein interaction network[5]. The robustness of the approach is tested in a system containing a high percentage of unclassified proteins and also in cases of deletion and insertion of specific protein interactions.

Two-hybrid experiments allow the reconstruction of the binary interactions among a set of proteins in a given proteome[5]. Our approach visualizes the protein-protein interaction data as a connectivity graph in which the nodes represent the proteins and the edges represent interactions among proteins[11,12] (see **Supplementary Fig. 1** online). Here we explore the concept that interacting proteins may belong to at least one common functional class, and thus knowledge of the functional classification of a subset of the proteins involved in the network may lead to an accurate prediction of the functional classification of the remaining subset of uncharacterized proteins. In principle, every protein could be assigned to one or more functional classes drawn from a set of F possible classes. F is the total number of functions considered and depends on the functional classification scheme used. The more stringent the definition of function used in the classification scheme the greater the number F. Because the functional classification for only a small subset of proteins has been characterized, there remain many proteins for which a function σ, chosen among all F possible functions, must still be determined.

A common approach involves assigning a function to an unclassified protein on the basis of the most common function(s) present among the classified interacting proteins, also known as the 'majority rule' assignment[7,8]. The majority rule derives from the empirical observation that 70–80% of interacting protein pairs share at least one function. In most cases, however, only a few unclassified proteins interact with more than one protein of known function[13]. In addition, in these few cases, the interacting proteins with known functions do not generally share functionalities (**Fig. 1**). In this respect, the majority rule assignment is inconclusive because the analysis does not include the links among proteins of unknown function. The result is that much of the information contained in a reconstructed protein-protein interaction network is not used. More importantly, the final configuration of functions assigned to unclassified proteins should be consistent with the rules used to determine the functions themselves. An unclassified protein with one or more unclassified partner(s) must be assigned functions that are consistent with those assigned to its unclassified partners. These constraints define a process by which assignment of function to an unknown protein influences the majority rule assignment in a self-consistent and iterative manner.

The functional prediction strategy described here is based on global optimization principles. A score that counts the number of interacting partners with the same functional assignment (see Methods section for details) is associated to any given assignment (configuration) of functions for the whole set of unclassified proteins. The score is lower in configurations that maximize the presence of the same functional annotation in interacting proteins. The contribution to the total score of a given functional assignment is computed from the number of classified and unclassified neighbor proteins with that function. Hence, determination of the functions of all unclassified proteins in a network becomes a global optimization problem that can no longer be solved on the basis of the local environment . The optimal function assignment corresponds to the configuration of the lowest score for the whole network. In statistical mechanics, this corresponds to minimizing the energy of a Pott's model with nonhomogeneous boundary conditions[14], the latter being represented by the proteins with known function. The resulting computational problem is 'frustrated'—that is, it has no single solution because of the improbability of satisfying all

[1]Department of Physics, University of Notre Dame, Notre Dame, Indiana 46556, USA. [2]International School for Advanced Studies (SISSA) and INFM, V. Beirut 2-4, 34014 Trieste, Italy. [3]The Abdus Salam International Centre for Theoretical Physics, P.O. Box 586, 34100 Trieste, Italy. [4]Laboratoire de Physique Théorique (UMR du CNRS 8627), Bâtiment 210 Université de Paris-Sud 91405 Orsay Cedex, France. Correspondence should be addressed to A.V. (avazque1@nd.edu).

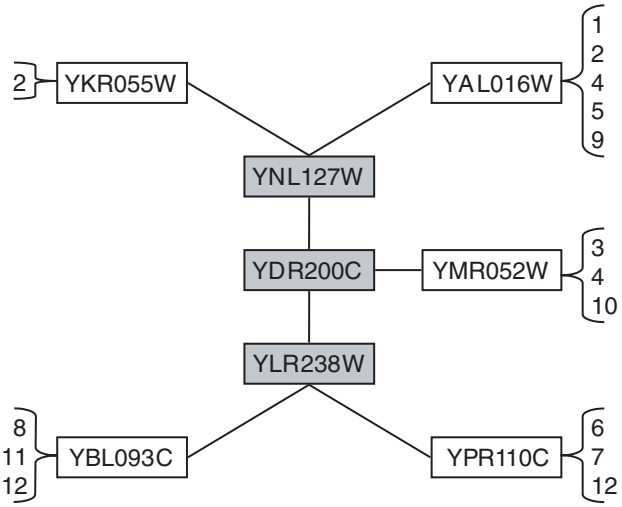

Figure 1 Illustration of the method. Subgraph of the protein interaction network of the yeast *Saccharomyces cerevisiae*. Proteins in gray boxes are unclassified (unknown function); the others are classified proteins (functions in brackets) and are labeled according to the following criteria: 1, cell growth; 2, budding, cell polarity and filament formation; 3, pheromone response, mating-type determination, sex-specific proteins; 4, cell cycle checkpoint proteins; 5, cytokinesis; 6, rRNA synthesis; 7, tRNA synthesis; 8, transcriptional control; 9, other transcription activities; 10, other pheromone response activities; 11, stress response; 12, nuclear organization. Given one of these proteins of unknown function, if we take as a prediction the function that appears more often in the neighbor proteins of known function, then we obtain the following classification (from top to bottom) YNL127W (2), YDR200C (3,4,10) and YLR238W (12). Our method, however, considers also the interactions among unclassified proteins. If we iterate once more the 'majority rule' by taking into account the interactions among the three unclassified proteins, we obtain the following classification: YNL127W (2,4), YDR200C (3,4,10) and YLR238W (12). This way we determined another possible function for YNL127W.

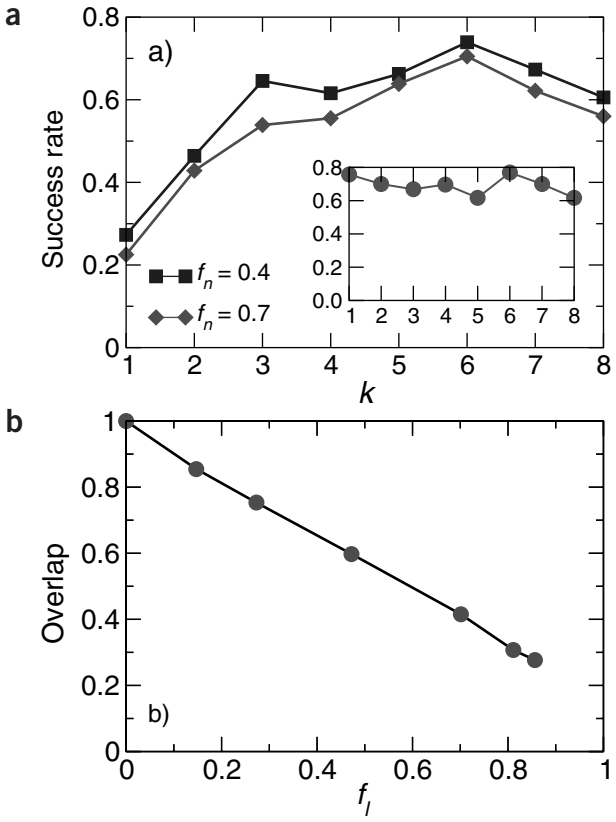

Figure 2 Statistical reliability of the method. (**a**) Self-consistency test. The success rate of the method after a fraction f_n of classified proteins has been unclassified. Each point represents the probability that the functional classification of proteins with k interacting partners, defined here as the top ranking for occurrence in the list of putative functions generated by our method, coincides with their real classification. We report the success rate for the values $f_n = 0.4$ and $f_n = 0.7$ in the upper and lower curve, respectively. The prediction quality for poorly connected nodes (degrees 1 and 2) decreases to just 30%, and it is degrading more rapidly than for highly connected proteins. This is because the corresponding proteins occupy a very marginal position in which the method cannot take full advantage of the global connectivity properties of the graph. In the inset we report the data for $f_n \to 0$, that is, when only a single protein is set unclassified. In this case it is possible to see that even for poorly connected proteins the method gives a very good statistical reliability of the corresponding predictions. (**b**) Tolerance to errors. Overlap of all $\Theta_i(f_l)$ averaged over all unclassified proteins between the functional predictions obtained using the original network and another with a degree of dissimilarity f_l defined as the percentage of edges between protein couples that are different in the two networks. The analysis shows that a moderate amount of misplaced interactions do not preclude a reliable function assignment. Higher numbers of errors lower the overlap, signaling that the two networks provide rather different configurations of functional assignment. The curve shows a decreasing linear trend that when extrapolated to $f_l = 1$ gives a predictably small value of the overlap (<15%). Extrapolation to $f_l = 1$ is inappropriate because a complete dissimilarity between the original and the scrambled network is hardly achievable by a random rewiring, and therefore unjustified in the present context.

the constraints imposed by classified proteins on their interacting, unclassified partners. Instead, multiple equivalent or nearly equivalent optimal solutions are generated that contain a minimal amount of interacting proteins with different functions. The existence of multiple solutions allows the objective assignment of multiple functions to most unclassified proteins (**Fig. 1**). Depending on the complexity of the underlying graph and on the boundary conditions, the score minimization represents a complicated computational task. In instances of this type, a 'simulated annealing' technique[16] (see Methods section) would be an appropriate tool to obtain the optimal solutions. Indeed, the optimization procedure is repeated several times to account for the nonuniqueness of the optimal configurations, and a functional classification prediction is made by taking those functions that occurred more often for each unclassified protein in the whole set of simulated annealing processes.

We have applied the functional prediction method outlined to the analysis of the yeast *S. cerevisiae* protein-protein interaction network. The interaction data were obtained from Schwikowski *et al*[7]. and contain $N = 1826$ proteins with $E = 2238$ identified interactions. The functional classification was obtained from the MIPS database[15]. The MIPS classification scheme contains $F = 424$ functional categories, plus two categories for proteins with no assigned function: 'CLASSIFICATION NOT YET CLEAR-CUT' and 'UNCLASSIFIED PROTEINS'. The data contain $n = 441$ proteins in these two last categories. We used our global optimization method to obtain the functional assignments of all the proteins listed within these two categories. The complete set of functional assignments can be found as **Supplementary Table 1** online. For each unclassified protein we report its degree, that is, the number of proteins directly connected to it and up to three of the most probable predicted functions as determined with our method. We

Table 1 Success rates for global optimization versus majority rule

K	n_k	MR1	GO1	GO2
2	328	0.40	0.46	0.61
3	205	0.55	0.65	0.76
4	102	0.60	0.62	0.77
5	72	0.58	0.66	0.86
6	41	0.66	0.74	0.89
7	28	0.58	0.67	0.94
$k>7$	85	0.69	0.74	0.94

Comparison of the success rate of the global optimization (GO) method proposed here and the majority rule (MR). To compute the success rate, we assume that a fraction $f_n = 0.4$ of the classified proteins are unclassified and then make functional predictions for them. The success rate is defined as the probability that the most ranked predicted function is the actual functional classification for the corresponding protein. Two different levels of functional classification have been used. In the finest level (1) we have taken the most stringent classification, containing 424 functional categories. In the coarse-grained level (2) we have taken the less detailed classification (metabolism, energy, cell growth and division, etc.), containing 20 functional categories. We show the success rate as a function of the number of interacting partners k (as a reference we also show how many proteins n_k have k interacting partners). The case $k = 1$ is not considered because the MR method finds only a trivial implementation in this case. The comparison of the values for $k \geq 2$ clearly indicates that the GO method is more effective, with a higher percentage of correct predictions.

attribute a higher level of 'certainty' to those functions with a higher percentage of occurrences.

A fundamental issue concerning protein function prediction is the assessment of the method reliability in light of the incomplete knowledge of the interaction network. To determine the confidence limit of the method, we determined the rate of successful predictions attained for a fraction f_n of classified proteins that were analyzed as unclassified proteins. This way we obtained a quantitative estimate of the reliability of our predictions as a function of the amount of information available about the network. We show graphically (**Fig. 2a**) the rate] of successful predictions as a function of the degree (number of interacting partners) of the proteins for different values of f_n using the most stringent functional classification scheme available (424 functional classes). For unclassified proteins with degree larger than 2, a correct prediction can be made between 60 and 70% of the time, independently of the degree of the protein involved in the prediction, and despite the loss of a substantial part of information on known classifications (up to $f_n = 0.4$). A brief examination of **Supplementary Table 2** online will permit a visual inspection of the test for $f_n = 0.4$. A quantitative account of the better performance of our method with respect to local optimization methods is presented in **Table 1**, where we also report predictions obtained with a coarser classification scheme. The more coarse-grained the classification, the higher the success rate. Not surprisingly, adopting a coarser classification scheme leads to an increase of the various rate of success (last column), because the number of degrees of freedom the method must deal with is drastically reduced. This has to be balanced with the parallel diminution of the information content of predictions. These results demonstrate the considerable and robust predictive power of this statistical method, even with a reduced amount of information (larger number of unclassified proteins).

A concern in implementing network-based predictive methods is the topological accuracy of the protein network. It is known that protein-protein interaction data obtained from two-hybrid experiments contain a certain number of false positive and negative results that could, in principle, compromise the quality of the predictions by incorporating spurious connectivities into the network (false or missing edges). The effect of this uncertainty on prediction accuracy can be modeled by 'rewiring' a certain fraction of protein interactions—that is, removing every reported interaction with a probability q and drawing new interactions among proteins that do not interact according to the available data. Thus we obtain a new network that is dissimilar to a certain degree, depending upon q, from the original one. The degree of dissimilarity f_l is measured as the percentage of edges between protein pairs that are different in the two networks, the original and the scrambled one. Note that moving one link in general implies that the connectivity pattern of four nodes is affected and that f_l thus has a nontrivial dependence on q. We implemented our method on the modified network, determining a new list of putative functions for each unclassified protein, together with the relative probability (or frequency) of occurrence of the putative functions themselves. For convenience we imagine these lists (one for each unclassified protein) to contain all possible functions, and associate a zero probability to those functions that have never occurred in the implementation of the method. We call $p_{is}(f_l)$ the probability that the unclassified protein i belongs to the functional class s, in the network with a degree of dissimilarity f_l with the original one. The case $p_{is}(0)$ then corresponds to the functional classification obtained using the original network. A quantitative comparison with the predictions made using the original network is provided by the overlap function $\Theta_i(f_l)$ defined as follows: $\Theta_i(f_l) = \Sigma_s [p_{is}(0)p_{is}(f_l)]^{1/2}$, which equals 1 when $p_{is}(f_l) = p_i(0)$ for all s. We compute the average of $\Theta_i(f_l)$ restricted to unclassified proteins with k interacting partners, and observe that it varies little with the node degree. We plot the average of $\Theta_i(f_l)$ over all unclassified proteins as a function of f_l (**Fig. 2b**). For a 10% dissimilarity, the overlap is ~0.85%. Because each displaced edge corresponds to three to four proteins with different interactions, this plot shows that even if ~30–40% of proteins have at least one misplaced interaction due to erroneous experimental results, the determination of the proteins' functions can still be effective. Of course higher levels of errors decrease the overlap, signaling that the two networks provide rather different configurations of functional assignment.

The method we propose can be used as a general tool for the assignment of protein function and demonstrates that protein-protein interaction data can be an effective framework to deduce the function of unclassified proteins. The method also allows determination of multiple functions and takes into account self-consistently the effect of unclassified proteins in the final assignment configuration. Finally, the validity tests conducted show that the method tolerates the inherent imperfection and the incomplete nature of the protein networks.

METHODS

A function σ_i chosen among the F ($F = 424$ in the finest MIPS classification scheme) possible ones is assigned to each unclassified protein $i = 1, 2, ..., n$, to globally minimize the following score function:

$$E = -\Sigma_{i,j} J_{ij} \delta(\sigma_i,\sigma_j) - \Sigma_i h_i(\sigma_i) \quad (1)$$

where J_{ij} is the adjacency matrix of the interaction network for the unclassified proteins (J_{ij} is equal to 1 if protein i and j interact and are unclassified, 0 otherwise), $\delta(i,j)$ is the discrete δ function and $h_i(\sigma_i)$ is the number of classified partners of protein i with function σ_i.

The majority rule[5,6] seeks to minimize only the second term on the right-hand side of equation (1). This can be achieved with local methods (that is, considering successively and independently each protein). Here, the contribution to the total score of assigning a protein i to functional class σ_i depends also on the assignment made for all other proteins, resulting in a more complicated

computational task. The advantage is that the underlying requirement that 'interaction requires a common function' is applied also to interactions between formerly unclassified proteins, information that is ignored in the majority rule approach.

To overcome the computational difficulties and find the configuration or configurations that minimize E, we conduct simulated annealing[16] introducing an effective temperature T. We start with an initial random configuration σ_i. Then, at each Monte Carlo step, we select one protein at random and change its state from σ_i to σ'_i, where σ'_i is selected at random among the possible states of protein i with the constraint $\sigma'_i \neq \sigma_i$. We then compute the score difference $\Delta E = E' - E$ between these two configurations. If $\Delta E \leq 0$, we accept the new configuration. If $\Delta E > 0$, we accept the new configuration with probability $r = \exp(-\Delta E/T)$ or keep the original configuration with probability $1 - r$. This Monte Carlo step is repeated until E reaches a stationary value. Thereafter, T is decreased by a small amount (for the simulations presented here, the inverse of T was increased at constant steps of size 0.01; no significant difference was observed for smaller increments). These two processes, equilibration at a given T and decrease of T, are repeated until the protein states stabilize. These protein states become the predicted functional classification. Because the minimum energy solution is not unique, the simulated annealing process is been repeated several times and starting from different initial configurations (100 times for the simulations presented here; no significant change was observed for a larger number of realizations). Finally, we computed the fraction of times (p_{is}) the protein i was observed in the final state s, which gives us an estimate of the probability that protein i belongs to the functional classification s.

Note: Supplementary information is available on the Nature Biotechnology website.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Received 6 August 2002; accepted 21 February 2003
Published online 12 May 2003; doi:10.1038/nbt825

1. Hodgman, T.C. A historical perspective on gene/protein functional assignment. *Bioinformatics* **16**, 10–15 (2000).
2. Zhang, M.Q. Promoter analysis of co-regulated genes in the yeast genome. *Comput. Chem.* **23**, 233–250 (1999).
3. Harrington, H.C., Rosenow, C. & Retief, J. Monitoring gene expression using DNA microarrays. *Curr. Opin. Microbiol.* **3**, 285–291 (2000).
4. Pellegrini, M., Marcotte, E., Thompson, M.J., Eisenberg, D. & Yeates, T.O. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Nat. Acad. Sci. USA* **96**, 4285–4288 (1999).
5. Uetz, P. *et al.* A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**, 623–627 (2000).
6. Ito, T. *et al.* 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. Nat. Acad. Sci. USA* **98**, 4569–1147 (2001).
7. Schwikowski, B., Uetz, P. & Fields, S. A network of protein-protein interactions in yeast. *Nat. Biotechnol.* **18**, 1257–1261 (2000).
8. Hishigaki, H., Nakai, K., Ono, T., Tanigami, A. & Tagaki, T. Assessment of prediction accuracy of protein function from protein-protein interaction data. *Yeast* **18**, 523–531 (2001).
9. Gavin, A. *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141–147 (2002).
10. Ho, Y. *et al.* Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415**, 180–183 (2002).
11. Wagner, A. Robustness again mutations in genetic networks of yeast. *Nat. Genet.* **24**, 355–361 (2000).
12. Jeong, H., Mason, S.P., Barabasi, A.L. & Oltwai, Z.W. Lethality and centrality in protein networks. *Nature* **411**, 41 (2001).
13. Meyer, M.L. & Hieter, P. Protein networks—built by association. *Nat. Biotechnol.* **18**, 1242–1243 (2000).
14. Wu, F.Y. The Potts Model. *Rev. Mod. Phys.* **54**, 235–268 (1982).
15. The MIPS Comprehensive Yeast Genome Database (CYGD), http://mips.gsf.de/proj/yeast/CYGD/db/.
16. Kirkpatrick, S., Gelatt, C.D. & Vecchi, M.P. Optimization by simulated annealing. *Science* **220**, 621–680 (1983).