28th NATIONAL RADIO SCIENCE CONFERENCE
(NRSC 2011)
April 26-28, 2011, National Telecommunication Institute, Egypt
C25

Yeast Protein Function Motif Extraction Based on Sequence Alignment

Khaled Sayed[#1], *Nahed Solouma*[*2], *Yasser Kadah*[#3]

[#]Bio-Electronics Department MTI Modern University for Technology and Information khaled.sayed@k-space.org

[#2]Laser application department Cairo University

[#3]Biomedical Engineering Department Cairo University

ABSTRACT

Protein function prediction is one of the most important problems in the field of proteomics since it leads to determining cell functions. 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. 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

Keywords: Protein sequence, MSA, Motif, Consensus.

I. INTRODUCTION

Proteins are macromolecules response for doing many functions. They are the main building blocks and functional molecules of the cell. Every 3 bases of RNA (codon) 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 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 for these clusters is the similarity in doing specific function. Many methods have been developed to predict the protein functions as analyzing gene expression patterns [3,4], phylo-genetic profiles [5,6,7], protein sequences [8, 9], protein domains [10, 11], and protein interaction networks [12-16], and estimated correlation [17].

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 [18], the protein sequences will be our interest in this study. Each protein function category (cluster) has group of proteins is defined and their protein sequences are collected. Many data sources 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 motif for each sub-function category. The technique has been applied to Yeast protein sequences. The extracted consensuses are collected and considered as feature for each sub-function category and protein function prediction process has been verified. 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.

II. METHODOLOGY

Protein sequence search in BLAST or NCBI is considered one of the multiple diverse sources in identification the proteins and determining their functions [19]. In this study, an integrated method has been used between different data sources to get the annotated protein sequences. These sequences which have the same sub-function

category were collected. And multiple sequence alignment has been used to extract specific motif (consensus) for each sub-function category.

A. Protein Sequence Collection

lthough 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 databases.

As shown in Table-1, a sample of protein names 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). So comparison between the protein names and data sources has been performed to identify all data about the protein. It can be shown in table-2, some missed cells which loss the corresponding names for these proteins. As example; protein code DIP: 239N equal Q27272 code in SWISS PROT equal A49067 in PIR database. But DIP 772 did not have corresponding code in SWISS PROT. Table-2 indicated the relations over different data sources for yeast proteins. Relating to the distinguished names of proteins (Gene name, Accession number, and ORF) and the different places for databases, the standard core for protein should be given (Accession number).

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

Protein Name	Protein sequence
BNI1	MLKNSGSKHSNSKESHSNSSSGIFQNLKRLANSNATNSNTGSPTYASQQQHSPVGNEVSTSPASSSS……
BNI4	MSDSISDSKSSELLNSTFYSSTSINTLDHARTFRNSLILKEISDQSLNSSIKPCESVLDRDVESSVLQ……
BNI5	MGLDQDKIKKRLSQIEIDINQMNQMIDENLQLVEPAEDEAVEDNVKDTGVVDAVKVAETALFSGND….
BUD2	MSSNNEPAQSRTSYFKLNEFLSNVKHYKNTFKGEIQWCNNLSLNDWKTHYLQITSTGALTHSIDELTA….
BUD3	MEKDLSSLYSEKKDKENDETLFNIKLSKSVVETTPLNGHSLFDDDKSLSDWTDNVFTQSVFYHGSDD…
BUD4	MAQDIDKLARDEEKPVKLSSSPLKFTLKSTQPLLSYPESPIHRSSIEIETNYDDEDEEEEDAYTCLTQS….
BUD5	MRTAVPQLLEATACVSRECPLVKRSQDIKRARKRLLSDWYRLGADANMDAVLLVVNSAWRFLAVWR…
BUD6	MKMAVDDPTYGTPKIKRTASSSSSIETTVTKLLMSTKHLLQVLTQWSKGTTSGRLVSDAYVQLGNDF…

Table 2: Different data sources for protein names and their codes

DIP interaction	DIP code	SP code	PIR code	GI code	DIP code	SP code	PIR code	GI code
DIP:193E	DIP:239N	SWP:Q27272	PIR:A49067	GI:1079142	DIP:368N	SWP:P04637	PIR:DNHU53	GI:8400738
DIP:196E	DIP:237N	SWP:P47825	PIR:A48184	GI:477148	DIP:36N	SWP:P08047	PIR:A29635	GI:88887
DIP:199E	DIP:772N		PIR:S41672	GI:1085161	DIP:368N	SWP:P04637	PIR:DNHU53	GI:8400738
DIP:207E	DIP:387N	SWP:P12428	PIR:FYFFB	GI:72497	DIP:388N	SWP:P10090	PIR:FYFFW	GI:17136592
DIP:229E	DIP:237N	SWP:P47825	PIR:A48184	GI:477148	DIP:570N	SWP:P03254	PIR:Q2AD2	GI:74182
DIP:271E	DIP:121N	SWP:P19538	PIR:A38926	GI:24638496	DIP:754N	SWP:P41044	PIR:S37695	GI:17136674
DIP:272E	DIP:492N		PIR:JC4234	GI:17137760	DIP:121N	SWP:P19538	PIR:A38926	GI:24638496
DIP:273E	DIP:45N		PIR:A31225	GI:24639671	DIP:526N		PIR:JU0092	GI:17136554
DIP:274E	DIP:54N	SWP:P13677	PIR:A32392	GI:17136716	DIP:526N		PIR:JU0092	GI:17136554
DIP:275E	DIP:769N		PIR:S40691	GI:2119474	DIP:526N		PIR:JU0092	GI:17136554
DIP:276E	DIP:537N	SWP:P07181	PIR:MCFF	GI:17647231	DIP:526N		PIR:JU0092	GI:17136554
DIP:342E	DIP:187N	SWP:P10083	PIR:A43731	GI:17136654	DIP:73N	SWP:P18491	PIR:A34688	GI:24654863
DIP:344E	DIP:40N	SWP:P16371	PIR:A30047	GI:24650241	DIP:637N		PIR:S06956	GI:85137
DIP:345E	DIP:325N	SWP:P10084	PIR:B43731	GI:17136616	DIP:356N	SWP:Q01068	PIR:D46177	GI:24650229

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 all possible protein sequences related for each specific sub-function category in one place. The collected protein sequences were in average 41% of the total number of protein sequences. An example for the collected number of sequences, for Biochemical ATPase sub-function category which had 247 proteins, 112 protein sequences have been collected. But for Biochemical protein motor sub-category which had 17 proteins, 12 protein sequences have been collected.

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) [20], the multiple sequence alignment has produced good results. Also it has been used in determining the interactions protein [21] and probabilistic approach [22].

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-1, the alignment process after applying MSA to the FASTA format protein sequences.

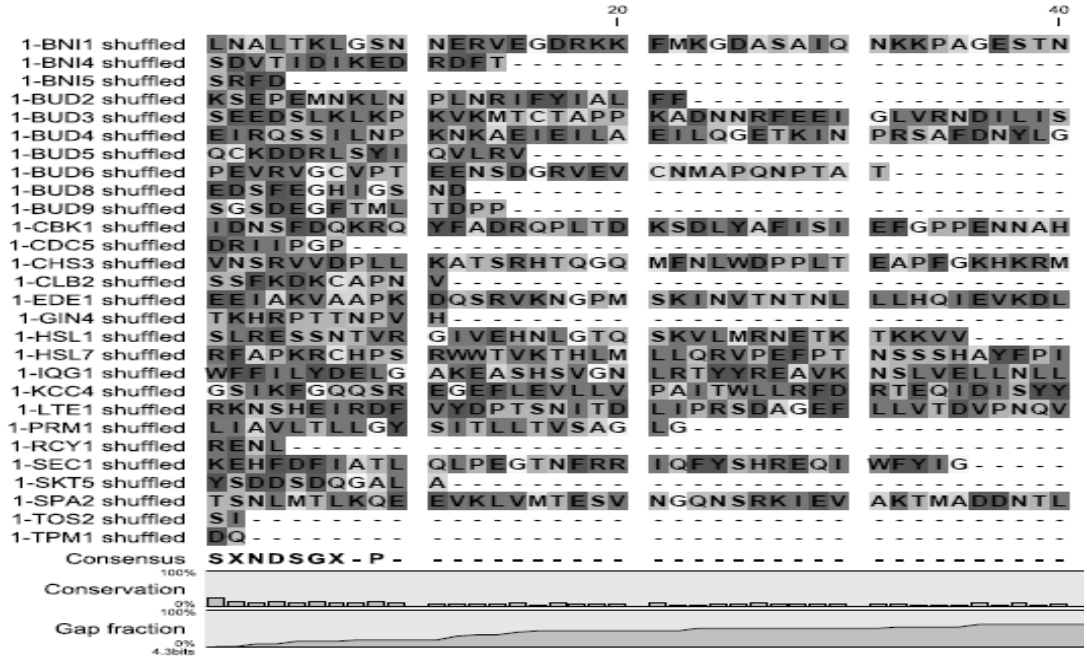

Fig. 1: MSA over collected proteins. Specific motif (consensus) has been extracted from the first 10 proteins locations.

As shown in the button part of figure-1, the conservation and gap fraction have been gotten. The conservation means the strength (most frequent) of amino acids in one place but 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. The first ten amino acids have high density relationships so the conservation level is increased and gab fraction is low percentage. So consensus can be created as [SxNDSGx-P]. The extracted letters can be divided into three parts. Capital letters SNDSGP which means the first letters of most proper amino acids but (-) means gap (no amino acid in this location) and (x) means any amino acid can be found in this location. On the other hand, unrelated sequences have poor relations so they have high gap fraction.

Although the motif was clear for each specific area, the motif extraction was difficult process to be automatically extracted. Since the manual extraction was very exhaustive process, a detected threshold has been created for consensus spectrum. This threshold was detected relating to the maximum conservation percentage of the sequence alignment. The relation between the conservation percentage and alignment position has been created as shown in figure-2. The threshold was around 20% which means any alignment conservation more than threshold (~0.2) will be as motif location (measure for the function). Figure-2 shows 11 peaks more than the determined threshold where there is just one as maxima.

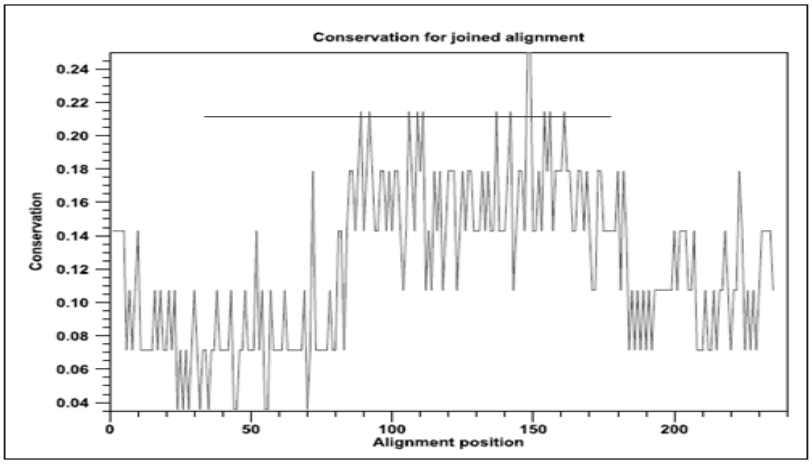

Fig. 2: the relation between the conservation and alignment position.

It can be noted, the threshold value for determining the consensus is different for each sub-function category according to the sequences alignment strength as shown in figure-3. This value can be determined visually from the spectral graph or from the two dimensional data array between the sequence position and frequent percentage.

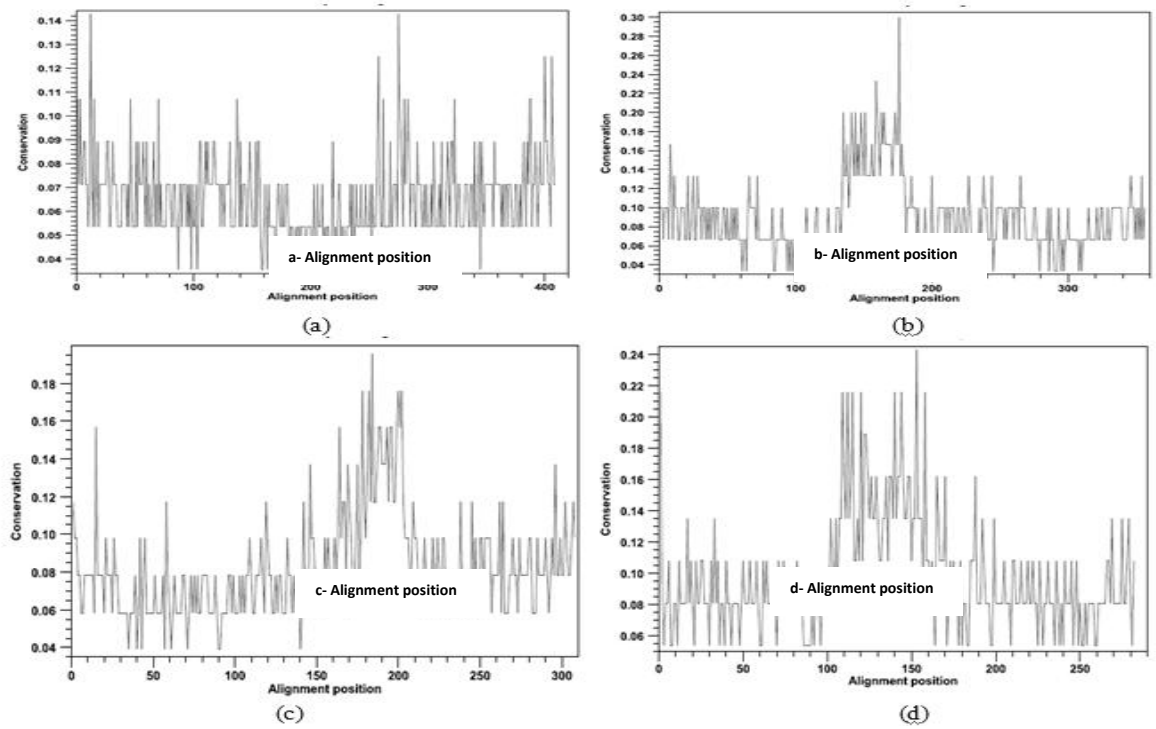

Fig. 3: four different alignment graphs spectrum (consensus versus the position) for Bio-chemical sub-function categories.

III. Results

In this study, the multiple sequence alignment has been applied to all protein sequences of Yeast. The protein sequences were divided into three main categories including 127 sub-functions. For each sub-function, protein sequences have been collected and sequence alignment is done to extract specific motif. This motif is considered as feature (signature) for this function. As shown in table-3, some motifs have been collected for each sub-function category. These motifs were 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 and the mathematical methods estimated that protein has function (**A**), it can be said, protein has high confidence to have this function (high probability). Also it can be noted, there are some sub-functions have no motifs. The reasons in such case are as The following: 1)- the matching of alignment is less than the suggested threshold, 2)- the sequences have other functions may change the main sequence of the protein, the group of proteins (collected sequences) does not express about the function. As shown in table-3, function ID 2 has no clear motifs because in figure-3a, the maximum amplitude does not reach for the required threshold. But the extracted motifs for protein functions 3, 8, and 5 as match as the peaks of the spectrum of the figure-3, b, c, d respectively.

Table 3: Yeast protein functions and its extracted motifs indicating the start and end positions

Function ID	Function Name	Starting position	End position	Consensus
2	Amino-acid metabolism	--	--	----------
3	Carbohydrate metabolism	138	176	TS-----ATELSR-R--T-A-AN---LEDL--------I
8	Cell structure	182	203	LDSRSSEXSEAALST---ESES
5	Cell cycle control	111	140	SLSSNXTLNTXEXESS--SEELKXTTRSEQSRRSTSLKI---SSES-E

IV. 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.

V. Future Work

As future work, the new techniques will be applied to the proteins that contain just one function. Because the motif is a measure for each protein function, the multiple sequence alignment should not be applied to sequences have more one function.

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

[4] 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.

[5] M. Morin " Phylogenetic Networks Simulation, Characterization, and Reconstruction" New Mexico, 2007.
[6] 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.
[7]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.
[8]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.

[9] 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.
[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] K. S. Ahmed, N. H. Soloma., and Y. M. Kadah, "Estimation of the correlation between protein sub-function categories based on overlapping proteins". In 27th National Radio Science conference (NRSC), 2010.
[18] 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.
[19] http://newscenter.cancer.gov
[20] 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.
[21] 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.
[22] S. Maslov and K. Sneppen, "Specificity and stability in topology of protein networks," Science, vol. 296, pp. 910-3, May 3 2002.