Determining The Relations Between Protein Sub-Function Categories Based On Overlapping Proteins

Khaled S. Ahmed, Nahed H. Solouma, Yasser M. Kadah

MTI, College of Engineering, Cairo University, Egypt

Received: September 18, 2010 / Accepted: October 25, 2010 / Published: January

Abstract: Protein functions and their relations are powerful tasks in the field of proteomics since it can lead to knowing cell functions and activities. The relations between the proteins functions do not be considered into protein function/interaction prediction; although they have critical rule in improvement the accuracy of protein functions. In this paper, we propose a technique for determining the relation between the protein functions. The technique is based on the overlapping number of proteins that to determine the correlation between the sub-function categories as well as improve the protein function prediction process. The proposed method was applied to Yeast functions proteome and the results revealed great improvement in increasing the degree of certainty and accuracy for protein function prediction.

Key words: Protein, function category, function relations.

1. Introduction

Although estimating the protein functions correlations is very important, many researchers are interested in determining the individual protein functions not the relations between these functions. Protein functions may be predicted from sequences [1, 2], gene expression [3, 4], protein domains [5, 6], protein localizations [7, 8, 9], and protein-protein interactions [10- 14].

In most cases, obtaining 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 that representing the relation between each function and other functions within the same category, depending on the number of overlapping proteins. The proposed method was applied to 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 [10, 12, 13]. Both integrated algorithms of protein-protein interactions and function relations are used to increase the accuracy of protein function prediction. The paper is classified as follows. The proposed algorithm is explained in section II. Section III presents the results of this work together with their discussion. Finally, the paper ends with a conclusion and future work.

2. Methodology

A proteome network can be described as a complex system of proteins linked by interactions. The computational analysis of these networks begins with the representation of the PPI network structure. The simplest representation takes the form of a network graph consisting of nodes and edges [15]. Proteins are represented as nodes in the graph and two proteins that interact physically are represented as adjacent nodes

Corresponding authors: Yasser Kadah, Ph.D., professor, research fields: Image processing, Signal processing. ymk@k-space.org. Nahed Solouma, Ph.D., associate professor, research fields: laser application. nsolouma@k-space.org. Khaled Sayed, M.Sc. Teaching assistance, research fields: Biomedical, Bioinformatics, Proteomics. Khaled.sayed@k-space.org.

connected by an edge [16]. Based on this graphical representation, various computational approaches, such as data mining, machine learning, and statistical approaches can be performed to reveal the PPI networks at different levels. Each protein may have more than one function. And can be seed (self dependent), temporary participate in certain function or in-complex. Also it may have more interactions or be alone. 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. The proposed algorithm applied to yeast protein functions which 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 (YPDatabase).

The protein function prediction methods have been more accurate by knowing the annotated proteins, and major functions among the sub-function category. The study determines the overlapped number of proteins between each two sub-functions 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 (ATPase) which has 247 proteins and the second function (ATP-binding cassette) contains 31 proteins. It can be noted that all proteins of second category have been found in the first category. So there are 31 proteins have the two function categories 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 numbers 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 found 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-1. A part of yeast sub-functions categories and their numbers of proteins.

Function category	Function name	# proteins
Cellular role	Aging	39
Cellular role	Amino-acid metabolism	218
Cellular role	Carbohydrate metabolism	254
Cellular role	Cell adhesion	4
Cellular role	Cell cycle control	213
Cellular role	Cell polarity	216
Cellular role	Cell stress	331
Cellular role	Cell structure	120
Cellular role	Cell wall maintenance	184
Cellular role	Chromatin/chromosome structure	274
Cellular role	Cyto kinesis	40
Cellular role	DNA repair	154
Cell location	Bud neck	61
Cell location	Cell ends	6
Cell location	Cell wall	70
Cell location	Centrosome/spindle pole body	72
Cell location	Contractile ring	3
Cell location	Cytoplasmic	755
Cell location	Cytoskeletal	107
Cell location	Endoplasmic reticulum	225
Cell location	Endosome/Endosomal vesicles	36
Cell location	Extracellular (excluding cell wall)	34
Biochemical	ATPase	247
Biochemical	ATP-binding cassette	31
Biochemical	Activator	46
Biochemical	Active "transporter," primary	93
Biochemical	Active "transporter," secondary	201
Biochemical	Adhesin/agglutinin	7
Biochemical	Anchor Protein	13
Biochemical	Chaperones	90
Biochemical	Complex assembly protein	76

Table-2. First 26 sub-functions of Yeast Biochemical function. The red cells (diagonal) shows the number of proteins in each sub-function and the green and blue cells show the overlapping cross section for high correlated functions.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
1	247	31	3	66	0	0	0	0	9	3	23	0	0	64	0	0	4	0	4	83	224	5	1	0	2	0
2	0	31	0	23	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	31	0	0	0	0	0
3	0	0	46	0	0	0	0	0	1	1	0	0	0	9	0	0	0	0	2	1	2	4	0	0	1	1
4	0	0	0	93	4	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	66	0	0	0	0	0
5	0	0	0	0	201	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	7	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	13	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
8	0	0	0	0	0	0	0	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	90	1	4	0	0	1	0	0	0	0	28	0	8	3	15	0	0	0
10	0	0	0	0	0	0	0	0	0	76	2	0	0	3	1	0	0	0	0	0	2	2	0	0	1	0
11	0	0	0	0	0	0	0	0	0	0	23	0	0	0	0	0	0	0	0	0	23	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	23	0	0	0	0	0	0	0	0	0	1	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	24	7	0	0	0	0	0	0	5	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	283	0	0	1	0	0	41	70	8	3	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	30	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	26	0	0	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	61	0	0	0	58	1	0	0	0	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	23	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	33	0	4	3	1	0	0	0
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	84	82	2	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	640	3	1	0	5	5
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	69	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	48	0	0	0
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0
25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	98	1
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	90

Table-3 the (9) direct relationships between sub-Bio chemical categories related to the threshold equal (0.85)

Function category number-1	Function category number-2	Weight
1	2	1
1	11	1
1	20	0.99
1	21	0.90
2	21	1
9	19	0.85
11	21	1
17	21	0.95
20	21	0.98

2.1 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) [17]. As shown the first top left cell indicates the sub-function category number one and contains 247 that mean the first sub-function category contains 247 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 0.85 direct relationships between the two sub-function categories can be estimated. As illustrated in Table-3; the method can determine 9 direct relationships (more than the threshold) between 57 functions in Bio-chemical sub-function categories. It can be noted that the threshold value is big number to express the correlation between the two sub-function categories. If the threshold decreases till 0.70 as [17], there are more than 15 direct relations.

The technique indicates that there is a direct relationship between sub-function category 1 (ATPase) and the second function (ATP-binding cassette) with weight equal 100% towards sub-function category one. It means that if protein has function category 2 it should have function category one. In the fourth raw of table-3, the weight is equal to the 0.9 that means 90% of the proteins which have sub-function category 21 (Hydrolase) have sub-function category 1 (ATPase). 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.

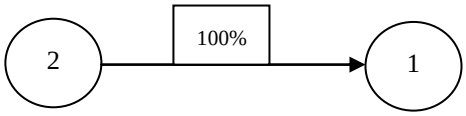

Fig. 1 The directed relation between the two sub-function categories 2, 1 and its weight equal (100%).

Because the directed relationship cannot give a wide screen for the relations between the sub-function categories, so the study has studied the indirect relationships and anti-correlations between the proteins sub-functions category.

2.2 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∩ B = n1 proteins,
A∩ C = n2 proteins and
B∩ 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 ≠n3)
- n1 ≠ n2

2.2.1 [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, it 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.

2.2.2. [n1=n2]

2.2.2.1 [n1= n2 = n3]

The same proteins found in the three sub-function categories *(A∩B∩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.

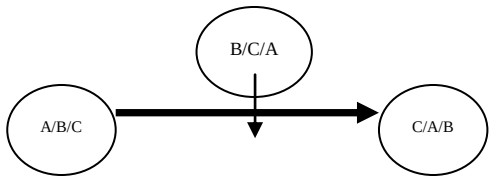

Fig. 2 shows the conditional relationship between the sub-function categories.

2.2.2.2 [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.

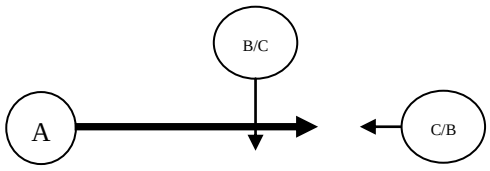

Fig. 3. the anti correlation between the two sub-functions category B, C given sub-function category A.

2.2.2.3 [n1= n2 ≠ n3]

If the protein is in two sub-function categories and is not in the third so it leads to *uncertainty* case.

2.2.3 [n1< > n2]

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

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

2.2.3.2 [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*.

2.2.3.3 [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.

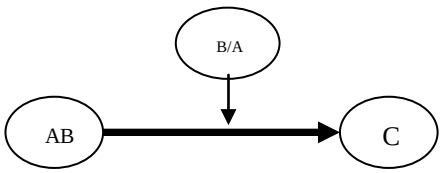

Fig.4. Shows that if protein has function A/B and has conditional function B/A respectively it will have function C.

3. 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 counting 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 1 (ATPase) contains 247 proteins (yeast protein function database) and has directed relation with sub-function category 2 (ATP-binding cassette) regarding to our technique with weight 100%. After applying the combination between neighbor counting method and the studied technique, it has been found that the results are 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) in addition the number of the true positive is roughly the same or less (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 1	Neighbor method + function relation (1,2)
# True positive	47 proteins	2
# False positive	141 proteins	15
# True negative	200 proteins	29

By applying the Chi-square method to get the correlation between the two sub-function categories 1 and 2, it has been found that for each prediction, the score values of the two functions are the highest scores.

4. 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. As future work, the function correlations can be enhanced by integrating the overlapping number of proteins with the number of clusters interactions.

5. References

[1] Harrington E, Singh A: **Quantitative assessment of protein function prediction from meta genomics shotgun sequences.** *PNAS* 2007, 104(35):13913–13918.

[2] Spriggs R, Murakami Y: **Protein function annotation from sequence: prediction of residues interacting with RNA.** *Bioinformatics* 2009, 25 (12) 1492–1497.

[3] Zhao XM, Wang Y, Chen L, Aihara K: **Gene function prediction using labeled and unlabeled data**. *BMC Bioinformatics* 2008, 9:57-71.

[4] Zhao H, Wu B: **DNA-Protein Binding and gene expression patterns**. *Lecture Notes-Monograph Series*, Statistics and Science: A Festschrift for Terry Speed 2003, 40: 259-274.

[5] Friedberg I : **Automated protein function prediction---the genomic challenge.** *Bioinformatics* 2006, 7(3) 225-242.

[6] Nariai N, Kolaczyk E : **Probabilistic Protein Function Prediction from Heterogeneous Genome-Wide Data.** *Plos one* **2007.**

[7] Morin M: **Phylogenetic Networks Simulation, Characterization, and Reconstruction**" New Mexico, 2007.

[8] Sun J, Zhao Z: **Construction of phylogenetic profiles based on the genetic distance of hundreds of genomes**. *Biochem Biophys Res Commun* 2007, **355**(3):849-853.

[9] Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO: **Assigning protein functions by comparative genome analysis: protein phylogenetic profiles**. *Proc Natl Acad Sci U S A* 1999, **96**(8):4285-4288.

[10] Schwikowski B, Uetz P, Fields S: **A network of protein-protein interactions in yeast**. *Nat Biotechnol* 2000, **18**(12):1257-1261.

[11] Sharan R: **Analysis of biological networks: Protein-protein interaction networks – functional Annotation**, lecture note *2006*.

[12] Hishigaki H, Nakai K, Ono T, Tanigami A, Takagi T: **Assessment of prediction accuracy of protein function from protein--protein interaction data**. *Yeast* 2001, **18**(6):523-531.

[13] Deng M, Zhang K, Mehta S, Chen T, Sun F: **Prediction of protein function using protein-protein interaction data**. *J Comput Biol* 2003, **10**(6):947-960.

[14] Khaled, S, Nahed, S, and Yasser, K.: **Comparison between different methods for protein function prediction**. NRC, 2009.

[15] Wagner A: **How the global structure of protein interaction networks evolves**. *Proc Biol Sci* 2003, **270**(1514):457-466.

[16] Aidong Z : *PROTEIN INTERACTION NETWORKS: Computational Analysis*. New York, Press 2009.

[17] Khaled, S., Nahed, S., and Yasser, K. :**Estimation of the correlation between protein sub-function categories based on overlapping proteins.** NRSC, 2010.