Academy of Scientific Research and Technology
27th National Radio Science Conference
Faculty of Electronic Engineering, Menoufia Univ., Menouf, Egypt
16-18 March 2010

Estimation Of The Correlation Between Protein Sub-Function Categories Based On Overlapping Proteins

Khaled S. Ahmed[1,2], Nahed H. Solouma[3], Yasser M. Kadah[2]
[1]Department of Biomedical Engineering, MTI University for Technology and Science, Katamia , Egypt Khaled.sayed@k-space.org
[2] Department of Biomedical Engineering, Cairo University, Giza, Egypt ymk@k-space.org
[3] *NILES, Cairo University* nsolouma@niles.edu.eg

Abstract

Unknown protein function prediction is a hot and recent area of research. Protein functions may be predicted from sequences, gene expression, protein domains, protein localizations, protein structure and protein-protein interactions. In most cases, the researchers are interested in estimating the individual protein functions but not the relations between these functions. Obtaining some information about the relations between different functions is of great importance, since this would increase the certainty of protein function prediction. In this paper, we provide a new method for protein function prediction based on the relations between the functions of other interacted proteins. We tried to estimate a value representing the relation between each function and the other functions within the same category, depending on the number of overlapping proteins. The proposed method was applied to the yeast proteome and the results were promising. To increase the degree of certainty, the proposed method could be integrated with the PPI-based function prediction methods. Both integrated algorithms of protein-protein interactions and function relations are used to increase the accuracy of protein function prediction.

1. Introduction

At a cellular level, networking is what we are all about, genes generate proteins but proteins work in complex instead, they bind to each other and interact. Complete genome sequencing has enabled scientists to generate lists of proteins that sequenced organisms are likely to make. They have begun isolating protein complexes and mapping protein-protein interactions by experimental methods to understand how cells function [1], [2]. A lot of methods have been developed to predict protein functions based on different information sources as protein structure, sequences, protein domain, protein-protein interactions, genetic interactions and gene expression analysis [3]-[11]. The accuracy of prediction can be enhanced by integrating multiple sources of information [12] or collecting relations between the known functions. As known protein may have a lot of functions and some functions may have correlations among them. If protein has certain function, it should have another one or should not have others. In spite of protein function prediction is one of the most important problems in the proteomics and estimating the probability of protein to have certain function is very difficult task, Most of the prediction methods do not take the correlation or relations between the functions into

considerations. Most computational techniques are used to predict the protein functions from protein-protein interaction networks. If protein has common function and has interacted with another one the second protein has high probability to have this function. In this study we try to explore some relations between the protein functions to improve the protein function prediction process. Regarding to the fact, the interacted proteins have common function (major function) (Brown et al. 2000; Eisen et al. 1998; Pavlidis et al. 2001). So the relations between the protein functions in the same function category are created according to the overlapping proteins in the known functions. And these relations enhance the prediction process and introduce promised results. This technique is integrated with both protein function prediction methods, neighborhood method and chi-square technique and the results were better than without integration and the accuracy is increased compared with absolute techniques.

2. Methodology

Proteins can be acted as network. Protein is as a node and interaction is as edge of the network. Each protein may have more than one function. And can be seed (self dependent), temporary participate in certain function and in-complex. Also it may have more interactions or seed. Modern techniques try to explore the protein-protein interactions network to predict the un-known protein functions. The used technique tries to explore the correlations or relationships between the proteins functions in yeast and estimate significance values for every relationship among the proteins. These estimated values (functions relation weights) can be integrated with the computational methods of protein function prediction to enhance the prediction results. Yeast protein functions can be divided into three categories Cellular role functions (C.R) (contains 43 sub-function category), Cell location functions (C.L) (contains 29 sub-function category) and Bio-chemical functions (Bio-ch) (contains 57 sub-function category) as shown in Table-1. Yeast proteins defined in the Yeast Proteome Database. (YP Database) http://www.incyte.com/sequence/proteome/databases/YPD. shtml).

Table1. Yeast sub-function categories, function name and the number of proteins for each function.

ID	Function category	Function name	#(p)	Function category	Function name	#(p)	Function category	Function name	#(p)
1	C.R	Aging	39	C.L	Bud neck	61	Bio-ch	ATPase	247
2	C.R	Amino-acid metabolism	218	C.L	Cell ends	6	Bio-ch	ATP-binding cassette	31
3	C.R	Carbohydrate metabolism	254	C.L	Cell wall	70	Bio-ch	Activator	46
4	C.R	Cell adhesion	4	C.L	Centrosome/spindle pole body	72	Bio-ch	Active "transporter," primary	93
5	C.R	Cell cycle control	213	C.L	Contractile ring	3	Bio-ch	Active "transporter," secondary	201
6	C.R	Cell polarity	216	C.L	Cytoplasmic	755	Bio-ch	Adhesin/agglutinin	7
7	C.R	Cell stress	331	C.L	Cytoskeletal	107	Bio-ch	Anchor Protein	13
8	C.R	Cell structure	120	C.L	Endoplasmic reticulum	225	Bio-ch	Channel [passive transporter]	15
9	C.R	Cell wall maintenance	184	C.L	Endosome/Endosomal vesicles	36	Bio-ch	Chaperones	90
10	C.R	Chromatin/chromosome structure	274	C.L	Extracellular (excluding cell wall)	34	Bio-ch	Complex assembly protein	76
11	C.R	Cytokinesis	40	C.L	Golgi	93	Bio-ch	Conserved ATPase domain	23
12	C.R	DNA repair	154	C.L	Lipid particles	14	Bio-ch	Cyclin	23
13	C.R	DNA synthesis	105	C.L	Lysosome/vacuole	94	Bio-ch	DNA polymerase or subunit	24
14	C.R	Differentiation	104	C.L	Microsomal fraction	19	Bio-ch	DNA-binding protein	283
15	C.R	Energy generation	290	C.L	Mitochondrial	479	Bio-ch	Docking protein	30
16	C.R	Lipid, fatty-acid and sterol metabolism	206	C.L	Mitochondrial inner membrane	155	Bio-ch	GTPase activating protein	26

The protein function prediction methods have more accuracy by knowing the annotated proteins, and major functions among the sub-category. The study determines the overlapped number of proteins between each two sub-function as shown in Table 2. Each row indicates the sub-function category contains number of proteins. As example the left top cell indicates the first function (Aging) which has 39 proteins and the second function (Amino-acid metabolism) contains 218 proteins and just one protein which has the two functions in the same time (the cross section cell). Each cross section cell between the two red cells shows the overlapped number of proteins between those two sub- function categories.

If n1, n2 are the number of proteins that have sub-functions category A and B respectively. If n2 less than n1 and n2 (the proteins of the second category) are proteins in both sub functions category A and B, so we say that there is a relation between sub- function category A and sub- function category B. After collecting these proteins as shown in Table 2, we can estimate the relations between the protein function sub-categories. These relations among the sub-functions category can be divided into direct and indirect relations.

Table 2. The Cellular role function categories, number of proteins in each function and the overlapping proteins between each two sub-function categories. The red cells (diagonal) shows the number of proteins in each sub-function and the green cell (3) shows the cross section between sub-function 4 and sub-function 17.

Cellular Role sub-categories

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
39	1	4	0	6	0	6	0	3	13	0	4	0	6	0	1	1
0	218	4	0	7	5	19	0	1	4	0	2	0	3	4	2	4
0	0	254	1	5	6	36	2	10	2	0	0	0	4	29	7	3
0	0	0	4	0	1	1	0	0	0	0	0	0	0	0	0	3
0	0	0	0	213	12	25	10	9	35	3	22	13	7	13	0	14
0	0	0	0	0	216	16	26	23	10	12	3	2	9	3	10	23
0	0	0	0	0	0	331	11	29	12	1	5	0	17	24	8	17
0	0	0	0	0	0	0	120	5	6	4	1	0	0	6	0	8
0	0	0	0	0	0	0	0	184	2	5	3	3	10	3	12	12
0	0	0	0	0	0	0	0	0	274	1	44	32	6	2	1	11
0	0	0	0	0	0	0	0	0	0	40	1	2	2	0	2	2
0	0	0	0	0	0	0	0	0	0	0	154	30	1	1	0	3
0	0	0	0	0	0	0	0	0	0	0	0	105	2	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	104	4	3	11
0	0	0	0	0	0	0	0	0	0	0	0	0	0	290	6	3
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	206	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	132

3. Direct relationships

Method collects all sub-function categories on the two axes as shown in Table 2. And puts the number of overlapped proteins in each cross section cell (square) and compares this number (cell) with the smaller number of the two surrounding sub-categories (red cells). As shown the first top left cell indicates the sub-function category number one and contains 39 as number that means the first sub-function category contains 39 proteins. And the rest cells in the first row indicate the overlapping number of proteins between the first sub-function and residuals of the same sub-functions category according to the column number. Percentage between each cell

number and the smaller number of the two surrounding sub-function categories will be calculated, by determining threshold equal to 0.7 direct relationships between the two sub-function categories can be estimated. As illustrated in Table-3; the method can determine 4 direct relationships between 43 functions in cellular role sub-function categories. It can be noted that the threshold value is big number to express about the correlation between the two sub-function categories.

Table 3. The direct relationships between sub-Cellular role categories

Function category number-1	Function category number-2	Weight
4	17	0.75
40	7	1
19	43	0.725
37	36	0.726

From the previous table, technique indicates that there is a direct relationship between sub-function category 4 (Celladhesion) and sub-function category 17 (Matingresponse) by weight 0.75 that means if protein Pi has sub-function category 4 it may have sub-function category 17 by probability equal to 0.75 or by another words a 75% of proteins that have sub-function category 4, they have sub-function category 17. Also in the third raw of table-3 the weight is equal to the unity that means all the proteins which have sub-function category 40 (Recombination) have sub-function category 7 (Cellstress). This technique converts the undirected graph of physical interactions between the proteins (protein interaction network) into directed graph between the sub-function categories which have been taken into consideration to enhance the accuracy of protein function prediction. As shown in figure-1 the arrow illustrates the direction between the sub-function categories.

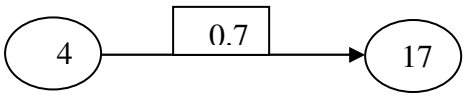

Fig. 1. The directed relation between the two sub-function categories 4, 17 and its weight equal (0.75).

As shown in table-3, there are found just 4 relations their values more than the determined threshold (0.7) which is high enough to grantee strong relation. Because the directed relationship cannot give a wide screen for the relations between the sub-function categories, so our study has studied the indirect relationships and anti-correlations between the proteins sub-functions category.

4. Indirect relationships

Because the few number of direct relationships, the study puts some conditions to estimate the indirect relationships or uncorrelated functions. If there are three sub - functions category A, B and C and each function contains number of proteins X1, X2 and X3 respectively and

If $A \cap B = n1$ proteins,
$A \cap C = n2$ proteins and
$B \cap C = n3$ proteins

The next combinations can be collected

- $n1 = 0$ and / or $n2=0$
- $n1 = n2$
  - a- ($n1=n2=n3$)
  - b- ($n1=n2$ and $n3=0$)
  - c- ($n1=n2 \neq n3$)
- $n1 \neq n2$

4.1 [n1= 0 and/or n2=0]

If the number of proteins in the cross section between two sub-function category is zero (no overlapped proteins are found) that leads to *uncertainty* case**.** We cannot say that there is anti correlation between these two sub-functions category which have intersection by zero. So, we should calculate the indirect relationship between two or more function categories if they interact in the same number of proteins for the same function category.

4.2. [n1=n2]

4.2.1 [n1= n2 = n3]

the same proteins found in the three sub-function categories *($A \cap B \cap C = n1 = n2 = n3$)* that leads to there is a *correlation* between B, C toward A and so on. If number m of proteins have functions B and C, they should have function A. as shown in figure-2 if protein has the two sub-functions category it should have the third one or by the statistical view p (A\B,C) = 1 the probability of protein to have sub-function category A conditional sub-function categories B and C equal the unity as shown in figure-2 if protein has A (first function) and conditional B (second function) it should have the third one C.

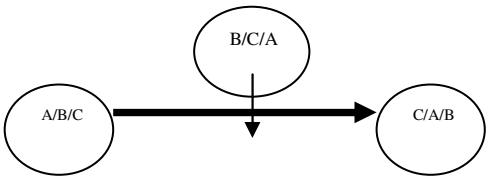

Fig. 2. Shows the conditional relationship between the sub-function categories.

4.2.2 [n1= n2] but no protein has the two sub-functions category that leads to *anti correlation* between those two sub-function categories. If protein has function A and B it should not have function C as shown in figure-3 or in the statistical view P (B\A, C) = P (C\A, B) = 0 the probability of protein to have the third function conditional the two functions is zero.

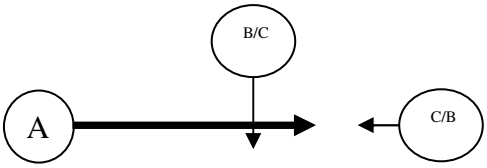

Fig. 3. Shows the anti correlation between the two sub-functions category B, C given sub-function category A.

4.2.3 [n1= n2 ≠ n3] if the protein is in two sub-function categories and is not in the third so it leads to *uncertainty* case.

4.3 [n1< > n2]

If the number of proteins is not the same in the two sub-function categories, it should have three combinations condition n1< n2.

4.3.1 [n1< n2] and there is no intersection between the proteins of the two sub-function categories that leads to *uncertainty* case.

4.3.2 [n1< n2] and some of n1 is found in n2 (some proteins of the first sub-function category are found in the second sub-function category) that also leads to *uncertainty*.

4.3.3 [n1< n2] and all of n1 are found in n2 that leads to function category B is correlated to function C when the protein has function category A as shown in figure-4 example for correlation between A\B towards C or function B is dependent on function A.

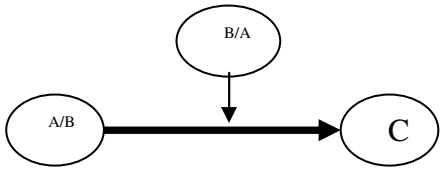

Fig.4. Shows that if protein has function A or function B and given it has another one (B or A) respectively that leads it has function C.

An illustrated example; if protein has function 4 as a basic function category and has direct relationship with function category 17 by weight 0.75 and indirect relationship with function categories 6, 7, 28 respectively the technique can illustrate the relationships as shown in figure-5.

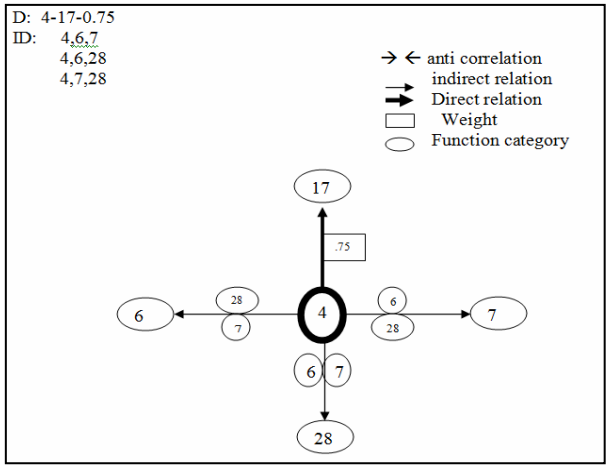

Fig. 5. Shows an example for complete combination between sub-function category 4 and the rest functions of the cellular role category in Yeast where D (directed interaction), ID (indirect interactions) and function 4 is the studied function.

5. Results and Discussion

The function relation technique has integrated with the traditional methods of protein function prediction. And improved results have been collected than previous. As known in neighborhood method the function with high frequency has been taken as first one then the others (less frequency) without taking the relation between the functions into consideration. Now the function has accepted by the highest frequency and bigger number of correlation or relations. For example sub-function category 19 (Membrane fusion) contains 40 proteins (yeast protein

function database) and has directed relation with sub-function category 43 (Vesicular transport) regarding to our technique with weight 0.725 and there are 71 interactions (MIPs). 28 interactions are between two proteins that have the same sub-function (19) and 43 have interactions between one has the function and the other not. After applying the combination between neighbor method and the studied technique, we have found the next results as shown in table 4. It can be noted that the numbers of the false positive and true negative in combination technique are less than in single mode (neighbor method) by addition the number of the true positive is roughly the same (the difference in the values according to the weight value). The studied technique has clear addition on the accuracy of the prediction.

Table 4. Shows the comparison between the neighborhood method and the combination algorithm

	Neighbor method For sub-function 19	Neighbor method + studied technique (19,43)
# True positive	15 proteins	14
# False positive	17 proteins	7
# True negative	25 proteins	7

By applying the Chi-square method to get the correlation between the two sub-function categories 19 and 43 we have found for each prediction the score values are almost the same for the predicted protein

Table 5. Shows the results of applying Chi-square technique on the two sub-function categories and the score values are very near. In some cells a wide range between values are shown, but they are still the highest values in the technique.

Protein ID YPDatabase	Protein name YPDatabase	Score value (19)	Score value (43)
955	ERV25	0.0062	19.08
3069	SEC17	20.9	21.37
3073	SEC22	0.39	32.1
3093	SED5	30	13.06
3240	SNC1	78.21	38.16
3409	SSO1	20	20
3410	SSO2	20	20
3619	TLG2	316.8	88.757
3756	UFE1	+VE	+VE
3805	VAM7	15	15

6. Conclusion

In this study, a novel technique has been introduced to get the relations between the functions in the same function category for yeast. By applying the technique on all the functions categories and mixing the results with any method of protein function predictions as neighborhood and Chi-square methods, an enhanced results and increasing of the accuracy has been achieved.

7. Reference

[1] Paul N. MacDonald "Two-Hybrid Systems Methods and Protocols", Totowa, New Jersey 2001.
[2] ERIC M. PHIZICKY1S Mar "Protein-Protein Interactions: Methods for Detection and Analysis" Vol. 59, No. 1p. 94–123, 1995.
[3] Minghua Deng, Kuizhang, Shipra Mehta, Ting Chen, "*Prediction of Protein Function Using Protein–Protein*
Interaction Data" journal of computational biology Mary Ann Liebert, Vol 10, pp. 947–960, 2003.
[4] A. Vazquez, A. Flammini, A. Maritan, and A. Vespignani. "Global protein function prediction from protein-protein interaction networks". *National Biotechnology*, vol 21 pp 697–700, 2003.

[5] Costanzo, M.C., Crawford, M.E., Hirschman, J.E., Kranz, J.E., Olsen, P., Robertson, L.S.,Skrzypek, M.S., Braun, B.R., Hopkins, K.L., Kondu, P., Lengieza, C., Lew-Smith, J.E., Tillberg, M., Garrels, J.I YPD*TM*, PombePD*TM*, and WormPD*TM*" model organism volumes of the BioKnowledge library, an integrated resource for protein information" . *Nucleic Acids,* vol 29, pp 75- 79,2001.
[6] Gavin, A., B¨*o*che, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A., Cruciat, C. et al. "Functional organization of the yeast proteome by systematic analysis of protein complexes" *Nature, pp*415-141 and 147,2002.
[7] Michele L. and Andrea Pagnani "Predicting protein functions with message passing algorithms" *Bioinformatics* Vol. 21 no. 2, pp 239–247, 2005.
[8] Roded Sharan "Analysis of Biological Networks: Protein-protein Interaction Networks – Functional Annotation"
London, 2006.
[9] B. Schwikowski, P. Uetz, and S. Fields. "A network of protein-protein interactions in yeast. National Biotechnology, vol 18(12), pp 1257–1261, 2000.
[10] H. Hishigaki, K. Nakai, T. Ono, A. Tanigami, and T. Takagi. "Assessment of prediction accuracy of protein function from protein-protein interaction data". Yeast, vol 18(6), pp523–531, 2001.
[11] Aidong Zhang ,*Protein Interaction Networks : Computational Analysis*, Cambridge, New York, 2009
[12] Yin Liu1, Inyoung Kim, Hongyu Zhao "Protein interaction predictions from diverse sources" Drug Discovery Today, Vol13 pp 409-416, May 2008.