Military Technical College
Kobry El-Kobbah,
Cairo, Egypt

8th International Conference
on Electrical Engineering
ICEENG 2012

Protein function Motif extraction based on single function category sequence alignment in yeast

By

Khaled Sayed Ahmed*

Abstract:
Protein function prediction is one of the most vital problems in the field of proteomics since it leads to determining cell functions and identifying the diseases and their effect. Since proteome is divided into clusters, each cluster (group of proteins) should have common characteristics. One of these characteristics is to have the same functions. In this study we try to extract motifs for each sub-function category of yeast proteins. The technique is based on applying multiple sequence alignment (MSA) to all yeast protein function categories. The protein sequences are collected from different data sources as DIP, PIR, and SWISS PROT and CLC program is used to apply the sequence alignment. The technique is applied to proteins having single function only that because multi function proteins can be affected by the functions correlation. Threshold is determined for every protein function category to indicate the most common frequent amino acids to be a feature for this category. After implementing the algorithm, sequence is verified with some proteins have the correct functions and the gained results are good. The technique is considered as verification method for protein function prediction. And reference database table is constructed based on the extracted motifs.

Keywords:
Sequence alignment, Motif extraction, Consensus, Function category

---

* College of Engineering, Modern University for Technology and Information, Cairo, Egypt

1. Introduction:

The most important macromolecules in cells are the proteins which response for doing many functions. Proteins are the main building blocks and functional molecules of the cell. Every codon (3 bases of RNA nucleotides {A, G, C, U}) correspond to one amino acid which is arranged to build the protein. Proteins are consisted of sequence of amino acids which are the basic units of structure [1]. When the 20 amino acids (natural components) are sequenced in different numbers and different orders, infinite number of proteins can be created. If the length of amino acids is more than 40 ones, it called protein otherwise called multi peptide [2].

The sequence of amino acids is response for constructing the folding shape of protein (3D structure) as well as its main functions. In particular, proteins transmit regulatory signals throughout the cell, catalyze a tremendous number of chemical reactions, and are critical for the stability of numerous cellular structures. As known, each group of proteins having specific characteristic is called cluster (group). One example of these clusters is the similarity in doing specific function. Many methods have been developed to predict the protein functions as protein sequences [3, 4], analyzing gene expression patterns [5,6], phylo-genetic profiles [7,8,9], protein domains [10, 11], and protein interaction networks [12-16], estimated correlation [17,18], and weighted interactions [19].

Since most of the prediction methods depend on the protein sequences and the fact that if two proteins have similar sequences, they may have the same function [20]. A previous study [21] has been introduced exploring the protein sequence alignments. This study has extracted motifs for certain group of function categories with low accuracy. The reason was due to collecting a huge number of protein sequences carrying more than one function category. The other functions have affected the studied one. In this study, the collected protein sequences will have just one / single function category.

Each protein function category (cluster) has group of proteins is defined and their protein sequences are collected. Many data sources (database) as DIP, PIR, and SWISS PROT are used to get these sequences. Accurate multiple sequence alignment technique is performed using Bio-CLC program. So in this paper, we introduce technique using multiple sequence alignment to extract (certain features) motif for each sub-function category. The technique has been applied to Yeast protein sequences. The extracted consensuses are collected and considered as features/ signatures for every sub-function category. The protein function prediction process has been verified and compared by true functions through NCBI. 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

Certainly, yeast proteins should be divided into clusters. And protein sequences should be collected. There are two common methods to collect these sequences:1)-through web engine NCBI (one of the multiple diverse sources in identification the proteins and determining their functions) or 2)-downloading the protein sequences from protein sequence database. In this study, an integrated method has been used between different data sources to get the annotated protein sequences. The protein sequences which have the same sub-function category were collected. And multiple sequence alignment (MSA) has been provided to extract specific motif (consensus) for each sub-function category.

A. Protein Sequence Collection

Although BLAST and NCBI web sites were used to get the protein data, it was very exhaustive process to gain the protein sequences manually. A group of databases as DIP, PIR, SWISS-PROT, and MIPS have been integrated to collect these sequences. This integration has been performed, since all annotated proteins have not been found in one database. Although DIP (Database of Interacting Proteins) was the most famous data source used to get the sequences of yeast proteins, it missed for some proteins which collected from other (resources) databases.

Table (1) Sample of protein names and their sequences from DIP database

Protein Name	#AA	Protein sequence
BAG7	409	MFNMNLLSTPSSEEGSPQNRSSSMSSVEGKKDRDTFTNLQNEFDGKVFGVSLEESLKVAQEEVIIQKSTN EIGSIPVVIAKSGKYLKENALDTTGIFRIAGSNKRVRELQAVFSKPPDYGRKFEGWCDFNVHDIATLLKR YLNSLSEPLVPLALYDIFRNPILENPKINEHKEQIIKDYEDIYMLLPQQNRHLILYLAALLNLFARNEKK NLMSASNLAAIVQPSLLSHPKDEMCPKEYEASRTVIEFLILHASDIIPNTEKANKDTMPHAGTVAKFNNI TVPEMAIDSDEEDFVHPSIDDHMLPRSRALSDSNNFTIHHHHHHHHALFPSPIDFDNNGLSVPRSFKGRT LSAESLSPRLSKLLGNVGNSSNTGIKDPTERVPRGEHKTKHKQRQSWLRRLTSPSRTQP
RGA2	1009	MSADPINDQSSLCVRCNKSIASSQVYELESKKWHDQCFTCYKCDKKLNADSDFLVLDIGTLICYDCSDKC TNCGDKIDDTAIILPSSNEAYCSNCFRCCRCSNRIKNLKYAKTKRGLCCMDCHEKLLRKKQLLLENQTKN SSKEDFPIKLPERSVKRPLSPTRINGKSDVSTNNTAISKNLVSSNEDQQLTPQVLVSQERDESSLNDNND NDNSKDREETSSHARTVSIDDILNSTLEHDSNSIEEQSLVDNEDYINKMGEDVTYRLLKPQRANRDSIVV KDPRIPNSNSNANRFFSIYDKEETDKDDTDNKENEIIVNTPRNSTDKITSPLNSPMAVQMNEEVEPPHGL ALTLSEATKENNKSSQGIQTSTSKSMNHVSPITRTDTVEMKTSTSSSTLRLSDNGSFSRPQTADNLLPHK KVAPSPNKKLSRSFSLKSKNFVHNLKSKTSEMLDPKHPHHSTSIQESDTHSGWGVSSTHTNIRKSKAKKN PVSRGQSDSTIYNTLPQHGNFTVPEFNHKKAQSSLGSISKKQNSNDTATNRRINGSFTSSSSGHHIAMFR TPPLESGPLFKRPSLSSESAHHRSSSLQTSRSTNALLEDDSTKVDATDESATSLEKDFYFTELTLRKLKL DVRELEGTKKKLLQDVENLRLAKERLLNDVDNLTREKDKQSASSRESLEQKENIATSITVKSPSSNSDRK GSISNASPKPRFWKIFSSAKDHQVGDLESQQRSPNSSSGGTTNIAQKEISSPKLIRVHDELPSPGKVPLS PSPKRLDYTPDGSHLYGSSLQARCAYEKSTVPIIIRCCIDRIEKDDIGLNMEGLYRKSGSQTLVEEIENE FAQNNSLHSDTLSPKLNALLNQDIHAVASVLKRYLRKLPDPVLSFSIYDALIDLVRNNQLIERLPLNNDK FLDSPQKVTIYEMVLKSLLEIFKILPVEHQEVLKVLAAHIGKVRRCSERNLMNLHNLSLVFAPSLIHDFD GEKDIVDMKERNYIVEFILGNYRDIFKQA
RGD2	714	MLSFCDYFWSEDLVSGLDVLFDRLYHGCEQCDLFIQLFASRMQFEVSHGRQLFGIEAGMDNLKAVQEDED EGVTVSRALRGILQEMSQEGTHHLTIASNIESLVLQPFSKWCIEHRERIQYSEKTLLTNVNNFRKSKKYV GKLEKEYFNKCRQLEEFKRTHFNEDELANAMKSLKIQNKYEEDVAREKDHRFFNRIAGIDFDYKTMKETL QLLLTKLPKTDYKLPLISYSLSNTNNGGEITKFLLDHMSLKDIDQAETFGQDLLNLGFLKYCNGVGNTFV NSKKFQYQWKNTAYMFANVPMPGSEEPTTGESLISRFNNWDGSSAKEIIQSKIGNDQGAAKIQAPHISDN ERTLFRMMDALAASDKKYYQECFKMDALRCSVEELLIDHLSFMEKCESDRLNAIKKATLDFCSTLGNKIS SLRLCIDKMLTLENDIDPTADLLQLLVKYKTGSFKPQAIVYNNYYNPGSFQNFGVDLETRCRLDKKVVPL IISSIFSYMDKIYPDLPNDKVRTSIWTDSVKLSLTHQLRNLLNKQQFHNEGEIFDILSTSKLEPSTIASV VKIYLLELPDPLIPNDVSDILRVLYLDYPPLVETALQNSTSSPENQQDDDNEEGFDTKRIRGLYTTLSSL SKPHIATLDAITTHFYRLIKILKMGENGNEVADEFTVSISQEFANCIIQSKITDDNEIGFKIFYDLLTHK KQIFHELKRQNSKN

As shown in Table (1), a sample of protein names (Gene name) and parts of their sequences has been indicated. It can be noted that, the protein names were gene names which means the DIP database dealt with gene names not the international name (accession number). The second and third columns indicated the amino acid counter and the protein sequences respectively. A comparative study between the protein names and data sources has been performed to validate the data of proteins that relating to the distinguished names of proteins (Gene name, Accession number, and ORF) and the different places for databases.

Yeast protein functions have been divided into three categories: Bio-chemical functions (contains 57 sub-function categories), Cellular role functions (contains 43 sub-function categories), and Cell location (contains 29 sub-function categories). The study collected protein sequences related for every specific sub-function category in one place. The collected protein sequences had just one sub-function category. As shown in figure (1), number of proteins having one function category has been introduced. It can be noted that for Biochemical function category_34 (Oxidoreductase) which contains 263 proteins, we have found about 27 [DLD2-FDH1-FRE5-FRE6-MET12-PIG2-SPS100-STR2-STV1-TAH18-UGA2-YBR014C-YDL010W-YDR133C-YDR199W-YDR286C-YEL070W- YGL198W-YIR035C-YJR039W-YJR078W-YJR149W-YKL102C-YLR364W-YML125C-YNL155W-YNR073C-YOL024W-YPL276W-ZTA1] proteins have this singe function category only. The rest (236 proteins) participate in other function categories. Also Biochemical function category-53 (Transferase) which has 533 proteins, only four proteins has this function only. Functional category_14 (DNA-binding protein) has four proteins only and function category_16 (GTPase activating protein) has three proteins.

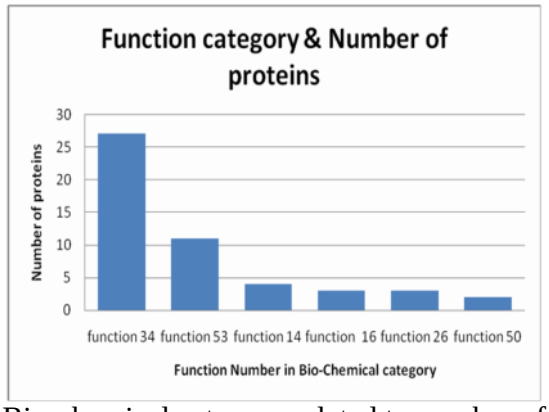

Figure (1) The Yeast protein Bio-chemical category related to number of proteins have single function

Although this study is performed to extract specific signature or motifs for every function category, it can be noted that group of functions have some specific proteins (associated proteins) as shown in figure (2). This note leads us to apply the same

technique (multiple sequence alignment) to the proteins which have the same two function categories.

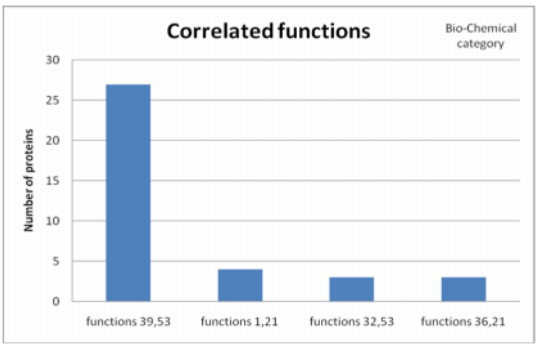

Figure (2) the Common functions related to number of proteins

B. ***Multiple sequence alignment***

Although there were many methods used in motif extraction as Deterministic algorithm (match or mismatch), Probabilistic algorithm, Combination between Deterministic and Probabilistic presentation and M-PST (mismatch probabilistic suffix tree) [22], the multiple sequence alignment has produced good results. Also it has been used in determining the interacted proteins [23] and probabilistic approach [24]. In this study, CLC BIO package has been used to perform MSA (multiple sequence alignment) for all collected protein sequences that have the same function. As shown in figure (3), the alignment process after applying MSA to the FASTA format proteins shown in table-2.

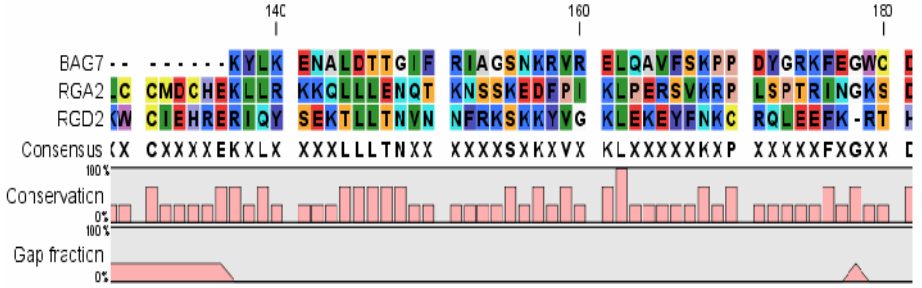

Figure (3) the multiple sequence alignment of three collected sequences of GTPase activating protein category

The button part of figure-3 indicates the conservation and gap fraction of the alignment process. The conservation means the strength of the alignment (most frequent) of amino acids found in one place. On the other hand, the gap fraction means the difference between amino acids in this location. High conservation and low gap fraction are good indication for extracting the consensus (motif). If the amino acids have high density relationships so the conservation level is increased and gab fraction is decreased (low percentage). The consensus part reflects the alignment strength. Consensus as [SxNDSGx-P] for example will be understood as letters. These extracted letters can be divided into three parts: 1)- Capital letters SNDSGP which means the first letters of

most proper amino acids, 2)- (- sign) means gap (no amino acid in this location) and 3)- (x) means any amino acid can be found in this location. On the other hand, unrelated sequences have poor relations and high gap fraction. The previous figure shows low level gab fraction and high level conservation which reflects the strength of the alignment in this area. Starting from AA position (137) till position 178 can be considered as motif for this function.

Table (2) Proteins have Biochemical function category_16 only (GTPase activating)

	Gene name	Different protein names		
1	BAG7	YOR3320	O3320	YOR134W
2	RGA2	D9481.4	YDR379W	
3	RGD2	YFL047W		

Another example for Biochemical function category_16 is shown in figure (4) which indicates the result of 27 sequence alignment. The gab fraction of this alignment is not constant through amino acid sequence. The motif can be extracted from position 817 or position 838 reaching for 853. The judge in this situation is to validate the extracted sequence through NCBI. Here after validation, the motif can be taken starting from position 838.

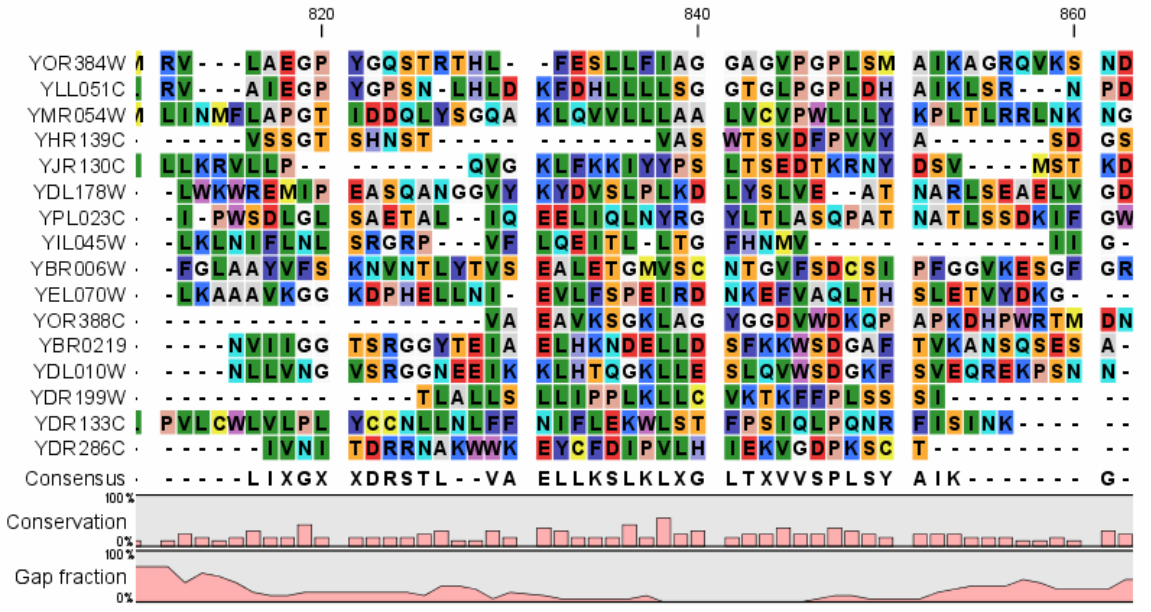

Figure (4) the sequence alignment of Biochemical function category_34 (Oxidoreductase)

3. Results

In this study, the multiple sequence alignment is applied to yeast function categories. The protein sequences are divided into three main categories including 127 sub-

functions. For each sub-function, protein sequences are collected and sequence alignment is performed to extract specific motif. This motif is considered as feature (signature) for this function category. As shown in table-3, two motifs are indicated for function categories 16, and 34 respectively. These motifs are considered as identified features for each function. And it can be used to verify the predicted functions.
If motif of function (**A***)* for example is found in protein sequence (xx) and the mathematical methods estimated that protein (xx) has this function (**A**), it can be said, protein has high confidence to have this function (high probability).
In this study, the suggested technique overcomes the previous problems faced in collecting many sequences having more than one function. it can be said that the function category has no expressed motif (specific features), if the collected sequences have poor alignment.

Table (3) the extracted motifs relating to the function category

Function ID	Function Name	Starting position	End position	Consensus
16	GTPase activating	137	178	EK-L----LLLTN------S-K-V-KL-----K-P-----F
34	Carbohydrate metabolism	838	853	LKLXGLTXVVSPLSYAIK

4. CONCLUSIONS

Herein, the multiple sequence alignment is applied to each group of protein sequences have specific function. This alignment is done to extract motif to be as identified feature (signature) for this function. These motifs are collected and used for verification process of protein function prediction. Each function can be predicted from any mathematical method, can be verified using this method by motif extraction search.

5. FUTURE WORK

A new technique can be applied to the proteins that contain just one function (collecting protein domains response for performing the function). The protein domains may response for doing the function or construction the 3D structure (tertiary structure).

References:

[1] I. M. KAPETANOVIC, S. ROSENFELD, AND G. IZMIRLIAN, "*OVERVIEW OF COMMONLY USED BIOINFORMATICS METHODS AND THEIR APPLICATIONS*," ANN N Y ACAD SCI, VOL. 1020, PP. 10-21, MAY 2004.

[2] J. Yang. Jingyi Yang, Jitender S. Deogun, Zhaohui Sun *"A New Scheme for Protein Sequence Motif Extraction"*.HICSS (2005).

[3] E. D. Harrington, A. H. Singh, T. Doerks, I. Letunic, C. von Mering, L. J. Jensen, J. Raes, 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.

[4] 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.

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

[6] 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.

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

[8] J. Sun and Z. Zhao, *"Construction of phylogenetic profiles based on the genetic distance of hundreds of genomes,"* Biochem Biophys Res Commun, vol. 355, pp. 849-53, Apr 13 2007.

[9] M. Pellegrini, E. Marcotte *"Assigning protein functions by comparative genome analysis: protein phylogenetic profile"* s. Proc Natl Acad Sci U S A , vol.96, pp.:4285-8, 1999.

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

[11] 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.

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

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

[14] 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.

[15] 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.

[16] K. S. Ahmed, N. H. Soloma., and Y. M. Kadah,*" Comparison between different methods for protein function prediction*" .in 1st International Joint Conference (NRC), 2009.

[17] Khaled S. Ahmed, Nahed H. Solouma, Yasser M. Kadah, *"Determining The Relations Between Protein Sub Function Categories Based On Overlapping Proteins," Journal of Communication and Computer*, vol. 8, no. 3, pp. 240-245, 2011.

[18] Khaled S, Nahed S., Yasser K: "*Exploring Protein Functions Correlation Based On Overlapping Proteins and Cluster Interactions*". *1st middle east conference for biomedical engineering, Sharjah, 2011.*

[19] K. Ahmed, 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.

[20] R. Sharan, S. Suthram, R. M. Kelley, T. Kuhn, S. McCuine, P. Uetz, T. Sittler, and T. Ideker, *"Conserved patterns of protein interaction in multiple species,"* Proc Natl Acad Sci U S A, vol. 102, pp. 1974-9, Feb 8, 2005.

[21] Khaled S. Ahmed, Yasser M. Kadah *" Yeast Protein Function Motif Extraction Based on Sequence Alignment"* NRSC, 2011 (in press)

[22] A. C. Gavin, M. Bosche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. M. Rick, A. M. Michon, R. R. Copley, A. Edelmann, E. Querfurth, V. Rybin, G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer, and G. Superti-Furga, *"Functional organization of the yeast proteome by systematic analysis of protein complexes,"* Nature, vol. 415, pp. 141-7, Jan 10 2002.

[23] Y. Ho, A. Gruhler, A. Heilbut, G. D. Bader, L. Moore, S. L. Adams, A. Millar, P. Taylor, K. Bennett, P. A. Nielsen, K. J. Rasmussen, J. R. Andersen, L. E. Johansen, L. H. Hansen, H. Jespersen, A. Podtelejnikov, E. Nielsen, J. Crawford, V. Poulsen, B. D. Sorensen, J. Matthiesen, R. C. Hendrickson, F. Gleeson, T. Pawson, M. F. Moran, D. Durocher, M. Mann, C. W. Hogue, D. Figeys, and M. Tyers, *"Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry,"* Nature, vol. 415, pp. 180-3, Jan 10 2002.

[24] S. Maslov and K. Sneppen, *"Specificity and stability in topology of protein networks,"* Science, vol. 296, pp. 910-3, May 3 2002.