Prediction of Protein–Protein Interactions Using Evolutionary and Structural Relationships

Nazar Zaki

***Abstract*—One of the central problems in modern biology is to identify the complete set of interactions among the proteins in a cell. The structural interaction of proteins and their domains in networks is one of the most basic molecular mechanisms for biological cells, and structural evidence indicates that, interacting pairs of close homologs usually interact in the same way. In this article, we make use of both evolutionary and structural relationships to predict interaction between protein pairs solely by amino acid sequence information. High quality core set of 150 yeast proteins obtained from the Database of Interacting Proteins (DIP) was considered to test the accuracy of the proposed method. The strongest prediction of the method reached over 70% accuracy. These results show great potential for the proposed method.**

***Index Terms*—Protein-protein interaction, pairwise alignment, protein domain, inter-domain linker regions.**

I. INTRODUCTION

A. The importance of protein–protein interactions

The term protein-protein interaction refers to the association of protein molecules and the study of these associations from the perspective of biochemistry, signal transduction and networks. Protein-protein interactions occur at almost every level of cell function, in the structure of sub-cellular organelles, the transport machinery across the various biological membranes, the packaging of chromatin, the network of sub-membrane filaments, muscle contraction, signal transduction, and regulation of gene expression, to name a few [1]. Abnormal protein-protein interactions have implications in a number of neurological disorders; include Creutzfeld-Jacob and Alzheimer's diseases. Because of the importance of protein-protein interactions in cell development and disease, the topic has been studied extensively for many years and a large number of approaches to detect protein-protein interactions have been developed. Each of these approaches has strengths and weaknesses, especially with regard to the sensitivity and specificity of the method.

B. Current methods to predict protein–protein interactions

One of the major goals in functional genomics is to determine protein interaction networks for whole organisms, and many of the experimental methods have been applied to study this problem. Co-immunoprecipitation is considered to be the gold standard assay for protein-protein interactions, especially when it is performed with endogenous proteins [2]. The protein of interest is isolated with a specific antibody. Interaction partners that stick to this protein are subsequently identified by western blotting. Interactions detected by this approach are considered to be real, but this method can only verify interactions between suspected interaction partners. Thus, this method is not a screening approach to identify unknown protein-protein interactions.

The yeast two-hybrid screen investigates the interaction between artificial fusion proteins inside the nucleus of yeast [3]. This approach can identify binding partners of a protein in an unbiased manner, but this method suffers from high false-positive rate which makes it necessary to verify the identified interactions by co-immunoprecipitation.

Tandem Affinity Purification (TAP) detects interactions within the correct cellular environment [4], which is a big advantage over the yeast two-hybrid approach. However, the TAP tag method requires two successive steps of protein purification, and thus this method cannot readily detect transient protein-protein interactions. This method is not an efficient means to detect physical protein-protein interactions that exist in different cellular environments either, which is especially important when studying the interaction network in an organism's genome a very significant in the post-genomic era.

Quantitative immunoprecipitation combined with knock-down (QUICK) relies on co-immunoprecipitation, quantitative mass spectrometry (SILAC) and RNA interference (RNAi) was introduced. This method detects interactions among endogenous non-tagged proteins [5], and thus this method's results have the same high confidence as co-immunoprecipitation. However, this method also depends on the availability of suitable antibodies.

These high-throughput methods have contributed tremendously in the creation of databases containing large sets of protein interactions, such as Database of Interacting Proteins (DIP) [6], MIPS [7] (developed at the Martinsried Institute for Protein Sequences) and Human Protein Reference Database (HPRD) [8]. In addition, several *in silico* methods have been developed to predict protein–protein interactions based on features such as gene context [9]. These include gene fusion [10], gene neighborhood [11] and phylogenetic profiles [12]. However, most of the *in silico* methods seek to predict functional association, which often implies but is not restricted to physical binding.

Despite the availability of the mentioned methods of predicting protein-protein interaction, the accuracy and coverage of these techniques have proven to be limited.

Nazar Zaki is an Assistant Professor with the College of Information Technology, United Arab Emirates University. Al-Ain 17555 UAE, (phone: +971-50-7332135; fax: +971-3-7626309; e-mail: nzaki@uaeu.ac.ae).

Computational approaches remain essential both to assist in the design and validation of the experimental studies and for the prediction of interaction partners and detailed structures of protein complexes [13].

C. Computational approaches to predict protein-protein interaction

Some of the earliest techniques predict interacting proteins through the similarity of expression profiles [14], coordination of occurrence of gene products in genomes, description of similarity of phylogenetic profiles [12] or trees [15], and studying the patterns of domain fusion [16]. However, it has been noted that these methods predict protein–protein interactions in a general sense, meaning joint involvement in a certain biological process, and not necessarily actual physical interaction [17].

Most of the recent works focus on employing the protein domain knowledge to predict the protein-protein interaction [18]-[22]. The motivation for this choice is that molecular interactions are typically mediated by a great variety of interacting domains [23]. It is thus logical to assume that the patterns of domain occurrence in interacting proteins provide useful information for training protein-protein interaction prediction methods [24]. An emerging new approach in the protein interactions field is to take advantage of structural information to predict physical binding [25]-[26]. Although the total number of complexes of known structure is relatively small, it is possible to expand this set by considering evolutionary relationships between proteins. It has been shown that in most cases close homologs (>30% sequence identity) physically interact in the same way with each other. However, conservation of a particular interaction depends on the conservation of the interface between interacting partners [27].

In this paper, we propose to predict protein-protein interaction using only sequence information. The proposed method combines evolutionary and structural relationships between protein pair to predict the interaction between them. Evolutionary relationships will be incorporated by measuring the similarity between protein pair using Pairwise Alignment. Structural relationships will be incorporated in terms of protein domain knowledge. We are encouraged by the fact that compositions of contacting residues in protein sequence are unique, and that incorporating evolutionary and predicted structural information improves the prediction of protein–protein interactions [28].

II. METHOD

In this paper, we present a simple yet effective method to predict protein-protein interaction solely by amino acid sequence information. Figure 1, illustrates the overview of the proposed method. It consists of three main steps: (a) extracting the evolutionary relationships by measuring regions of similarity that may reflect functional, structural or evolutionary relationships between protein sequences (b) downsize the protein sequences of interest by predicting and eliminating inter-domain linker regions (c) scanning and detecting domain matches in all the protein sequences of interest. Two proteins may interact if they share similar domains.

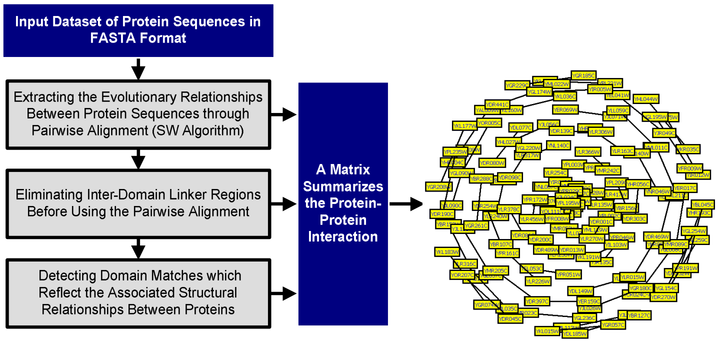

Input Dataset of Protein Sequences in FASTA Format
Extracting the Evolutionary Relationships Between Protein Sequences through Pairwise Alignment (SW Algorithm)
Eliminating Inter-Domain Linker Regions Before Using the Pairwise Alignment
Detecting Domain Matches which Reflect the Associated Structural Relationships Between Proteins
A Matrix Summarizes the Protein-Protein Interaction

Fig. 1. Overview of the proposed method.

A. Similarity Measures between Protein Sequences Using Pairwise Alignment

One of the most fundamental tools in the field of bioinformatics is sequence alignment. By aligning sequences to one another, it is possible to evaluate how similar the sequences are and identify conserved regions in sets of related sequences. Therefore, our proposed method starts by measuring the protein-protein interaction sequence similarity, which reflects the evolutionary and homology relationships. Two protein sequences may interact by the mean of the amino acid similarities they contain [24]. This work is motivated by the observation that the Smith-Waterman (SW), algorithm [29], which measures the similarity score between two sequences by a local gapped alignment, provides a relevant measure of similarity between protein sequences. This similarity incorporates biological knowledge about protein evolutionary structural relationships [30].

The Smith-Waterman similarity score $SW(x_1, x_2)$ between two protein sequences x_1 and x_2 is the score of the best local alignment with gaps between the two protein sequences computed by the Smith-Waterman dynamic programming algorithm. Let us denote by μ a possible local alignment between x_1 and x_2, defined by a number n of aligned residues, and by the indices

$1 \le i_1 < \ldots < i_n \le |x_1|$ and $1 \le j_1 < \ldots < j_n \le |x_2|$ of the aligned residues in x_1 and x_2 respectively. Let us also denote by $\prod(x_1, x_2)$ the set of all possible local alignments between x_1 and x_2, and by $p(x_1, x_2, \mu)$ the score of the local alignment $\mu \in \prod(x_1, x_2)$ between x_1 and x_2, the Smith-Waterman score $SW(x_1, x_2)$ between sequences x_1 and x_2 can be written as:

$$SW(x_1, x_2) = \max_{\mu \in \prod(x_1, x_2)} p(x_1, x_2, \mu) \qquad (1)$$

The similarity matrix can be calculated as follow:

$$Matrix\_a = \begin{bmatrix} SW(x_1,x_1) & SW(x_1,x_2) & \ldots & SW(x_1,x_m) \\ SW(x_2,x_1) & SW(x_2,x_2) & \ldots & SW(x_2,x_m) \\ : & : & : & : \\ SW(x_m,x_1) & SW(x_m,x_2) & \ldots & SW(x_m,x_m) \end{bmatrix} \qquad (2)$$

where m is the number of the protein sequences.

For example, suppose we have the following randomly selected protein-protein interaction dataset:

YDR190C, YPL235W, YDR441C, YML022W, YLL059C, YML011C, YGR281W and YPR021C represented by x_1, x_2, x_3, x_4, x_5, x_6, x_7 and x_8 respectively. The interaction between these 8 proteins is shown in Figure 2.

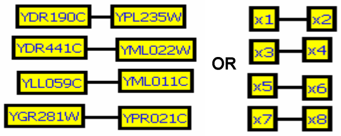


Fig. 2. The interaction between the randomly selected proteins.

Then the SW similarity score matrix $Matrix\_a$ will be calculated as:

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
x_1	X	465	28	30	25	30	34	29
x_2	465	X	30	24	32	33	50	47
x_3	28	30	X	553	29	27	32	29
x_4	30	24	553	X	29	20	25	40
x_5	25	32	29	29	X	24	28	49
x_6	25	33	27	20	24	X	25	26
x_7	34	50	32	27	28	26	X	36
x_8	29	47	29	40	49	26	36	X

From $Matrix\_a$, higher score may reflect interaction between two proteins. $SW(x_1, x_2)$ and $SW(x_2, x_1)$ scores are equal to 465; $SW(x_3, x_4)$ and $SW(x_4, x_3)$ scores are equal to 553, which confirm the interaction possibility. However, $SW(x_5, x_6)$ and $SW(x_6, x_5)$ scores are equal to 24; $SW(x_7, x_8)$ and $SW(x_8, x_7)$ scores are equal to 36, which are not the highest scores. To correct these errors more biological information is needed, which lead us to the second part of our method.

B. Identify and Eliminating Inter-domain Linker Regions

The results could be further enhanced by incorporating inter-domain linker regions knowledge. The next step of our algorithm is to predict inter-domain linker regions solely by amino acid sequence information. Our intention here is to identify and eliminate all the inter-domain linker regions from the protein sequences of interest. By doing this step, we are actually downsizing the protein sequence to shorter ones with only domains, which may produce better alignment scores. In this case, the prediction is made by using linker index deduced from a data set of domain/linker segments from SWISS-PROT database [31]. DomCut developed by Suyama *et al* [32] is employed to predict linker regions among functional domains based on the difference in amino acid composition between domain and linker regions. Following [32], we defined the linker index S_i for amino acid residue i and it is calculated as follows:

$$S_i = -\ln\left(\frac{f_i^{Linker}}{f_i^{Domain}}\right) \qquad (3)$$

Where f_i^{Linker} is the frequency of amino acid residue i in the linker region and f_i^{Domain} is the frequency of amino acid residue i in the domain region. The negative value of S_i means that the amino acid preferably exists in a linker region. A threshold value is needed to separate linker regions as shown in Figure 3. Amino acids with linker score greater than the set threshold value will be eliminated from the protein sequence of interest.

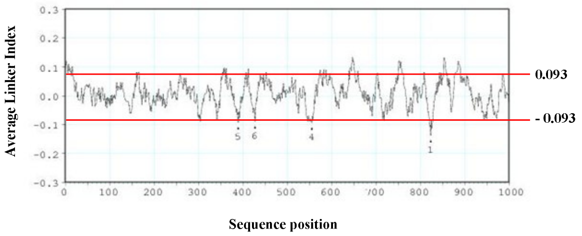


Fig. 3. An example of linker preference profile generated using Domcut. In this case, linker regions greater than the threshold value 0.093 will be eliminated from the protein sequence.

When applying the second part of the method, the matrix $Matrix\_a$ will be calculated as follows:

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
x_1	X	504	30	30	25	23	34	27
x_2	504	X	30	21	32	32	50	36
x_3	30	30	X	775	29	24	38	29
x_4	30	21	775	X	19	21	53	37
x_5	25	32	29	19	X	28	28	24
x_6	23	32	24	21	28	X	23	27
x_7	34	50	38	53	28	23	X	339
x_8	27	36	29	37	24	27	339	X

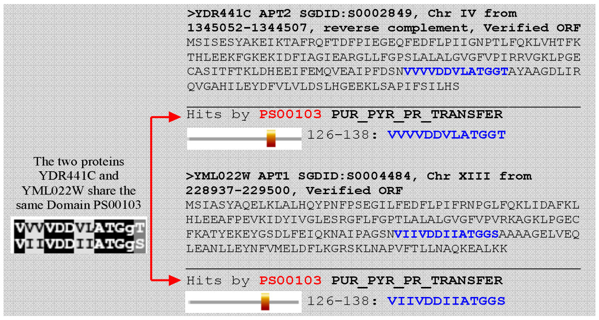
From $Matrix\_a$, its clearly noted that, more evidence is shown to confirm the interaction possibility between proteins x_7 and x_8, and therefore, the result is furthermore enhanced. In the following part of the method, protein domain knowledge will be incorporated in $Matrix\_a$ for better accuracy.

C. Detecting Domain Matches and Associated Structural Relationships in Proteins

In this part of the method, protein domains knowledge will be incorporated in $Matrix\_a$. Protein domains are highly informative for predicting protein-protein interaction as it reflects the potential structural relationships between them. In this implementation, we employed ps_scan [33] to scan one or several patterns, rules and profiles from PROSITE against our protein sequences of interest. Running ps_scan through the 8 proteins identifies the following Domains:

YDR441C (x_3) → PS00103
YML022W (x_4) → PS00103
YGR281W (x_7) → PS00211, PS50893 and PS50929
YPR021C (x_8) → PS50929

Which reveals structural relationships between proteins x_3 and x_4; and proteins x_7 and x_8. This step is illustrated in Figure 4.



```
>YDR441C APT2 SGDID:S0002849, Chr IV from
1345052-1344507, reverse complement, Verified ORF
MSISESYAKEIKTAFRQFTDFPIEGEQFEDFLPIIGNPTLFQKLVHTFK
THLEEKFGKEKIDFIAGIEARGLLFGPSLALALGVGFVPIRRVGKLPGE
CASITFTKLDHEEIFEMQVEAIPFDSNVVVVDDVLATGGTAYAAGDLIR
QVGAHILEYDFVLVLDSLHGEEKLSAPIFSILHS

Hits by PS00103 PUR_PYR_PR_TRANSFER
126-138: VVVVDDVLATGGT

>YML022W APT1 SGDID:S0004484, Chr XIII from
228937-229500, Verified ORF
MSIASYAQELKLALHQYPNFPSEGILFEDFLPIFRNPGLFQKLIDAFKL
HLEEAFPEVKIDYIVGLESRGFLFGPTLALALGVGFVPVRKAGKLPGEC
FKATYEKEYGSDLFEIQKNAIPAGSNVIIVDDIIATGGSAAAAGELVEQ
LEANLLEYNFVMELDFLKGRSKLNAPVFTLLNAQKEALKK

Hits by PS00103 PUR_PYR_PR_TRANSFER
126-138: VIIVDDIIATGGS
```

The two proteins YDR441C and YML022W share the same Domain PS00103

```
VVVVDDVLATGgT
VIIVDDIIATGgS
```

Fig. 4. An example of domain PS00103 found in proteins YDR441C and YML022W.

Based on this relationship, $SW(x_3, x_4)$ and $SW(x_7, x_8)$ will be calculated as follow:

$$SW(x_3, x_4) = SW(x_3, x_4) + k * 300 \qquad (4)$$

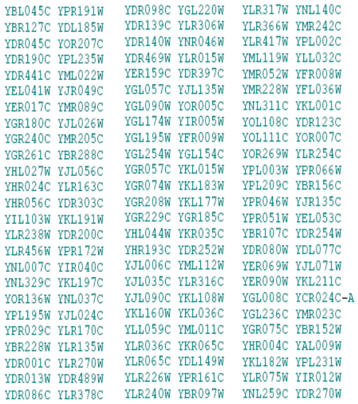
$$SW(x_7, x_8) = SW(x_7, x_8) + k * 300 \qquad (5)$$

Where k is the number of Domains the two interacting proteins share. Unfortunately, these results have not added more accuracy in this case, however, it confirmed the interacting possibilities between proteins x_3 and x_4; x_7 and x_8.

III. Experimentl Work

To test our method, we obtained the protein-protein interaction data from the Database of Interacting Proteins (DIP). The DIP database catalogs experimentally determined interactions between proteins. It combines information from a variety of sources to create a single, consistent set of protein-protein interactions in *Saccharomyces cerevisiae*. The data stored within the DIP database were curated, both, manually by expert curators and also automatically using computational approaches that utilize the knowledge about the protein-protein interaction networks. This knowledge is extracted from the most reliable, core subset of the DIP data [34]. The DIP version we used contains 4749 proteins involved in 15675 interactions for which there is domain information [6]. However, only high quality core set of 2609 yeast proteins was considered in this experimental work. This core set is involved in 6355 interactions, which have been determined by at least one small-scale experiment or two independent experiments [35]. Furthermore, we selected proteins interacts with only one protein and not involved in any other interactions. This process results in a dataset of 150 proteins with 75 positive interactions as shown in Figure 5. The intention here is to design a method capable of predicting protein interaction partner, which facilitate a way to construct protein-protein interaction using only protein sequences information.

We started our experimental work by measuring the protein-protein sequence interaction similarity using Smith-Waterman algorithm as implemented in FASTA [36]. The default parameters are used: gap opening penalty and extension penalties of 13 and 1, respectively, and a substitution matrix BLOSUM62 matrix. Various types of substitution matrices have been used over the years. Some were based on theoretical considerations, however, the most successful, based on analysis of alignments of numerous homologs of well-studies proteins from many different species [37]. The choice of which substitution matrix to use is not trivial because there is no one correct scoring scheme for all circumstances. The BLOSUM matrix is another very common used amino acid substitution matrix that depends on data from actual substitutions. This procedure produces the matrix $Matrix\_a_{150X150}$. This matrix was then enhanced by incorporating inter-domain linker regions information. In this case, only well defined domains with sequence length ranging from 50 to 500 residues were considered. We skipped all the frequently matching (unspecific) domains. A trashed value of 0.093 is used to separate the linker regions. Any residue generates an index greater than the threshold value results in eliminating it. This procedure downsized the protein sequences without losing the biological information. In fact, running the SW algorithm on a sequence having pure domains, results in better accuracy. A linker preference profile is generated using the linker index values along an amino acid sequence using a siding window. A window of size $w = 15$ is used because it gives the best performance.



YBL045C	YPR191W	YDR098C	YGL220W	YLR317W	YNL140C
YBR127C	YDL185W	YDR139C	YLR306W	YLR366W	YMR242C
YDR045C	YOR207C	YDR140W	YNR046W	YLR417W	YPL002C
YDR190C	YPL235W	YDR469W	YLR015W	YML119W	YLL032C
YDR441C	YML022W	YER159C	YDR397C	YMR052W	YFR008W
YEL041W	YJR049C	YGL057C	YJL135W	YMR228W	YFL036W
YER017C	YMR089C	YGL090W	YOR005C	YNL311C	YKL001C
YGR180C	YJL026W	YGL174W	YIR005W	YOL108C	YDR123C
YGR240C	YMR205C	YGL195W	YFR009W	YOL111C	YOR007C
YGR261C	YBR288C	YGL254W	YGL154C	YOR269W	YLR254C
YHL027W	YJL056C	YGR057C	YKL015W	YPL003W	YPR066W
YHR024C	YLR163C	YGR074W	YKL183W	YPL209C	YBR156C
YHR056C	YDR303C	YGR208W	YKL177W	YPR046W	YJR135C
YIL103W	YKL191W	YGR229C	YGR185C	YPR051W	YEL053C
YLR238W	YDR200C	YHL044W	YKR035C	YBR107C	YDR254W
YLR456W	YPR172W	YHR193C	YDR252W	YDR080W	YDL077C
YNL007C	YIR040C	YJL006C	YML112W	YER069W	YJL071W
YNL329C	YKL197C	YJL035C	YLR316C	YER090W	YKL211C
YOR136W	YNL037C	YJL090C	YKL108W	YGL008C	YCR024C-A
YPL195W	YJL024C	YKL160W	YKL036C	YGL236C	YMR023C
YPR029C	YLR170C	YLL059C	YML011C	YGR075C	YBR152W
YBR228W	YLR135W	YLR036C	YKR065C	YHR004C	YAL009W
YDR001C	YLR270W	YLR065C	YDL149W	YKL182W	YPL231W
YDR013W	YDR489W	YLR226W	YPR161C	YLR075W	YIR012W
YDR086C	YLR378C	YLR240W	YBR097W	YNL259C	YDR270W

Fig. 5. Dataset of core interaction proteins used in the experimental work.

Further more, protein domains knowledge will be incorporated in $Matrix\_a_{150X150}$. In this implementation, ps_scan [33] is used to scan one or several patterns, rules and profiles from PROSITE against the 150 protein sequences. All frequently matching (unspecific) patterns and profiles are skipped. The ps_scan requires two compiled external programs from the PFTOOLS package : "pfscan" used to scan a sequence against a profile library and "psa2msa" which is necessary for the "-o msa" output format only. The $Matrix\_a_{150X150}$ is then used to predict the protein interaction network. Two proteins may interact if the similarity score between them is the highest.

IV. RESULTS AND DISCUSSION

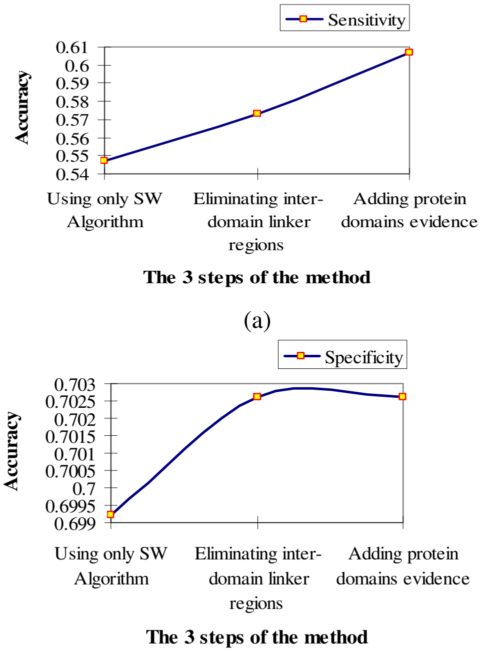
The performance of the proposed method is measured by how well it can predict the protein-protein interaction network. Prediction accuracy, whose value is the ratio of the number of correctly predicted interactions between protein pairs to the total number of interactions and non-interactions possibilities in network, is the best index for evaluating the performance of a predictor. However, approximately 20% of the data are truly interacting proteins, which leads to a rather unbalanced distribution of interacting and non-interacting cases.

To assess our method objectively, another two indices are introduced in this paper, namely specificity and sensitivity commonly used in the evaluation of information retrieval. A high sensitivity means that many of the interactions that occur in reality are detected by the method. A high specificity indicates that most of the interactions detected by the screen are also occurring in reality. Sensitivity and specificity are combined measures of true positive (tp), true negative (tn), false positive (fp) and false negative (fn) and can be expressed as:

Sensitivity (Sens) = $\dfrac{tp}{tp + fn}$, Specificity (Spec) = $\dfrac{tn}{tn + fp}$

Where, tp = interacting two protein sequences classified interacting,
fn = non-interacting two protein sequences classified interacting,
fp = interacting two protein sequences classified non-interacting,
tn = non- interacting two protein sequences classified non-interacting

Based on the above mentioned performance measures, our algorithm was able to achieve encouraging results. In Figures 6 and 7, we summarized the sensitivity and specificity results based on the three stages of the method. The figures clearly show improvement in sensitivity but not much in specificity and that's because of the big number of non-interacting possibilities.



(a)

(b)

Fig. 6. Sensitivity and specificity results.

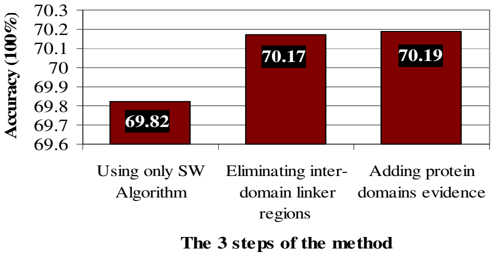


Fig. 7. Overall accuracy.

The overall performance evaluation results are summarized in Table 1.

Table 1: Overall performance evaluation

	tp	fp	tn	fn	Sens	Spec	RFP*	Accu
Similarity Measure Using SW Algorithm	82	68	15523	6677	0.5467	0.6992	0.3008	69.82
Eliminating Inter-domain Linker Regions	86	64	15597	6603	0.5733	0.7026	0.2974	70.17
Adding Structural Domain Evidence	91	59	15597	6603	0.6067	0.7026	0.2974	70.19

*Rate of False Positive (RFP), which defined as the fraction of negative test sequences that score as high as or better than the positive sequence RFP = $\dfrac{fp}{(fp + tn)}$ for $(fp + tn) > 0$ [38].

V. Conclusion

In this article we make use of both evolutionary and structural similarities among domains of known interacting proteins to predict putative protein interaction pairs. When tested on a sample data obtained from the Database of Interacting Proteins (DIP), the proposed method shows great potential and a new vision to predict protein-protein interaction. It proves that the combination of methods predicts domain boundaries or linker regions from different aspects and the evolutionary relationships would improve accuracy and reliability of the prediction as a whole. However, it is difficult to directly compare the accuracy of our proposed method because all of the other existing methods use different criteria for assessing the predictive power. Moreover, these existing methods use completely different characteristics in the prediction. One of the immediate future works is to consider the entire protein-protein interaction network and not to restrict our work on binary protein-protein interaction.

References

[1] Donaldson, I., Martin, J., de Bruijn, B., Wolting, C., Lay, V., Tuekam, B., Zhang, S., Baskin, B., Bader, G.D., Michalickova, K., Pawson, T., and Hogue, C.W. (2003). PreBIND and Textomy--mining the biomedical literature for protein-protein interactions using a support vector machine. BMC Bioinformatics. 4:11.

[2] Gharakhanian, E., Takahashi, J., Clever, J., and Kasamatsu, H. (1998). In vitro Assay for Protein-Protein Interaction: Carboxyl-Terminal 40 Residues of Simian Virus 40 Structural Protein VP3 Contain a Determinant for Interaction with VP1. PNAS 85(18):6607-6611.

[3] Bartel, P.L. and Fields, S. (eds) (1997). The yeast two-hybrid system. In Advances in Molecular Biology. Oxford University Press, New York.

[4] Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M., and Seraphin, B. (1999). A generic protein purification method for protein complex characterization and proteome exploration. Nature Biotechnology. 17(10):1030-1032.

[5] Selbach, M. and Mann, M. (2006). Protein interaction screening by quantitative immunoprecipitation combined with knockdown (QUICK). Nature Methods. 3:981-983.

[6] L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie, D. Eisenberg, "The Database of Interacting Proteins: 2004 update", Nucleic Acids Res. 2004 Jan 1;32, pp:449-51.

[7] Mewes,H.W. et al. (2004) MIPS: analysis and annotation of proteins from whole genomes. Nucleic Acids Res., 32 (Database issue), D41–D44.

[8] Peri,S. et al. (2004) Human protein reference database as a discovery resource for proteomics. Nucleic Acids Res., 32 (Database issue), D497–D501.

[9] Espadaler,J. et al. (2005) Detecting remotely related proteins by their interactions and sequence similarity. Proc. Natl Acad. Sci. USA, 102, 7151–7156.

[10] Marcotte,E. et al. (1999) Detecting protein function and protein–protein interactions from genome sequences. Science, 285, 751–753.

[11] Dandekar,T. et al. (1998) Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem. Sci., 23, 324–328.

[12] M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates, "Assigning protein functions by comparative genome analysis: protein phylogenetic profiles," In the proceedings of National Academy of Sciences, USA, vol. 96, 1999, pp: 4285–4288.

[13] A. Szilàgyi, V. Grimm, A. K Arakaki and J. Sholnick, "Prediction of physical protein-protein interactions", Phys. Biol. 2005, pp: 1-16.

[14] E. M. Marcotte, M. Pellegrini, M. J. Thompson, T. O. Yeates, and D. Eisenberg, "A combined algorithm for genome-wide prediction of protein function," Nature, vol. 402, 1999, pp: 83–86.

[15] F. Pazos and A. Valencia, "Similarity of phylogenetic trees as indicator of protein-protein interaction," Protein Engineering, vol. 14(9), 2001, pp: 609- 614.

[16] J. Enright, I. N. Ilipoulos, C. Kyrpides, and C. A. Ouzounis, "Protein interaction maps for complete genomes based on gene fusion events," Nature, vol. 402, 1999, pp: 86–90.

[17] D. Eisenberg, E. M. Marcotte, I. Xenarios, and T. O. Yeates, "Protein function in the post-genomic era," Nature, vol. 405, 2000, pp: 823-826.

[18] J. Wojcik and V. Schachter, "Protein-Protein interaction map inference using interacting domain profile pairs," Bioinformatics, vol. 17, 2001, pp: S296-S305.

[19] W. K. Kim, J. Park, and J. K. Suh, "Large scale statistical prediction of protein-protein interaction by potentially interacting domain (PID) pair," Genome Informatics, vol. 13, 2002, pp: 42-50.

[20] S. K. Ng, Z. Zhang, and S. H. Tan, "integrative approach for computationally inferring protein domain interactions," Bioinformatics, 19, 2002, pp: 923-929.

[21] S. M. Gomez, W. S. Noble, and A. Rzhetsky, "Learning to predict protein-protein interactions from protein sequences," Bioinformatics, 19, 2003, pp: 1875-1881.

[22] C. Huang, F. Morcos, S. P. Kanaan, S. Wuchty, A. Z. Chen, and J. A. Izaguirre, "Predicting Protein-Protein Interactions from Protein Domains Using a Set Cover Approach", IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 4, no. 1, 2007.

[23] T. Pawson and P. Nash, "Assembly of cell regulatory systems through protein interaction domains," Science, vol. 300, 2003, pp: 445-452.

[24] N. M. Zaki, S. Deris and H. Alashwal, "Protein Protein Interaction Detection Based on Substring Sensitivity Measure", International J. of Biomedical Sci., Vol. 1, 2006, pp: 148-154.

[25] P. Aloy and R. B. Russell, "InterPreTS: protein interaction prediction through tertiary structure", Bioinformatics, 19, 2003, pp: 161–162.

[26] L. Lu, "Multiprospector: an algorithm for the prediction of protein–protein interactions by multimeric threading", Proteins, 49, 2002, pp: 350–364.

[27] J. Espadaler, O. Romero-Isart, R. M. Jackson and B. Oliva1 "Prediction of protein–protein interactions using distant conservation of sequence patterns and structure relationships", Bioinformatics, Vol. 21 no. 16 2005, pp: 3360–3368.

[28] O. Keskin, "A new, structurally nonredundant, diverse data set of protein–protein interfaces and its implications", Protein Sci., 13, 2004, 1043–1055.

[29] T. Smith and M. Waterman, "Identification of common molecular subsequences", J. Mol. Bio., 147, 1981, pp: 195-197.

[30] H. Saigo, J. Vert, N. Ueda and T. Akutsu, "Protein homology detection using string alignment kernels," Bioinformatics, Vol. 20 no. 11, 2004, pp: 1682-1689.

[31] A. Bairoch and R. Apweiler, "The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000", Nucleic Acids Res., 28, 2000, pp: 45–48.

[32] M. Suyama and O. Ohara, "DomCut: prediction of inter-domain linker regions in amino acid sequences", Bioinformatics 19, 2003, pp: 673-674.

[33] A. Gattiker, E. Gasteiger, A. Bairoch, "ScanProsite: a reference implementation of a PROSITE scanning tool", Applied Bioinformatics, 1(2), 2002, pp: 107-108.

[34] I. Xenarios, L. Salwínski, X. J. Duan, P. Higney, S. Kim and D. Eisenberg, "DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions" , Nucleic Acids Research, Oxford University Press, vol. 30, 2002, pp: 303-305.

[35] C. M. Deane, L. Salwinski, I. Xenarios, and D. Eisenberg, "Protein interactions: two methods for assessment of the reliability of high throughput observations," Molecular & Cellular Proteomics, vol. 1(5), 2002, pp: 349-56.

[36] W. R. Pearson, "Rapid and sensitive sequence comparisons with FASTAP and FASTA Method", Enzymol, 183, 1985, pp: 63.

[37] Marketa, Z. and Jeremy, O. B., "Understanding bioinformatics", Garland Science, Taylor & Francis Group, LLC, 2008.

[38] N. M. Zaki, S. Deris, and R. M. Illias, "Performance analysis of string kernel on protein remote homology detection", Applied Bioinformatics, vol 4, (1), 2005, pp: 45-52.