Estimating Protein Functions Correlation Based on Overlapping Proteins and Cluster Interactions

Khaled S. Ahmed

Biomedical Department, Modern University for Technology and Information, Cairo, Egypt

Abstract Usually, the relations between the protein functions do not be considered into protein function/interaction prediction processes. In this paper, we present a new technique for determining the relation between the protein functions. The strategy is based on the overlapping number of proteins furthermore interactions over protein clusters 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 proteome and the results revealed great improvement in increasing the degree of certainty and accuracy for protein function prediction.

Keywords Protein Function, Correlation, Cluster Interaction

1. Introduction

Protein is a fundamental component of all living cells, it consists of sequences of Amino Acids (AAs) and performs a variety of biological tasks as Control physicochemical conditions inside the cell or transmit biological signals. Usually proteins can bind to each other and interact so they work in complex system. Protein complex isolation and mapping protein-protein interactions is considered one of the most important problems in proteomics. The target of these processes is to understand the cell functions and to have basic idea about the relations between the proteins functions. A lot of methods have been developed to predict protein functions based on different information sources as protein sequences[1,2], protein structure[3,4], protein-protein interactions[5,6,7], protein domain [8], genetic interactions and gene expression analysis[9,10]. The accuracy of prediction can be enhanced by integrating multiple sources of information[11,12] or collecting relations between the known functions[13]. Also a new technique is used to enhance the protein function prediction process depending on the weights of interactions[14]. Recently, the researchers introduced different methods to determine the probability of protein function prediction using the information extracted from PPI. Although these techniques are promising, they lack the addressing of effective problems such as determining the relations between the protein functions. In this paper, we introduce an integrated algorithm based on the overlapping proteins[12,15] and interactions among different protein clusters. The interactions are regarding to the fact, the interacted proteins have common function (major function) (Brown et al. 2000; Eisen et al. 1998; Pavlidis et al. 2001). As known, protein may have more than one function (up to 8 functions in Yeast Saccharomyces cerevisiae). Some of these functions may be correlated, anti correlated or independent. In this paper, an integrated technique will be introduced to determine the relations between protein functions. The technique is applied to the three function categories of yeast and integrated with protein function prediction method; neighborhood counting method and the results were better than without integration and the accuracy is increased compared to absolute techniques. The paper is organized 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

Protein may be seed (self dependent) or participate in certain function or in-complex (temporary or permanent). If protein has certain function F1 and it has another function as F2 but it should not have function F3, so it can be said that functions (F1, F2) has specific relations and functions (F1, F3) are anti correlated. The proposed technique is to explore the relation between the protein functions based on the overlapping proteins and interactions over the protein clusters.

A. Function Categories And Overlapping Proteins

As mentioned before, protein may have one function or more. Regarding the yeast (Saccharomyces cerevisiae) studied species, it has three function categories: Cell location functions (C.L) (contains 29 sub-function category), Cellu-

lar role functions (C.R) (contains 43 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. Each function category has certain number of proteins. And some of those proteins are involved in more than one sub-function category. Herein the overlapping number of proteins is calculated for each two different sub-function categories. And certain score is recorded related to the smaller number of proteins for the two categories. As shown in Table-1, the function category cellular role contains some sub-functions (up to 43 sub-functions) as: Amino-acid metabolism contains 218 proteins that means there are 218 proteins among yeast proteins have this function. As similar Carbohydrate metabolism function, there are 254 proteins having this function. Also the other function categories, cell location "Bud neck" contains 61 proteins and Biochemical "ATPase" contains 247 proteins. The target is to calculate the overlapping proteins between each two sub-function categories. The overlapping proteins mean that these proteins have the two sub-function categories. As mentioned in[12], the relations between the sub-functions are divided into direct and indirect relations depending on the suggested threshold value. The proposed technique tries to determine certain score for each sub-function pair. And integrate these scores with the extracted values of protein clusters interactions.

Table 1. yeast sub-function categories, function name and number of proteins for each function.

Function category	Function name	# proteins
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	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	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	Anchor Protein	13
Biochemical	Chaperones	90
Biochemical	Complex assembly protein	76

By applying the proposed technique on the yeast function category (Biochemical), it has found a lot of direct relations between the sub-function categories as shown in Table-2 and Table-3. Method collects all sub-function categories on the two axes as shown in Table-3 and puts the number of overlapped proteins in each cross section cell (square) then 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 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-2; the method can determine 9 direct relationships among 57 functions in biochemical sub-function categories. It can be noted that if the threshold value is decreased to 0.72, the direct relations between the sub-function categories will increase.

Table 2. the direct relations over biochemical sub-function categories when threshold greater than 0.85.

Fx_1 ID	Fx_2 ID	Fx_1 name	Fx_2 name	Score
1	2	ATPase	ATP-binding	1
1	11	ATPase	Conserved ATP	1
1	20	ATPase	Helicase	0.99
1	21	ATPase	Hydrolase	0.91
2	21	ATP-binding	Hydrolase	1
9	19	Chaperones	Heat shock protein	0.85
11	21	Conserved ATP	Hydrolase	1
17	21	GTP-binding protein/GTPase	Hydrolase	0.95
20	21	Helicase	Hydrolase	0.98

The direct relations between the functions mean correlation between those functions. For example if protein has function x, it should have y because there is high correlations between function x and function y. as shown in Table-2, there are 9 direct relations (green cells in Table-3). The score between four of them is 1 which means all proteins have the first function they have the second function. As shown protein has sub-function category_2 (ATP-binding) will have by default sub-function category_1 (ATPase) as shown in Figure-1 and each protein has sub-function category_11 (conserved ATP) will have sub-function category_21 (Hydrolase). If the threshold decreased into 0.72, a lot of direct relations can be created (blue cells) as relation between sub-function category_1 (ATPase) and sub-function category_4 (transporter) which has scored 0.72. These scores are collected and will be integrated with scores of cluster interactions.
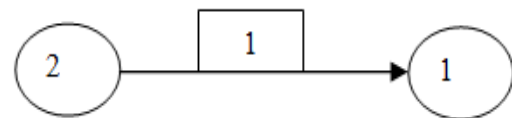

Figure 1. The *direct relations* between the Biochemical sub-function category_2 towards the sub-function category_1.

Table 3. biochemical function categories (first 26 sub-functions), number of proteins for each function and the overlapping proteins between each two sub-function categories. the red cells (diagonal) show number of proteins in each sub-function and the green cells show the overlapping cross section for 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

B. Protein Cluster Interaction

Proteins can be acted as network. The simplest representation takes the form of a network graph consisting of nodes and edges. Proteins are represented as nodes in the graph and two proteins that interact physically are represented as adjacent nodes connected by an edge. Each group of proteins doing certain functions called cluster (may have sequence similarity or not). So the network consists of groups of clusters. The clusters may be self assembled or have external interactions. The interactions may be from real interactions (physical interactions between proteins in the two different groups or clusters) or from overlapping proteins (same proteins are found in the two clusters and have self interactions).
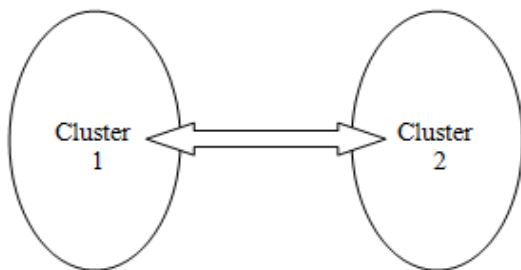

Figure 2. Shows two interacted clusters

As shown in figure-2, two clusters can interact and these interactions are bidirectional. Table-4 shows the cross section numbers of cluster interactions of Yeast Biochemical function categories. Each cell indicates the interactions number of the two indicating functions or clusters. For example proteins sub-function category_1 (cluster-1) interacts with proteins sub-function category_14 (cluster-14) by 49 interactions. The number of interactions is small comparing to the number of overlapping proteins (64) which exciting in the two clusters (1, 14). Although the interaction number is small, they have included 17 self interactions and the rest is correct interactions. Also there are some clusters have no interactions with any one (self interactions or external) as clusters (24, 6). The cause is, these clusters have group of proteins does not have the ability to interact with others or because the required function needs only one protein. The threshold that determines the strength of correlation for cluster interactions is very difficult to specify. The threshold can be estimated as certain number (specific) or as percentage of the number of proteins. The proposed technique suggests that the threshold is to be more than 10% of the number of proteins found in one of the two clusters.

C. Overlapping and Interaction Integration

In this paper, the scores of overlapping and cluster interactions will be integrated to determine the relation between the functions either positive (to participate in the same functions) or negative (anti correlations, if protein has one function, it should not have the other one) or independent (there is no relations between the studied functions). Herein, if the score of overlapping proteins is more than the threshold (0.85), it will be positive otherwise will be negative. Also for the cluster interactions that have more than 10% will be positive and other wise will be negative.

Table 4. biochemical function categories (first 26 sub-functions yellow colored), number of proteins for each functionsand the number of interactions between each two sub-function categories. the red cells show the number of proteins in each sub-function and the green cells show the higher number of interactions (cross section) for correlated functions.

| | | 247 | 31 | 46 | 93 | 201 | 7 | 13 | 15 | 90 | 76 | 23 | 23 | 24 | 283 | 30 | 26 | 61 | 23 | 33 | 84 | 640 | 69 | 48 | 6 | 98 | 90 |
		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
247	1	33	3	8	9	1	0	0	0	11	6	11	1	2	49	1	0	5	3	4	7	53	2	2	0	3	3
31	2	3	1	1	2	0	0	0	0	0	0	0	0	0	1	0	0	2	0	0	1	5	1	0	0	0	0
46	3	8	1	6	0	1	0	1	0	1	4	2	1	0	8	7	1	3	1	1	3	11	11	0	0	0	0
93	4	9	2	0	6	0	0	0	0	0	6	0	0	0	1	0	0	0	0	0	1	9	2	0	0	0	0
201	5	1	0	1	0	0	0	0	0	0	0	0	0	0	3	0	1	1	1	0	0	2	2	0	0	0	0
7	6	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	0
13	7	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
15	8	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
90	9	11	0	1	0	0	0	1	0	26	1	2	4	0	0	0	0	2	5	19	0	15	4	2	0	0	0
76	10	6	0	4	6	0	0	0	0	1	7	0	1	0	9	1	1	1	0	0	1	7	3	0	0	0	0
23	11	11	0	2	0	0	0	0	0	2	0	5	0	0	3	0	0	0	0	0	0	10	0	0	0	0	0
23	12	1	0	1	0	0	0	0	0	4	1	0	0	0	1	0	0	0	0	3	0	0	4	1	0	0	0
24	13	2	0	0	0	0	0	0	0	0	0	0	0	4	3	0	0	0	0	0	0	1	0	0	0	0	0
283	14	49	1	8	1	3	0	0	1	0	9	3	1	3	60	0	0	2	0	0	5	36	9	1	0	3	3
30	15	1	0	7	0	0	0	0	0	0	1	0	0	0	0	16	2	3	0	0	0	4	0	0	0	0	0
26	16	0	0	1	0	1	0	0	0	0	1	0	0	0	0	2	1	3	0	0	0	3	1	0	0	0	1
61	17	5	2	3	0	1	0	0	0	2	1	0	0	0	2	3	3	5	17	2	2	15	4	2	0	0	0
23	18	3	0	1	0	1	0	0	0	5	0	0	0	0	0	0	0	17	3	5	1	21	4	0	0	0	0
33	19	4	0	1	0	0	0	0	0	19	0	0	3	0	0	0	0	2	5	6	0	5	2	1	0	0	0
84	20	7	1	3	1	0	0	0	0	0	1	0	0	0	5	0	0	2	1	0	2	12	0	1	0	0	0
640	21	53	5	11	9	2	0	1	0	15	7	10	0	1	36	4	3	15	21	5	12	79	9	8	0	2	0
69	22	2	1	11	2	2	0	0	0	4	3	0	4	0	9	0	1	4	4	2	0	9	6	1	0	0	1
48	23	2	0	0	0	0	0	0	0	2	0	0	1	0	1	0	0	2	0	1	1	8	1	1	0	1	0
6	24	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	0
98	25	3	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	2	0	1	0	1	0
90	26	3	0	0	0	0	0	0	0	0	0	0	0	0	3	0	1	0	0	0	0	0	1	0	0	0	5

Table 5. yeast biochemical functions, estimated numbers of proteins as true positive (TP), true negative (TN), false positive (FP).

Function category	TP	TN	FP
1	47	200	141
2	2	29	15
4	9	84	18
9	25	65	29
11	6	17	10
14	67	216	194
18	4	19	36
19	7	26	16
20	4	80	45
21	91	549	395

3. Results

The function relation technique has integrated with the traditional method of protein function prediction (neighbor counting method). Improved results have been gained than previous. As known in neighborhood method, it finds the neighbor proteins and gets their assigned functions and the frequencies of occurrence of these functions. Then, these functions are arranged in descending order according to their frequencies. The first k functions are considered and assigned to the un-annotated protein. The authors in[18] used this technique with k equals to 3. By applying the proposed technique on the yeast function categories, the results are as shown in Table-5 and Table-6. The algorithm shows the increasing number of true positive (TP) and de-

creasing the true negative (TN) and false positive (FP). Table-5 shows each yeast Biochemical function category and its results. Function category_1 has 247 proteins, 47 of them identified as TP and the rest (200) identified as TN and there are 141 proteins identified as FP. On the other hand function category_2 has 2 proteins as (TP), 29 proteins as (TN) and 15 proteins as (FP). Also function category_11 has 6 proteins as (TP), 17 proteins as (TN) and 10 proteins as (FP). It can be noted that the integrated algorithm enhanced (increased) the numbers of TP and decreased the numbers of TN and FP. As shown in Table-6 the integration between function_1 and Function_2 (positive overlapping and positive interactions) shows the same numbers of function_2 (least one). And integration between function_1 and function_11 has 6 proteins as (TP, the same number of function_11 true positive) and decreases the number of FP (141 & 10 → 7). The integration process has been divided into for cases regarding to the states of overlapping and interactions. The collected cases are 1)- Positive overlapping & positive interactions (the score of overlapping more than the threshold (0.85) and the number of interactions are more than 10% of the minimum number of proteins in one category), 2)- positive overlapping and negative interactions, 3)- negative overlapping and positive interactions, and 4)- negative overlapping and negative interactions. It can be noted that in case of (positive & positive), enhanced results has been gained specially in increasing the TP and decreasing the TN and FP. Although the number of TP is small relating to one function of them, it is very accurate and equal

to the minimum number over the two functions. It is very clear that the numbers of TN and FP are decreased as in cases functions (1-21) which they have FP equal to 141 and 395 respectively and now it is 74.

Table 6. shows the cases of integrated functions relating to the overlapping number of proteins and number of interactions according to the determined threshold in algorithm.

F:x-y	Overlapping		Interactions		TP	TN	FP
1-2	31/31	+	3/31	(~)+	2	29	15
1-11	23/23	+	11/23	+	6	17	7
2-21	31/31	+	5/31	+	2	29	13
11-21	23/23	+	10/23	+	6	17	7
1-20	83/84	+	7/84	-	4	80	26
1-21	224/247	+	53/247	+	29	218	74
20-21	82/84	+	12/84	+	4	80	26
1-4	66/99	-	9/93	-	4		14
4-21	66/93	-	9/93	-	4		13
1-14	64/247	-	49/247	+	12		35
18-21	0/23	-	21/23	+	0		12

When the two scores are negative the results are poor which reflects or demonstrates the effect of overlapping and interactions. When one of them is positive and the other is negative, it has variety in results. The negative of interaction score fixes the number of TN and the negative of overlapping score increases the TN. We can conclude that the overlapping numbers of proteins and the number of interactions has affected the protein function prediction process in positive way. And the relations between the protein functions enhanced the degree of confidence.

4. Conclusions

In this paper, an integrated technique is introduced to estimate the correlations or relations between yeast protein functions. The technique depended on the overlapping number of proteins as well as number of interactions over the protein clusters. By applying the proposed algorithm on the collected data, the results have been improved; reducing the number of true negative and false positive furthermore increasing the true positive results. The results were good when the two measures were positive. Although the number of interactions was important for enhancement the results but the overlapping number was more critical. In protein function prediction problem, the effect of the function correlations has been indicated and the results were better than the absolute method (neighbor counting method without function correlation). As future work, considering the relations between protein functions into the different statistical algorithms is very important step.

REFERENCES

[1] D. Harrington, A. H. Singh, T. Doerks, I. Letunic, C. von Mering, and P. Bork, "Quantitative assessment of protein function prediction from metagenomics shotgun sequences," Proc Natl Acad Sci U S A, vol. 104, pp. 13913-8, Aug 28 2007

[2] R. V. Spriggs, Y. Murakami, and S. Jones, "Protein function annotation from sequence: prediction of residues interacting with RNA," Bioinformatics, vol. 25, pp. 1492-7, Jun 15 2009

[3] J. C. Whisstock and A. M. Lesk, "Prediction of protein function from protein sequence and structure," Q Rev Biophys, vol. 36, pp. 307-40, Aug 2003

[4] I. Friedberg, "Automated protein function prediction--the genomic challenge," Brief Bioinform, vol. 7, pp. 225-42, Sep 2006

[5] B. Schwikowski, and S. Fields, "A network of PPI in yeast," Nat Biotechnol, vol. 18, pp. 1257-61, Dec 2000

[6] H. Hishigaki, K. Nakai, T. Ono, and T. Takagi, "Assessment of prediction accuracy of protein function from protein--protein interaction data," Yeast, vol. 18, pp. 523-31, Apr 2001

[7] M. Deng, K. Zhang, S. Mehta, T. Chen, and F. Sun, "Prediction of protein function using protein-protein interaction data," J Comput Biol, vol. 10, pp. 947-60, 2003

[8] N. Nariai, E. D. Kolaczyk, and S. Kasif, "Probabilistic protein function prediction from heterogeneous genome-wide data," PLoS One, vol. 2, p. e337-344, 2007

[9] M. Zhao, and K. Aihara, "Gene function prediction using labeled and unlabeled data," BMC Bioinformatics, vol. 9, p. 57-71, 2008

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

[11] Y. Liu, and H. Zhao, "Protein interaction predictions from diverse sources," Drug Discov Today, vol. 13, pp. 409-16, May 2008

[12] K. Sayed, N. Soloma, and Y. Kadah, "Estimation of the correlation between protein sub-function categories based on overlapping proteins," Proc. 27th NRSC, Menouf, Egypt, March 2010

[13] A. Wagner, " How the global structure of protein interaction networks evolves," Proc Biol Sci vol. 270, 2003

[14] K. Sayed, N. Soloma, and Y. Kadah "Improving the prediction of yeast protein function using weighted protein-protein interactions," Theoretical Biology and Medical Modelling, vol. 8, 2011.

[15] K. Sayed, N. Soloma, and Y. Kadah "Determining The Relations Between Protein Sub-Function Categories Based On Overlapping Proteins," Journal of Communication and Computer, vol. 8, 2011