ISBN xxx-x-xxxxx-xxx-x
Proceedings of 2009 International IACSIT Conference on Machine Learning and Computing
Perth, Australia, July 10-12, 2009

Domain Linker Region Knowledge Contributes to Protein-protein Interaction Prediction

Nazar Zaki

[1] College of Information Technology, UAE University, Al Ain 17551, UAE

Abstract. Protein-protein interaction has proven to be a valuable biological knowledge and starting point for understanding how the cell internally works. In this paper, we propose a method for PPI prediction using only the primary structure information of protein sequence. The method is developed based on inter-domain linker region knowledge and a combination of pairwise similarity and support vector machine techniques. Two protein sequences may interact by the mean of the similarities between the domain-linker regions they contain. The method is tested in different datasets from yeast *saccharomyces cerevisiae* protein interaction and showed higher specificity, sensitivity and accuracy than the PIPE, MLE and decision forest methods.

Keywords: Protein-protein interaction, pairwise alignment, support vector machine, inter-domain linker region

1. Introduction

The term protein-protein interaction (PPI) refers to the association of protein molecules and the study of these associations from the perspective of biochemistry, signal transduction and networks. The prediction of PPI is one of the fundamental problems in computational biology since it can aid significantly in identifying the function of newly discovered proteins. Abnormal protein-protein interactions have implications in a number of neurological disorders; include Creutzfeld-Jacob and Alzheimer's diseases [1]. To solve this problem, many impressive experimental techniques have been developed to predict the physical interactions which could lead to the identification of the functional relationships between proteins. Most of the recent computational methods developed such as the Association Method (AM) [2], Maximum Likelihood Estimation (MLE) [3], Maximum Specificity Set Cover (MSSC) [4] and Domain-based Random Forest [5] have employed domain knowledge to predict the PPI. The motivation behind this employment is that molecular interactions are typically mediated by a great variety of interacting domains [6].

A very recent tool termed PIPE (Protein-Protein Interaction Prediction Engine) was also developed [1]. PIPE is based on the assumption that some of the interactions between proteins are mediated by a finite number of short polypeptide sequences. These sequences are typically shorter than the classical domains, and are used repeatedly in different proteins and contexts within the cell.

Most of the mentioned methods have common limitations:

- They are based on previously identified domains.
- Identifying domain is a long and computationally expensive process.
- They all focus only on domain structure.
- They are not universal because the accuracy and reliability of these methods depend on the domain information of the protein partners.
- They often have limited abilities to detect novel interactions and to differentiate them from false positives. A high rate of false negatives is another disadvantage associated with most of these methods.

In this paper, we introduce a simple yet novel method to predict PPI based on domain-linker region knowledge and using only protein primary structure. Two protein sequences may interact by the mean of the

similarities between domain-linker regions they contain. For each sequence the difference in amino acid composition between domain and linker regions is calculated. Amino acids with linker score less than the set threshold value are eliminated from the protein sequence of interest. By doing this step, we are actually downsizing the protein sequences to shorter ones with only domain-linker regions and without losing their generality. Pairwise technique is then used to measure the similarity between inter-domain linker regions. Two proteins are classified interacting if the inter-domain linker regions they contain produce similar scores when compared to a long subsequence of amino acids.

2. Method

The PPI based on domain-linker regions similarity (PPI-DLR) method consists of three major steps:

- Domain-linker region prediction: representing each protein sequence by the domain-linker regions it contains.
- Feature extraction step: representing each protein sequence by a vector of pairwise similarities against long subsequences of amino acids.
- Classification: taking as a kernel the dot product between these vector representations to be used in conjunction with SVM.

In the proceeding sections, we describe these steps.

2.1. Domain-linker region prediction

The first step of our algorithm is to predict inter-domain linker regions solely by amino acid sequence information. Our intention here is to identify all the inter-domain linker regions from the protein sequences of interest. By doing this step, the protein sequence will be shorter with only inter-domain linker regions, which may produce better pairwaise 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 [8]. DomCut developed by Suyama et al [7] is employed to predict linker regions among functional domains based on the difference in amino acid composition between domain and linker regions. Following [7], we defined the linker index S_i for amino acid residue i and it is calculated as $S_i = -\ln\left[\frac{f_i^{Linker}}{f_i^{Domain}}\right]$, 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 exists in a linker region. As shown in Fig 1, a threshold value is needed to separate linker regions. Amino acids with linker score less than the set threshold value will be eliminated from the protein sequence of interest. This step will result in significant downsizing of the protein sequence without losing its generality.

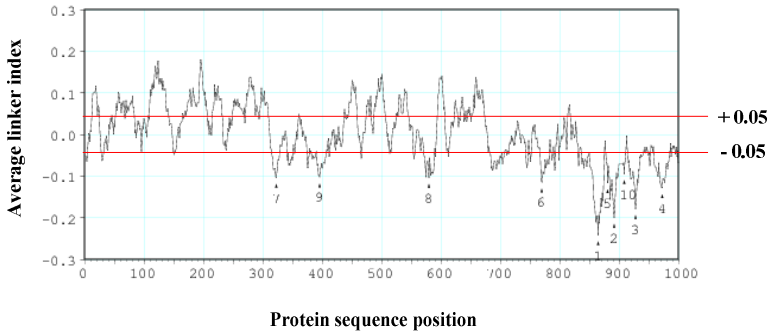

Fig. 1: An example of linker preference profile generated using Domcut. In this case, linker region less than a threshold value of 0.05 and more than – 0.05 will be eliminated from the protein sequence.

2.2. Protein feature extraction

In the feature extraction step, we represent a protein sequence by a fixed-length of feature vectors. Each coordinate of this feature vector is typically the E-value of the Smith–Waterman (SW) algorithm as implemented in Fasta [9]. The score is calculated by comparing each protein sequence in the dataset to subsequences of amino acids created by shifting a large window over the protein training sequences.

2.3. Smith-Waterman score

The Smith-Waterman score $SW(s_0, s_1)$ between protein sequences s_0 and the subsequence s_1 is the score of the best local alignment with gaps between the two protein sequences, computed by the SW dynamic programming algorithm [9].

Following Saigo et at. [10], let us denote by π a possible local alignment between protein sequences s_0 and the subsequence s_1, defined by a number n of aligned residues, and by the indices $1 \le i_1 < ... < i_n \le |s_0|$ and $1 \le j_1 < ... < j_n \le |s_1|$ of the aligned residues in s_0 and s_1 respectively. Let us also denote by $\prod(s_0, s_1)$ the set of all possible local alignments between s_0 and s_1, and by $p(s_0, s_1, \pi)$ the score of the local alignment $\pi \in \prod(s_0, s_1)$ between s_0 and s_1, the Smith-Waterman score $SW(s_0, s_1)$ between s_0 and s_1 can be written as $SW(s_0, s_1) = \max_{\pi \in \prod(s_0, s_1)} p(s_0, s_1, \pi)$.

Using a shifting window over the concatenated sequences of the training set may lead to generating a subsequence comprises of the end of one sequence and the beginning of the next sequence, however, this is not a problem since all protein sequence of interest score against the same subsequence.

We believe that the feature extraction is particularly significant step in our method to predict PPI. More meaningful features yield better generalization performance [11].

2.4. Classification Step

The problem is basically formulated as a two-class classification problem: both training and testing sets contain protein pairs belong to either "interacted" or "non-interacted". This representation is combined with support vector machine (SVM) to classify between the two sets. The overall algorithm is illustrated in Fig. 2.

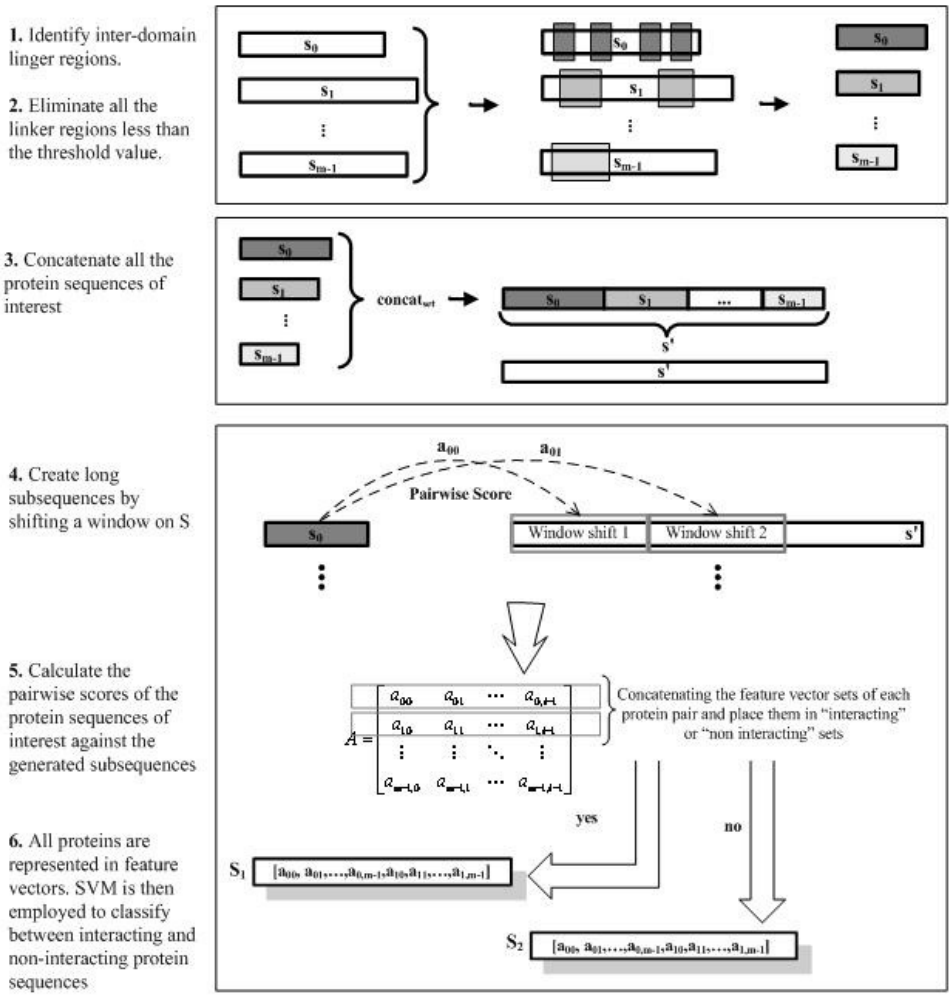

Fig. 2: Illustration of the PPI-DLR algorithm.

3. Experimental Work and Results

In our first experimental work, we assess the recognition ability of our method to classify between 100 interacted protein pairs (157 proteins) and 100 non-interacted protein pairs (77 proteins). The dataset was randomly selected by Sylvain et. al [1] from the Database of Interacting Proteins (DIP) [3]. 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 dataset was used to evaluate PIPE's accuracy [1]. It was generated from the yeast protein interaction literature for which at least three different lines of experimental evidence supported the interaction.

The experimental work starts by predicting the inter-domain linker regions within the 234 protein sequences. This step downsized the protein sequence tremendously. The downsized proteins are then used to create a long string of amino acids by concatenating all of the 234 protein sequences available in the dataset. By choosing a large different window sizes, we were able to generate different large subsequences. All of the downsized protein sequences in the dataset were scored against the generated subsequences using Smith–Waterman (SW) algorithm. The SW [12] has undergone two decades of empirical optimization in the field of bioinformatics. Thus, considerable prior knowledge is implicitly incorporated into the pairwise sequence similarity scores and hence into the PPI-DLR vector representation. For instance, if we have a protein sequence *s* then the corresponding score will be $F_s = f_{s_0}, f_{s_1}, .., f_{s_{m-1}}$ where $m-1$ is the total number of proteins and f_{s_i} is the E-value of the SW score between sequence s and the i^{th} subsequence. In this case, the default parameters are used: gap opening penalty and extension penalties of 13 and 3, respectively, and the BLOSUM 50 matrix. Based on prior biological knowledge about the interaction information between proteins, the feature vectors of two "interacted" proteins s_0 and s_1 are concatenated and added to the positive set, and the "non-interacting" proteins are also concatenated and added to the negative set.

Following the preparation of the dataset, we employed Gist SVM to discriminate between the "interacted" and "non-interacted" protein pairs using hold-one-out cross-validation to measure the accuracy. The Gist SVM software is implemented by Noble et al. and it is available at http://www.cs.columbia.edu/compbio/svm. In all experiments, Gaussian Radial Basis Function kernel (RBF kernel) was used. The RBF kernel allows pockets of data to be classified which is more powerful way than just using a linear dot product [13]. The function has the form $K(x, x_i) = e^{-\gamma \|x - x_i\|^2}$,where $x, x_i \in X$ and $\gamma > 0$. In all of the experimental work, the scaling parameter γ was set to 0.001.

The accuracies of our predictions are measured by specificity (SP) and sensitivity (SN). The specificity is defined as the ratio of the number of matched interactions between the predicted set, and the observed testing set, over the total number of predicted interactions. The sensitivity is defined as the ratio of the number of matched interactions, to the total number of observed interactions in the testing set [4]. The Receiver Operating Characteristics (ROC) and overall accuracy are also used. In Table 1, we record the classification results between the 100 interacted protein pairs and the 100 non-interacted protein pairs. A window of size 5000 produces the most accurate result. The average SN, SP, ROC and overall accuracy are 0.9667, 0.9338, 0.9864 and 0.9502, respectively.

Beside the accuracy superiority, PPI-DLR has two more advantages over PIPE. Firstly the PIPE method is computationally intensive and the evaluation of PIPE performance over the same dataset took around 1,000 hours of computation time compared to few minutes using PPI-DLR. Secondly it is also mentioned by the PIPE authors that their method is expected to be weak if it is used for detecting novel interactions among genome wide large-scale data sets. This is not the case for PPI-DLR as it's further tested on a large-scale data (in the proceeding section).

In the second experiment we furthermore split the 100 interacted protein pairs into 2 sets A (50 pairs) and B (50 pairs). We also split the 100 non-interacted protein pairs into 2 sets C (50 pairs) and D (50 pairs). We then combined A with C to create a training dataset and B with D to create a testing dataset. Similar set up as mentioned in the earlier experiment was followed. In Table 2 we show SP and SN calculated using different window sizes. The average SN, SP, ROC and overall accuracy are 0.9114, 0.6905, 0.8663 and 0.801, respectively.

Table 1. SP, SN, ROC and overall accuracy scores recorded from testing PPI-DLR on a dataset of 200 protein pairs based on different window size value.

WINDOW SIZE	SN	SP	ROC	ACCURACY
20000	0.90	0.90	0.959	0.900
19000	0.96	0.88	0.979	0.920
18000	1.00	0.96	0.995	0.980
17000	0.95	0.92	0.976	0.935
16000	0.88	0.91	0.973	0.895
15000	1.00	0.96	0.996	0.980
14000	0.90	0.97	0.977	0.935
13000	1.00	0.94	0.992	0.970
12000	0.90	0.97	0.978	0.935
11000	0.97	0.98	0.986	0.975
10000	1.00	0.95	0.998	0.975
9000	1.00	0.96	0.997	0.980
8000	0.98	0.97	0.986	0.975
7000	1.00	0.93	0.996	0.965
6000	1.00	0.96	0.998	0.980
5000	1.00	0.97	0.999	0.985
4000	0.98	0.96	0.994	0.970
3000	1.00	0.94	0.995	0.970
2000	0.97	0.95	0.993	0.960
1000	0.95	0.86	0.988	0.905
500	0.96	0.77	0.960	0.865

Table 2. SP and SN scores recorded from testing PPI-DLR on a testing dataset of 100 protein pairs based on different window size value

WINDOW SIZE	SN	SP	ROC	ACCURACY
20000	0.96	0.76	0.877	0.86
19000	0.66	0.84	0.871	0.75
18000	0.48	0.82	0.794	0.65
17000	0.62	0.70	0.788	0.66
16000	0.94	0.70	0.852	0.82
15000	0.62	0.78	0.820	0.70
14000	0.98	0.74	0.869	0.86
13000	0.98	0.76	0.885	0.87
12000	0.98	0.70	0.846	0.84
11000	0.98	0.74	0.873	0.86
10000	1.00	0.66	0.885	0.83
9000	0.94	0.74	0.845	0.84
8000	1.00	0.72	0.884	0.86
7000	1.00	0.70	0.874	0.85
6000	1.00	0.70	0.870	0.85
5000	1.00	0.66	0.863	0.83
4000	1.00	0.66	0.897	0.83
3000	1.00	0.56	0.887	0.78
2000	1.00	0.60	0.925	0.80
1000	1.00	0.52	0.905	0.76
500	1.00	0.44	0.886	0.72

In the third experimental work, we assess the recognition ability of our method on the dataset created by Xue-Wen et al. [5]. He initially obtained 15,409 interacting protein pairs in the yeast organism from DIP, 5719 pairs from Deng et al. [3] and 2238 pairs from Schwikowski et al. [14]. The datasets were then combined by removing the overlapping interaction pairs and excluded the pairs where at least one of the proteins has no domain information. Finally, 9834 protein interaction pairs remained among 3713 proteins,

and it was separated evenly (4917 pairs each) into training and testing datasets. Since non-interacting protein data are not available, the negative samples are randomly generated. A protein pair is considered to be a negative sample if the pair does not exist in the interaction set. A total of 8000 negative samples were generated and also separated into two halves. Both final training and testing datasets contain 8917 samples, 4917 positive and 4000 negative samples.

For comparison purpose, we tested two other state-of-the-art sequence based methods, maximum likelihood estimation (MLE) developed by Deng et al. [3] and domain-based random forest of decision trees, developed by Xue-Wen et al. [5]. Results of the primary experiment are summarized in Figure 3. The figure also shows performance comparison between PPI-DLR and other two state-of-the-art sequence based methods; Maximum Likelihood Estimation (MLE) and Domain-based random forest of decision trees. Higher SP, SN and overall accuracy correspond to more accurate PPI detection performance. Using any of these performance measures, the PPI-DLR method performs better than the other two methods.

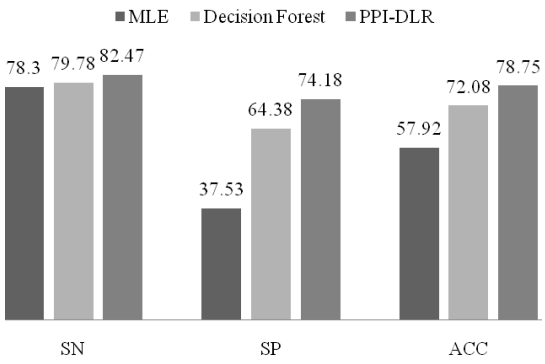

Fig. 3: SP, SN and accuracy scores recorded from testing PPI-DLR on a testing dataset of 4917 interacted proteins and 4000 non-interacted proteins based on window size of 5000.

4. Discussion and Conclusion

The method presented here is based on the assumption that two proteins may interact if the inter-domain linker regions they contain are similar. It is understood that, this assumption excludes the applicability of interactions of proteins which are not similar or evolutionary related to each other. However, the main contribution of this paper is to show that pairwise sequence comparison in conjunction with domain linker knowledge can be extremely powerful. Moreover, we are motivated by the fact that SW alignment score provides a relevant measure of similarity between proteins. The experimental results have shown that PPI-DLR method applied on different datasets from the yeast saccharomyces cerevisiae protein interaction literature can predict PPIs with higher specificity, sensitivity and accuracy than the PIPE, MLE and decision forest methods.

The remarkable accuracy of our method follows from the use of two widely used and powerful algorithms. On one hand, the SVM algorithm is based on a sound mathematical framework and much of its power comes from its criterion for selecting a separating hyperplane that maintains a maximum margin from any point in the training set [11]. On the other hand, SW scores have been developed to quantify the similarity of biological sequences. Their parameters have been optimized over the past two decades to provide relevant measures of similarity between sequences and they now represent core tools in computational biology. In the future we will combine more accurate domain linker region methods. Our technique could also be used to solve different computational biology problems such as protein remote homology detection.

5. References

[1] Sylvain, P. Frank, D. Albert, C. Jim, C. Alex, D. Andrew, E. Marinella, G. Jack, G. Mathew, J. Nevan, K. Xuemei, L. Ashkan, G. (2006). PIPE: a protein-protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs. *BMC Bioinformatics, 7*: 365.

[2] Sprinzak, E. Margalit, H. (2001). Correlated sequence-signatures as markers of protein-protein interaction. *J. Mol Biol., 311*: 681–692.

[3] Deng, M. Mehta, S. Sun, F. Cheng, T. (2002). Inferring domain-domain interactions from protein-protein interactions. *Genome Res., 12*, 1540-1548

[4] Huang, T. W. Tien, A. C. Huang, W. S. Lee, Y. C. Peng, C. L. Tseng, H. H. Kao, C. Y. Huang, C. Y. (2004). POINT: a database for the predic-tion of protein-protein interactions based on the orthologous interactome. *Bioinformatics, 20*: 3273-3276

[5] Xue-Wen, C. Mei, L. (2005). Prediction of protein–protein interac-tions using random decision forest framework. *Bioinformatics, 21*: 4394–4400.

[6] Pawson, T. Nash, P. (2003). Assembly of cell regulatory systems through protein interaction domains. *Science, 300*: 445-452.

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

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

[9] Pearson, W. R. Lipman, D. L. (1988). Improved tools for biological sequence comparison. *PNAS, 85*: 2444-2448.

[10] Saigo, H. Vert, J. Ueda, N. Akutsu, T. (2004). Protein homology detection using string alignment kernels. *Bioinformatics, 20* :1682-1689.

[11] Zaki, N. M. Deris, S. Illias, R. (2004). Feature Extraction for Protein Homologies Detection Using Markov Models Combining Scores. *Inter. J. on Computational Intelligence and Applications, 4*:1-12.

[12] Smith, T. Waterman, M. (1981). Identification of common molecular subsequences. J. Mol. Bio., 147: 195-197.

[13] Zaki, N. M. Deris, S. Alashwal, H. (2006). Protein-protein Interaction Detection Based on Substring Sensitivity Measure. *Inter. J. of Biomedical Sciences, 1*:148-154.

[14] Schwikowski, B. (2000). A network of protein–protein interactions in yeast. *Nat. Biotechnology, 18*: 1257–1261.