2009 International Conference on Machine Learning and Computing
IPCSIT vol.3 (2011) © (2011) IACSIT Press, Singapore

Domain Linker Region Knowledge Contributes to Protein-protein Interaction Prediction

Nazar Zaki, Piers Campbell

College of Information Technology, UAE University, Al Ain 17551, UAE

Abstract. Protein-protein interaction has proven to be a valuable piece of biological knowledge and a starting point for understanding the internal workings of the cell. In this paper, we propose a novel method for protein-protein interaction prediction using only the primary structural information of the 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 means of the similarities between the domain-linker regions they contain. The method is tested on different datasets from yeast saccharomyces cerevisiae protein interaction and showed higher specificity, sensitivity and accuracy than the maximum likelihood estimation, protein-protein interaction prediction engine 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 as it can aid significantly in identifying the function of newly discovered proteins. Abnormal protein-protein interactions have implications in a number of neurological disorders [1]. Most of the recent computational methods developed such as Protein-Protein Interaction Prediction Engine (PIPE) [1], 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 these employments is that molecular interactions are typically mediated by a great variety of interacting domains [6]. However, identifying domains is a computationally expensive process.

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 means 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. The Pairwise technique is then used to measure the similarity between inter-domain linker regions. Two proteins are classified as 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 which are illustrated in Fig. 1. In the subsequent sections, we describe these steps.

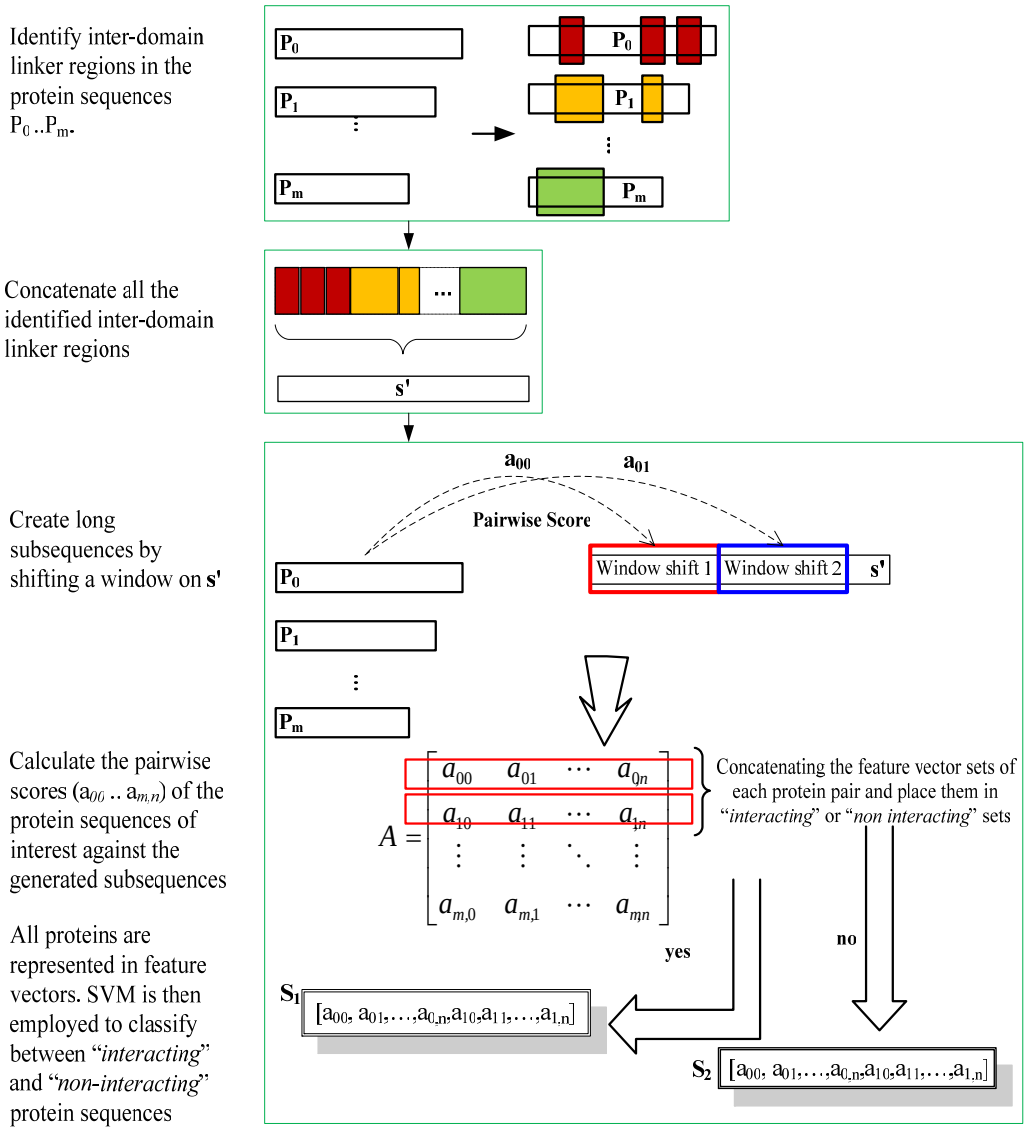

Fig. 1: Illustration of the PPI-DLR algorithm.

2.1. Domain-linker region prediction

The first step of our algorithm is to predict inter-domain linker regions solely from amino acid sequence information. Our intention here is to identify all the inter-domain linker regions from the protein sequences of interest. By performing this step, the protein sequence will be shorter with only inter-domain linker regions, which may produce improved 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 [7]. DomCut developed by Suyama et al. [8] is employed to predict linker regions among functional domains based on the difference in amino acid composition between domain and linker regions. Following [8], we defined the linker index Si 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. A negative value of Si indicates that the amino acid exists in a linker region. A threshold value is needed to separate linker regions. Amino acids with a 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 loss of generality.

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].

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

We believe that the feature extraction is particularly significant step in our method to predict PPI, asmore meaningful features yield better generalization performance [10].

2.4. Classification Step

The problem is basically formulated as a two-class classification problem; both training and testing sets contain protein pairs belonging to either “interacted” or “non-interacted” sets. This representation is combined with support vector machine (SVM) to classify between the two sets.

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 commences by predicting the inter-domain linker regions within the 234 protein sequences, and 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 range of 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 [10] 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. 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 a more powerful technique than simply using a linear dot product [11].

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 good performance, PPI-DLR has two further 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 just a few minutes using PPI-DLR. Secondly, it is stated by PIPE's 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 demonstrated in our test on a large-scale dataset, as described below.

In the second experiment we further 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. A similar process as mentioned in the earlier experiment was followed. The average SN, SP, ROC and overall accuracy achieved are 0.9114, 0.6905, 0.8663 and 0.801, respectively.

In the third experimental work, we assessed 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. [12]. The datasets were then combined by removing the overlapping interaction pairs and excluding the pairs in which at least one of the proteins has no domain information. Ultimately, 9834 protein interaction pairs remained among 3713 proteins, and the dataset was separated evenly (4917 pairs each) into training and testing datasets. As 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 comparative purposes, 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 Fig. 2. The figure also shows a 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. A higher value for SP, SN and overall accuracy corresponds to greater accuracy in PPI detection performance. Using any of these performance measures, the PPI-DLR method performs significantly better than the other methods.

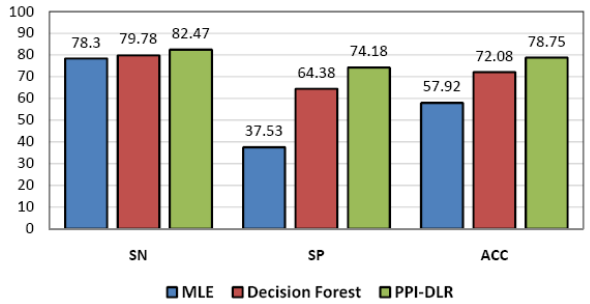

Fig. 2: SP, SN and accuracy scores recorded from testing PPI-DLR on a testing dataset of 4917 interacted proteins and 4000 non-interacted proteins based on window size of 5000.

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. We are motivated by the fact that SW alignment score provides a relevant measure of similarity between proteins. This measure was crucial in our recently published methods [13], [14]. The main improvement of this study over our previous method termed PPI_PS [13] is that we recorded the sensitivity of the protein sequences of interest against only the domain-linker regions and not the actual protein sequences. This is a more efficient technique, comparing protein sequences to shorter amino acid sequences and has the additional benefit of not compromising generality.

Most of the existing work including PPI-PS has employed domain knowledge to predict PPI which we intentionally tried to avoid in this study. The PPI-PS approach compared protein sequences to a long

subsequence of amino acids created by concatenating the actual protein sequences in the training data. These subsequences obviously contain protein domain regions which could make the classification easier.

The experimental results have shown that the PPI-DLR method when applied to 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. We further demonstrated that linker region knowledge on its own is a sufficient source of biological information and could assist in PPI prediction. The results published in [14] have highlighted the great potential in using linker regions to predict PPI when combined with domain knowledge. However, the method was applied to a set of proteins known to exhibit only unique pairwise interactions which is not sufficient to allow any definitive conclusions to be drawn.

It is important to mention that the idea of representing protein sequence via its similarity to a collection of other sequences is not novel. Liao et al. [15] Zaki et al. [16] have done similar work in their algorithm to detect protein remote homology. It is our intention to combine more accurate domain linker region methods in our future work.

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] 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.

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

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

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

[11] 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.

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

[13] Zaki, N.M. Lazarova-Molnar., L. El-Hajj, W., Campbell, P. (2009). Protein-protein interaction based on pairwise similarity. *BMC Bioinformatics, 10*:150.

[14] Zaki, N.M. (2009). Protein-protein interaction prediction using homology and inter-domain linker region information. *International Conference of Systems Biology and Bioengineering*, Imperial College, LNCS, Springer. 635-645.

[15] Liao, L. Noble, W.S. (2003). Combining Pairwise Sequence Similarity and Support Vector Machines for Detecting Remote Protein Evolutionary and Structural Relationships. *J. Comp. Biol. 10:*857-868.

[16] Zaki, N.M. Deris, S. Illias, R.M. (2005). Application of string kernels in protein sequence classification. *Applied Bioinformatics*. *4*:45—52.