FEBS Letters 580 (2006) 1891–1896

Faster and more accurate global protein function assignment from protein interaction networks using the MFGO algorithm☆

Shiwei Sun[a,1], Yi Zhao[a,1], Yishan Jiao[a], Yifei Yin[a], Lun Cai[a], Yong Zhang[b], Hongchao Lu[a], Runsheng Chen[a,b], Dongbo Bu[a,*]

[a] *Institute of Computing Technology, Chinese Academy of Sciences, Beijing, PR China*
[b] *Institute of Biophysics, Chinese Academy of Sciences, Beijing, PR China*

Received 5 December 2005; revised 11 February 2006; accepted 20 February 2006

Available online 28 February 2006

Edited by Robert B. Russell

Abstract Motivation **Predicting protein function accurately is an important issue in the post-genomic era. To achieve this goal, several approaches have been proposed deduce the function of unclassified proteins through sequence similarity, co-expression profiles, and other information. Among these methods, the global optimization method (GOM) is an interesting and powerful tool that assigns functions to unclassified proteins based on their positions in a physical interactions network [Vazquez, A., Flammini, A., Maritan, A. and Vespignani, A. (2003) Global protein function prediction from protein–protein interaction networks, Nat. Biotechnol., 21, 697–700]. To boost both the accuracy and speed of GOM, a new prediction method, MFGO (modified and faster global optimization) is presented in this paper, which employs local optimal repetition method to reduce calculation time, and takes account of topological structure information to achieve a more accurate prediction.**

Conclusion **On four proteins interaction datasets, including Vazquez dataset, YP dataset, DIP-core dataset, and SPK dataset, MFGO was tested and compared with the popular MR (majority rule) and GOM methods. Experimental results confirm MFGO's improvement on both speed and accuracy. Especially, MFGO method has a distinctive advantage in accurately predicting functions for proteins with few neighbors. Moreover, the robustness of the approach was validated both in a dataset containing a high percentage of unknown proteins and a disturbed dataset through random insertion and deletion. The analysis shows that a moderate amount of misplaced interactions do not preclude a reliable function assignment.**

Keywords: Protein interaction network; Function prediction; Global optimization

1. Introduction

Since the first complete genome was sequenced in 1995, genome sequences of more than 100 organisms and thousands of genes have been made available. This explosion in genome sequences has greatly increased the number of predicted proteins relative to that of experimentally characterized proteins. Understanding the function of the predicted proteins is instructional to the exploration of cellular and physiological mechanisms of organisms. However, experimentally determining protein function continues to be laborious and time-consuming since considerable resources are generally required. Hence, to develop a reliable computational method for protein function assignment is of great importance.

To date, a variety of algorithms have been proposed to deduce the function of proteins based on sequence similarity, clustering patterns of co-regulated phylogenetic profiles [5,11], protein complexes [6,8], and other information. In general, the methods using sequence similarity assume that proteins sharing similar primary sequences and secondary structures tend to possess similar or related functions. And other methods using clustering patterns of co-regulated phylogenetic profiles or protein complexes always assume that the proteins that function together in a pathway or structural complex are likely to evolve in a correlated fashion [11]. Though these methods successfully predict the functions for some proteins, they suffer from several limitations, which make themselves not very suitable for eukaryote genes and complete genomes, and their range of application relatively narrow [2].

Recently, the large-scale protein–protein interaction datasets enlighten several interesting and powerful strategies [2,4,7,8,13]. Moreover, functional associations between proteins was also explored through a combination of protein interaction and sequence homology, domain or structure information [5,9]. Breaking the above mentioned constraints, these methods enrich the area of protein functional prediction by utilizing topological interaction patterns of a protein–protein interaction network. Among these methods using a large-scale interactions, the popular 'majority rule' (MR) assignment method [12] assigns a protein with the function that is the most common among its neighbors. However, MR suffers from a disadvantage that only classified neighbors were taken into account, while the information of unclassified ones was neglected. To overcome this shortcoming, the global optimization method (GOM) was developed to utilize not only the information of classified neighbors of proteins, but also the unclassified neighbors [13]. By adding the ante-result of prediction to the protein–protein interaction network for the next round of prediction, GOM achieves a higher accuracy of prediction. In this method, the global optimization procedure is

☆ Availability: All predictions and software are freely available from http://www.bioinfo.org.cn/MFGO/.

*Corresponding author.
E-mail addresses: bdb@ict.ac.cn, dwsun@ict.ac.cn (D. Bu).

[1] These authors contributed equally to this manuscript.

0014-5793/$32.00
doi:10.1016/j.febslet.2006.02.053

repeated several times and a functional classification prediction is made by choosing those functions that occur most often for each unclassified protein in the whole set of simulated annealing processes. However, GOM performs slowly since the repeated simulated annealing process is very computationally intensive.

To boost both the accuracy and speed of GOM, a new approach, MFGO (modified and faster global optimization), is proposed in this paper, which avoids the high intensive computation of the repeated simulated annealing process and utilizes more topological information than GOM. In MFGO, the protein i is assigned with a functional vector $w_i = (w_{i1}, w_{i2}, \ldots, w_{im})$, where m denotes the number of protein functions, and each entry w_{is} denotes a weight to measure the confidence that the protein i has the function s. In other words, MFGO is essentially a continuous strategy to solve a similar objective function as GOM, merely relaxing the statement that only one entry can take the value $w_{is} = 1$ (all the others being 0) to $0 \leqslant w_{is} \leqslant 1$, $\sum_{s=1}^{m} w_{is}^2 = 1$. Our reasoning is as follows: first, it is possible that one protein may exert two or more functions with different intensities. Secondly, it is easy to see that the minimum k-Terminal [ermine] Cuts problem, which is NP-hard even for $k = 3$, can be reduced to this optimization problem under the original assumption, and continuous programming is much more efficient than a discrete optimization one to solve this problem.

In MFGO, the iterative local optimization algorithm starts at a random point and searches for optimized result step by step, which obviously eases the computational intensity compared to a discrete and global optimized algorithm. In addition, we take into account the influence of non-adjacent proteins with common neighbors. The above modifications bestow MFGO a significant improvement on both speed and accuracy of prediction. It should be noticed that the prediction accuracy decreases when the number of false interaction increases, though MFGO shows robustness for the false interaction, which means MFGO will work better on a more accurate interaction and annotation data.

2. Method

2.1. Datasets

On four protein–protein interaction datasets of yeast *Saccharomyces cerevisiae*, MFGO was tested and compared with GOM. Here, the functional classification information was obtained from the MIPS database [10], containing 424 functional categories, plus two categories for proteins with no assigned function, 'CLASSIFICATION NOT YET CLEAR-CUT' and 'UNCLASSIFIED PROTEINS'. For example, in Vazquez dataset [13] including 2238 identified interactions among 1826 proteins, 441 proteins are labeled as 'UNCLASSIFIED PROTEINS' according to MIPS database. MFGO method attempts to generate an accurate function assignment for these proteins with no assigned function.

The other three larger and more complex datasets we used for a further comparison are the *DIP-core dataset* [3] involving 6574 physical interactions among 2608 proteins, the *SPK dataset* [1,6,8,10] involving 13 344 physical interactions among 4537 proteins and the *YP dataset* with 11 855 high and medium confidence interactions among 2617 proteins [14] (see Supplements for more details).

2.2. Algorithm

Let n denotes the number of proteins in our dataset and m denotes the number of protein functions ($m = 424$ in the finest MIPS classification scheme). For each unclassified protein i ($1 \leqslant i \leqslant n$), let $W_i = (w_{i1}, w_{i2}, \ldots, w_{im})$ denote the functional vector of protein i satisfying $\sum_{s=1}^{m} w_{is}^2 = 1$, where w_{is} measures the tendency that protein i has the sth function. For each i ($1 \leqslant i \leqslant n$), let neighbor($i$) be the set of all proteins adjacent to protein i. Our goal is to maximize the following objective function:

$$E = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{s=1}^{m} \rho_{ij} w_{is} w_{js} + \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{s=1}^{m} \lambda_{ij} w_{is} w_{js}, \quad (1)$$

where for each adjacent pair (i,j), $\rho_{ij} = 1$, $\lambda_{ij} = 0$, and for each non-adjacent pair

$$(i,j), \quad \rho_{ij} = 0, \quad \lambda_{ij} = \frac{|\text{neighbor}(i) \cap \text{neighbor}(j)|}{|\text{neighbor}(i) \cup \text{neighbor}(j)|}.$$

The first term in the right hand of formula (1) is the same as GOM, while the second term is unique to MFGO. The second term accounts for those proteins without direct interaction between each other but sharing the same interactors, since it was reported that the more common interactors two proteins share, the more likely they are to be functionally related. Hence, MFGO takes into account more topological and biological information than GOM does.

To optimize this score function, we adopted an iterative local optimization algorithm as follows. For each classified protein i, w_{is} in the function vector $W_i = (w_{i1}, w_{i2}, \ldots, w_{im})$ was set to 1 if protein i has function s, otherwise 0. For all unclassified proteins, w_{is} is initialized randomly, and updated at each step as follows: $w'_{is} = \sum_{j=1}^{n}(\rho_{ij} + \lambda_{ij}) w_{js}$, $1 \leqslant s \leqslant m$, and then we normalized the functional vector W'_i to meet $\sum_{s=1}^{m} w_{is}^2 = 1$.

In each step, the distance between the previous and the current value of $W_i = (w_{i1}, w_{i2}, \ldots, w_{im})$ was calculated as follows $\Delta S = \sum_{s=1}^{m} (w'_{is} - w_{is})^2$. The iteration was repeated until no significant difference of distance was observed, that is, $\Delta S < \xi$ (We adopted $\xi = 1 \times 10^{-12}$ in this paper, and no significant influence was observed for smaller ξ). Finally, a functional vector $w_i = (w_{i1}, w_{i2}, \ldots, w_{im})$ is computed for each unclassified protein, and the weight w_{is} indicating the confidence that the protein i has the sth function. The details of MFGO are given in Fig. 1.

3. Results

Experimental results demonstrate that our continuous iterative algorithm is about 30 times faster than GOM, and is the highest accurate one compared with others, i.e., MR, GOM. In addition, the robustness of MFGO is also confirmed through experiment on disturbed interaction network. Similar results were also observed on another large-scale yeast protein–protein interaction. (All prediction results are available from http://www.bioinfo.org.cn/MFGO/.)

3.1. Speed-up effect of MFGO

In essence, the simulated annealing in GOM is a discrete optimization procedure, which assumes that each protein has only one function. In GOM, simulated annealing was repeated 100 times, and the most frequently reported function was chosen as the candidate, which makes GOM quite computationally intensive. It is well known that a protein may have more than one function. Hence, for each protein i a weight w_{is} can be assigned to measure its possibility of possessing function s ($1 \leqslant s \leqslant m$, $\sum_{s=1}^{m} w_{is}^2 = 1$, where m is the total number of functional categories), then the optimization problem is transformed into a much easier continuous programming one. Since the objective function is not convex, we cannot guarantee that a local maximum is also a global one. None of the standard routines for solving continuous programming problems seems applicable to this case, therefore we designed an iterative local optimization algorithm to solve it.

Algorithm MFGO

Input	The number of unclassified proteins n, the number of protein functions m, the adjacency matrix M of the protein interaction network.
Output	The functional vectors $W_1, W_2, \cdots, W_n$ of all unclassified proteins.
1.	Compute all the ρ_{ij} and λ_{ij}, $1 \le i, j \le n$.
2.	Generate the initial configuration $W_1, W_2, \cdots, W_n$ for a unclassified protein randomly
3.	Let $d = 100, \varepsilon = 0.01$.
4.	While $d > \varepsilon$
	(a) Let $d = 0$.
	(b) For i from 1 to n do
	(i) For each $1 \le s \le m$, compute $w'_{is} = \sum_{j=1}^{n} (\rho_{ij} + \lambda_{ij}) w_{is}$
	(ii) Normalize the functional vector W'_i according to $\sum_{s=1}^{m} w'^2_{is} = 1$.
	(iii) Compute $d = \sum_{s=1}^{m} (w'_{is} - w_{is})^2$. if $d > \varepsilon$ then let $l = i$ and break.
	(c) If $d > \varepsilon$ then update W_l by W'_l.
5.	Output the functional vectors $W_1, W_2, \cdots, W_n$ of all unclassified proteins.

Fig. 1. The modified and faster global optimization (MFGO) algorithm for protein function prediction.

To a certain extent, the iterative local optimization algorithm saves computational resources and enhances the speed of operation compared with GOM. Moreover, the convergence of the iteration process could be proved strictly (see Supplemental material or http://www.bioinfo.org.cn/MFGO/ for the details of proof).

On the same hardware platform (CPU: 2.8 GHz Intel Pentium IV, 2 GB RAM, Red Hat Linux 9.0), comparison was made between GOM and MFGO, both of which are implemented in Java. Because the original source code of GOM is unavailable, we ourselves implemented it according to the paper [13]. Both packages can be downloaded freely from our websites http://www.bioinfo.org.cn/MFGO/ for using and verifying.

On the Vazquez dataset, both of the algorithms were run 100 times to reduce the effect of randomness. As a result, we found that the speed of MFGO was up to 30 times higher than GOM averagely. See Table 1 for the details of the above comparison.

3.2. Enhanced accuracy of function prediction

To enhance the accuracy of protein function prediction, more topological and biological information than GOM are introduced into MFGO. GOM are based on the observation that the interacting proteins tend to share common function. Besides the above observation, MFGO adopted the prior knowledge that proteins without direct interaction but sharing common inter-actors would also tend to share common functions.

Similar to Vazquez et al., the reliability of the protein functional prediction was assessed on incomplete knowledge of the interaction network. To determine the confidence of function prediction, we compared the accuracy of successful predictions of the three methods on various datasets in which a fraction f_n of the classified proteins are assumed to be unclassified. Every unclassified protein is assigned with the function whose weight is the highest in function vector. The success rate is defined as the probability that the function is the actual functional classification for the corresponding proteins, which is a quantitative estimate of the reliability of our predictions. The accuracy of the three methods for different values of f_n using the most stringent functional classification scheme available (424 functional classes from MIPS) is shown in Fig. 2. From Fig. 2, we can see that the accuracies of the MFGO are always higher than those of GOM and MR.

The *leave-one-out* evaluation strategy, in which only one classified protein was denoted as unclassified, is well accepted for its closeness to real situation of application of prediction. On Vazquez dataset, the *leave-one-out* strategy produced the accuracy 56% for MFGO, which is higher than GOM (53%) and MR (50%).

Table 1
The absolute operational time of GOM and MFGO on the same hardware platform

Hardware platform	Time (s)	
	GOM	MFGO
CPU: 2800 MHz Intel Pentium RAM: 2000 MB Operation system: Linux 9.0 Environment: Java Dataset: Vazquez protein–protein network	1500	50

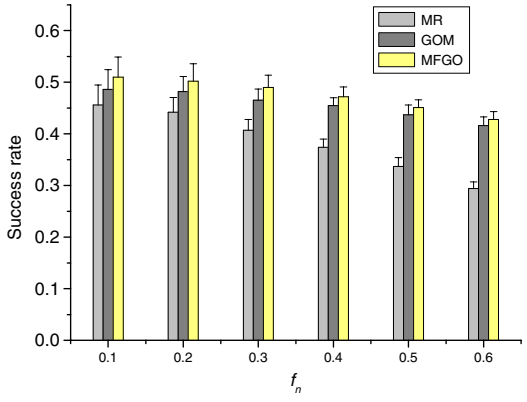

Fig. 2. f_n denotes the fraction of classified proteins assumed to be unclassified. The success rate is defined as the probability that function which weight is the most is the actual functional classification for the corresponding proteins. The gray, dark gray and yellow lines represent the success rates using MR, GOM and MFGO, respectively.

In order to eliminate the bias stemming from differences in data sources, we repeated our analysis on other three representative protein–protein interaction datasets, i.e., DIP-core, SPK and YP. The results indicate that different datasets brought a little influence on the prediction accuracy of the various methods, but the accuracy of MFGO was always the highest regardless of which data set is used. The detailed results are listed in Table 2 and Fig. 3A–D with $f_n = 0.2$ (see Supplementary materials for details).

For proteins with few neighbors, all the predicting methods perform poorly due to too little information. Since topological information involving non-adjacent proteins with common

Table 2
Protein functional prediction accuracies of the three different methods on four different interaction datasets

Data Set	Method	Degree								
		1	2	3	4	5	6	7	>7	Mean
Vazquez	MR	0.290	0.47	0.571	0.646	0.674	0.697	0.657	0.740	0.433
	GOM	0.327	0.498	0.636	0.671	0.711	0.750	0.674	0.750	0.477
	MFGO	0.350	0.547	0.678	0.679	0.740	0.770	0.682	0.750	0.504
DIP	MR	0.246	0.386	0.503	0.516	0.609	0.573	0.648	0.737	0.486
	GOM	0.299	0.420	0.524	0.556	0.654	0.604	0.683	0.744	0.519
	MFGO	0.343	0.430	0.553	0.556	0.652	0.625	0.683	0.759	0.540
YP	MR	0.307	0.360	0.494	0.466	0.456	0.536	0.566	0.618	0.480
	GOM	0.393	0.394	0.534	0.489	0.504	0.586	0.593	0.629	0.518
	MFGO	0.438	0.466	0.585	0.527	0.542	0.600	0.600	0.647	0.554
SPK	MR	0.236	0.30	0.386	0.444	0.324	0.479	0.523	0.521	0.373
	GOM	0.29	0.344	0.434	0.512	0.386	0.533	0.538	0.536	0.419
	MFGO	0.32	0.386	0.454	0.34	0.414	0.544	0.555	0.543	0.439

To compute the success rate, a fraction, f_n, of the classified proteins are assumed to be unclassified and then make functional predictions for them. The table shows the success rates of three methods in four datasets with $f_n = 0.2$. (MR: major rules method, GOM: global optimization method, MFGO: modified and faster global optimization method.)

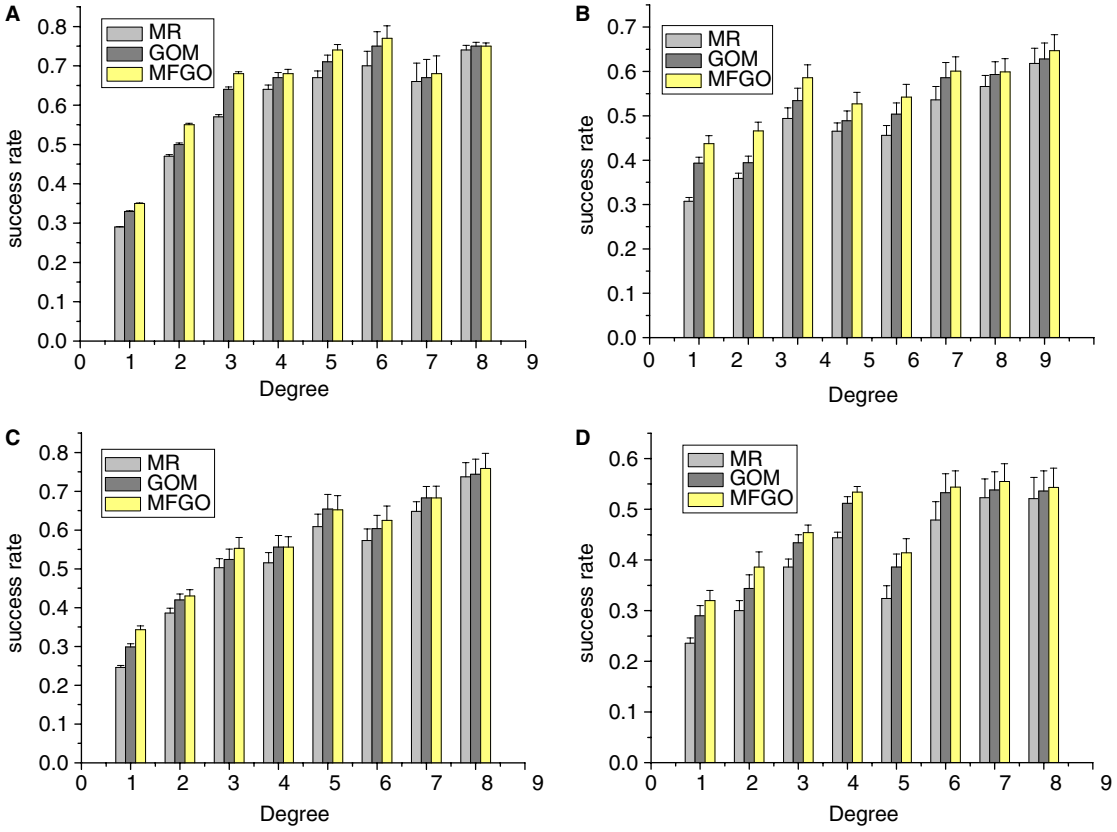

Fig. 3. Protein functional prediction accuracies of the three different methods on four different interaction datasets. To compute the success rate, a fraction, f_n, of the classified proteins are assumed to be unclassified and then make functional predictions for them. The X-axis represents the number of protein neighbors and the Y-axis represents the success rate. The gray, dark gray and yellow lines represent the success rates of MR, GOM and MFGO, respectively ($f_n = 0.2$): (A) the Vazquez dataset; (B) the YP-dataset; (C) the DIP-core dataset; (D) the SPK dataset.

neighbors are taken into account in MFGO, this strategy brings more information for unknown-function proteins, especially for low-degree ones. Hence, an increase of prediction accuracy is observed more obviously for MFGO for the low-degree proteins. For the unclassified proteins in the Vazquez's dataset with less than 3 interactions, the success rate of MFGO is about 8% higher than that of MR, and 4% higher than that of GOM on average ($f_n = 0.2$).

3.3. Robustness validation

For the function assignment algorithm, the tolerance to false interactions is the key factor of its robustness since experimental identified protein–protein interactions will necessarily contain a certain amount of false positives and negatives. The uncertainty of interaction can be simulated through reducing some randomly selected edges and adding the same amount of new edges at random. As a result we get a new network with a dissimilarity degree d_l compared with the original network, where d_l denotes the percentage of different edges between the two networks [13].

We ran MFGO on the modified network, and obtained a function vector for each unclassified protein. For each unclassified protein i ($1 \leqslant i \leqslant n$), $W_i(d_l) = (w_{i1}, w_{i2}, \ldots, w_{im})$ denotes the function vector of protein i satisfying $\sum_{s=1}^{m} w_{is}^2 = 1$, and $W_{is}(d_l)$ denotes the weight of the unclassified protein i belonging to the functional class s in a network with dissimilarity degree d_l (compared to the original network). Hence, $W_i(0)$ corresponds to the functional vector obtained from the original network. $W_{is} = 0$ means that the protein i has not been assigned the function s. A quantitative comparison of predictions through the two networks is described by the overlap function $\Theta_i(d_l)$ defined as follows:

$$\Theta_i(d_l) = \sum_s [W_{is}(0) W_{is}(d_l)].$$

The overlap function equals 1 when $W_{is}(d_l) = W_{is}(0)$ for all s.

We computed the average of $\Theta_i(d_l)$ over all unclassified proteins with different node degree, and observed that it varied little with the node degree. The average of $\Theta_i(d_l)$ of the predicted results between the modified and the original network are shown in Fig. 4 for different value of d_l. From Fig. 4, we can see that the overlap in predicted results decreases when d_l increase. And the overlap is about 90% when $d_l = 20\%$, and is still larger than 80% when $d_l = 50\%$. This suggests that even if more than 50% of proteins having at least one false interaction due to erroneous experimental results, assignment of protein function can still be effective. Because each displaced edge corresponds to three to four proteins with different interactions, this indicates that our approach can tolerate a considerable amount of error in the interaction data. It should be noticed that the prediction accuracy decreases when the number of false interaction increases, which means MFGO will work better on a more accurate interaction data.

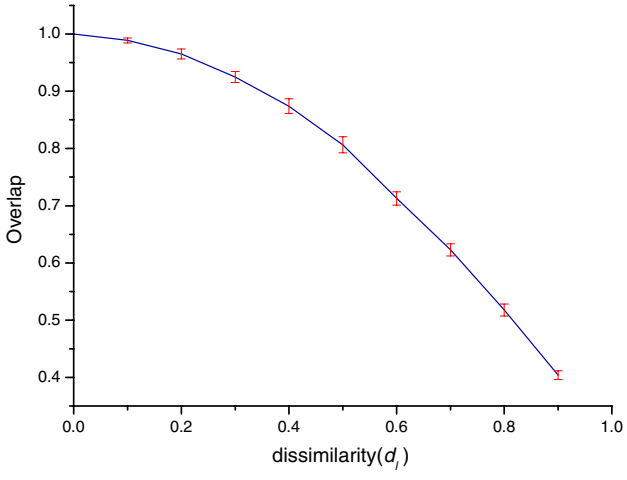

Fig. 4. MFGO's tolerance to interaction data errors. The figure shows the fraction of overlapping predictions $\Theta_i(d_l)$ between the modified and the original network with a degree of dissimilarity d_l, which is the percentage of different edges between the two networks. The analysis shows that a moderate amount of misplaced interactions do not preclude a reliable function assignment.

4. Discussion

Both the speed and the accuracy of our approach are greatly enhanced by the improvements introduced into GOM, implying that MFGO method may be a useful tool for functional prediction based on protein–protein interaction networks. In particular, experimental results show that MFGO has distinctive advantage of predicting function for proteins with few neighbors. Our work also shows a more accurate prediction can be achieved by taking into account of non-adjacent proteins with common neighbors.

The increase in predictions accuracy obtained in the succession from MR through GOM to MFGO shows a promising trend for the prediction of protein function based on protein interaction data. Moreover, we changed the prediction criteria so that one protein might have several functions in MFGO, which expands the area of its applications. Finally, the tests of accuracy and robustness show that our method tolerates a certain number of false positive and negative interactions arising from experimental data, and can also perform well with a certain incompleteness of the protein interaction network. In summary, when compared to previously published methods (MR and GOM), the performance of our algorithm is better in terms of accuracy and robustness, and the functional predictions seem more reliable.

Acknowledgements: This work was supported by National Sciences Foundation of China under Grants 60496320, 30500104 and 30570393, National Key Basic Research & Development Program 973 under Grants 2002CB713805 and 2003CB715900, and opening task of Shanghai Key Laboratory of Intelligent Information Processing Fudan University No. IIPL-04-001. Thanks Dr. Geir Skogerbo and Dr. Deng Minghua for their valuable help.

References

The MIPS comprehensive yeast genome database (CYGD): http://mips.gsf.de/proj/yeast/CYGD/db/

[1] Bader, G.D., Betel, D. and Hogue, C.W. (2003) BIND: the biomolecular interaction network database. Nucleic Acids Res 31, 248–250.

[2] Brun, C., Chevenet, F., Martin, D., Wojcik, J., Guenoche, A. and Jacq, B. (2003) Functional classification of proteins for the prediction of cellular function from a protein–protein interaction network. Genome Biol. 5, R6.

[3] Deane, C.M., Salwinski, L., Xenarios, I. and Eisenberg, D. (2002) Protein interactions: two methods for assessment of the reliability of high throughput observations. Mol. Cell. Proteom. 1, 349–356.

[4] Deng, M., Zhang, K., Mehta, S., Chen, T. and Sun, F. (2003) Prediction of protein function using protein–protein interaction data. J. Comput. Biol. 10, 947–960.
[5] Espadaler, J., Aragues, R., Eswar, N., Marti-Renom, M.A., Querol, E., Aviles, F.X., Sali, A. and Oliva, B. (2005) Detecting remotely related proteins by their interactions and sequence similarity. Proc. Natl. Acad. Sci. USA 102, 7151–7156.
[6] Gavin, A.C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.M., Cruciat, C.M., Remor, M., Hofert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M.A., Copley, R.R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G. and Superti-Furga, G. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415, 141–147.
[7] Hishigaki, H., Nakai, K., Ono, T., Tanigami, A. and Takagi, T. (2001) Assessment of prediction accuracy of protein function from protein–protein interaction data. Yeast 18, 523–531.
[8] Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S.L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., Yang, L., Wolting, C., Donaldson, I., Schandorff, S., Shewnarane, J., Vo, M., Taggart, J., Goudreault, M., Muskat, B., Alfarano, C., Dewar, D., Lin, Z., Michalickova, K., Willems, A.R., Sassi, H., Nielsen, P.A., Rasmussen, K.J., Andersen, J.R., Johansen, L.E., Hansen, L.H., Jespersen, H., Podtelejnikov, A., Nielsen, E., Crawford, J., Poulsen, V., Sorensen, B.D., Matthiesen, D., Hendrickson, R.C., Gleeson, F., Pawson, T., Moran, M.F., Durocher, D., Mann, M., Hogue, C.W., Figeys, D. and Tyers, M. (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. Nature 415, 180–183.
[9] Okada, K., Kanaya, S. and Asai, K. (2005) Accurate extraction of functional associations between proteins based on common interaction partners and common domains. Bioinformatics 21, 2043–2048.
[10] Pagel, P., Kovac, S., Oesterheld, M., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., Montrone, C., Mark, P., Stumpflen, V., Mewes, H.W., Ruepp, A. and Frishman, D. (2004) The MIPS mammalian protein–protein interaction database. Bioinformatics.
[11] Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D. and Yeates, T.O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc. Natl. Acad. Sci. USA 96, 4285–4288.
[12] Schwikowski, B., Uetz, P. and Fields, S. (2000) A network of protein–protein interactions in yeast. Nat. Biotechnol. 18, 1257–1261.
[13] Vazquez, A., Flammini, A., Maritan, A. and Vespignani, A. (2003) Global protein function prediction from protein–protein interaction networks. Nat. Biotechnol. 21, 697–700.
[14] von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S. and Bork, P. (2002) Comparative assessment of large-scale data sets of protein–protein interactions. Nature 417, 399–403.