International Journal of Biological, Life and Agricultural Sciences
ISSN: 2415-6612
Vol:5, No:3, 2011

Improving Protein-Protein Interaction Prediction by Using Encoding Strategies and Random Indices

Essam Al-Daoud

***Abstract*—**A New features are extracted and compared to improve the prediction of protein-protein interactions. The basic idea is to select and use the best set of features from the Tensor matrices that are produced by the frequency vectors of the protein sequences. Three set of features are compared, the first set is based on the indices that are the most common in the interacting proteins, the second set is based on the indices that tend to be common in the interacting and non-interacting proteins, and the third set is constructed by using random indices. Moreover, three encoding strategies are compared; that are based on the amino asides polarity, structure, and chemical properties. The experimental results indicate that the highest accuracy can be obtained by using random indices with chemical properties encoding strategy and support vector machine.

***Keywords*—**protein-protein interactions, random indices, encoding strategies, support vector machine.

I. Introduction

ALMOST every cellular process relies on interacting of two or more proteins in order to accomplish a specific task, therefore; predicting protein–protein interactions (PPI) represents a crucial step toward deciphering the biological processes, such as protein synthesis, signaling pathways, DNA replication, cell adhesion and regulation of metabolic. Drug discovery and understanding the functional roles of un-annotated protein are another area where protein–protein interaction prediction plays an important role. Moreover; Protein–protein interactions is only the way which allows to the cells to communicate with the internal and the external parts. High throughput experimental methods including two-hybrid screens, affinity purification, mass spectrometry, protein chip and hybrid approaches have been used in an attempt to discover protein-protein interactions pairs. However, experimental methods are time-consuming, expensive and exhibit high false positive and false negative rates. Therefore the current experimental methods are covering only a fraction of the complete protein-protein interaction networks [1, 2].

Protein-protein interaction database are generated by the experimental methods, and then collected together in specialized biological databases that allow the interactions to be searched, compared and studied further. The first of these databases was Database of Interacting Proteins (DIP) at UCLA. DIP contains 23146 proteins, 274 organisms and 71205 interactions. For instance, the number of the *Saccharomyces cerevisiae* (baker's yeast) interactions that are stored in DIP is 23855, the number of the *Drosophila melanogaster* (fruit fly) interactions pairs is 22975 and the number of the *Homo sapiens* (Human) interactions pairs is 3350 [3]. A large number of further database collections have been created such as BIND, MIPS, GRID, HIV Interaction DB and STRING [4] . Fig. 1 shows interactions that are generated by STRING database [5].

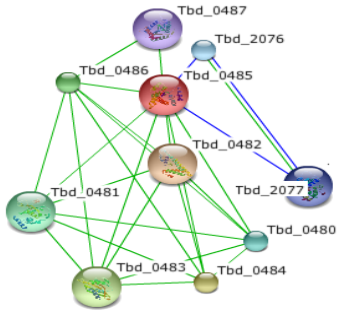

Fig. 1 PPI that are generated by STRING database

Unfortunately, PPI databases are contradictory and incomplete, the main reason is that the interacting proteins pairs (positive sets) are much less than the proteins pairs that do not interact (negative set). For example the estimated number of the *Yeast* positive set is roughly 80,000 pairs, while the number of the unrepeated pairs that are produced by 6000 proteins in the *Yeast* is about 18 million, which means that the estimated number of the *Yeast* negative set is about 17 million pairs [6].

The computational techniques are the alternative methods to predict protein-protein interactions. Several computational approaches have been applied to predict protein-protein interaction such as: homology modeling, interolog mapping, statistical potentials, threading of structural complexes, correlated mutations, and docking methods. In this paper a novel features are generated and compared to enhance the prediction accuracy.

E. Al-Daoud is with faculty of Science and Information Technology, Computer Science Department, Zarka Private University, Jordan (Tel +962-796680005, e-mail:essamdz@zpu.edu.jo).

II. Related Work

Several computational techniques have been developed. An important category is based on integration the PPI data source. Lan et al. [6] integrated direct and indirect genomic and proteomic data to construct a decision tree, and then the decision tree is used to predict protein pairs. Jansen et al. [7] used naïve Bayes and a fully-connected Bayesian network to combine direct and indirect data sources. QI et al. [8] constructed a random forest (several decision trees) from a training set by using direct and indirect information about protein-protein interactions. The resulting forest and the classification algorithms are used to classify the protein pairs. Yanjun *et al.* [9] collected 163 features from 17 different data source; the suggested sources can be divided into four types: Direct experimental, indirect data, based on the proteins sequences and functional properties of proteins. Li et al. [10] showed that conditional random fields outperformed the conventional classification methods, where 1276 non-redundant hetero complex chains are used as training and testing set. Jansen et al. [6] developed an approach based on Bayesian networks, their method naturally weights and combines the genomic features, and it can integrate often noisy experimental interaction data sets. The main disadvantage of this category is the difficulties of collecting the full protein-protein interaction data from all the sources, and the pre-knowledge about the proteins is not always available. Another important category have been used to solve prediction of PPI is based on protein sequences or DNA sequences. Espadaler et al. [11] predicted the protein interaction by using the similarities of the structures and the conservation history of interacting proteins pairs. Wang et al. [12] used spatial sequence profile, sequence information entropy and evolution rate, they suggested two stages models: support vector machine and Bayesian discrimination by considering the predicted labels of spatial neighbor residues. Wang et al. [13] suggested a new method based only on DNA sequences, they selected a suitable negative set to deal with the imbalance problem, the suggested method applied on *Plasmodium falciparum* and *Escherichiacoli.* Bakar et al. [14] predicted protein-protein interactions by using multiple independent fuzzy systems and the similarity of protein secondary structures. Their method consists of two main steps: similarity score computation, and similarity classification. The first step consists of three tasks: multiple-sequence alignment, secondary structure prediction and similarity measurement. In the classification stage; 1029 proteins of *Saccharomyces cerevisie* (baker's yeast) are used to train and test the generated first order Sugeno fuzzy system. Yaveroglu [15] predicted protein-protein interaction by using phylogenetic profiles, two proteins are combined by checking existence of homologs in different species and fitting the combined profile into a statistical model.

III. Construction of the Frequency Array

In this section a new general algorithms are suggested to extend the frequency vector that is introduced by [16]. Algorithm 1 can be used with any length of the amino acids subsequences.

Algorithm 1: Frequency Array construction

Repeat

$i = i+1$

$j = \sum_{k=0}^{m} AA_{i+k} * s^{m-k} + 1$

$freqVec_j = freqVec_j + 1$;

if $i+m+1 = length(AA)$ then stop

where *freqVec* is the frequency vector of an amino acid, m is the length of the amino acid subsequence, *s* depends on the representation system and AA is an amino acid sequence that is represented by one of the following three encoding strategies:

1-Based on the Amino acid polarity: the amino acid is identified as polar or non-polar. A further sub-classification of acidic-polar when the side chain contains a carboxylic acid, and basic-polar when the side chain contains an amino group [17]:

Non polar {'G', 'A', 'V', 'I', 'L','Y', 'M', 'P', 'F' } → 0
Polar{'S', 'T', 'N', 'Q','C', 'W'}→ 1
Acidic (Polar){'K', 'H', 'R'}→ 2
Basic (Polar){'D', 'E'} → 3

2- Based on structure of the side chain: the nature of the amino acid side chains has significant influence on the topography of the protein. The complex protein structures are generated by the bonds between amino acid side chains. Koolman suggested the following seven categories based on the amino acid structures [18]:

Alphatic {'G', 'A', 'V', 'I', 'L'} → 0
Sulfur-containing {'C', 'M'}→ 1
Aromatic {'F', 'Y', 'W'} → 2
Neutral {'S', 'T', 'N', 'Q'} → 3
Acidic {'D', 'E'}→ 4
Basic {'K', 'R', 'H'}→ 5
Special case {' P'}→ 6

3- Based on the Amino acids chemical properties. The amino acids can be categorized by using dipole scale and volume scale, where dipole scale can be divided into five levels and volume scale can be divided into two levels [16]:

Level one and one {'A','G','V'}→ 0
Level one and two {'I','L','F','P'}→ 1
Level two and two{'Y','N','Q','W'}→ 2
Level three and two {'H','M','T','S'} → 3
Level four and two {'R','K'}→ 4
Level five and two {'D','E'} → 5
Special case {'C'}→ 6

IV. Feature Extraction

Feature extraction will be adopted in this paper due two reasons: Firstly to analysis a large number of variables, a huge amount of memory and computation power are required. For example if the size of the vector *freqVec* is 343 then the size of the variables which are produced by Tensor product is 117659, which is considered very huge to be processed. Secondly a large number of variables is suspected to be notoriously redundant (much data, but not much information) which overfits the training sample and generalizes poorly to new samples. In this paper, three Feature sets are suggested and tested:

Indices =FindLargest(DiffMatrix, v) // To find the indices of the largest v values, which are more common in the interacting proteins

Indices =FindSmallest(Abs(DiffMatrix), v) // To find the indices of the common v values

Indices =Random // To generate v random indices

Thus the set of positive and negative patterns can be generated by using the following algorithm:

Foreach index *i* and *j* in the *indices* array
  Foreach interacting pairs *a* and *b*
    $P_k = freqVec_i^a \times freqVec_j^b$
    $T_k = 1$
Foreach index *i* and *j* in the *indices* array
  Foreach non-interacting pairs *c* and *d*
    $P_{k+t} = freqVec_i^c \times freqVec_j^d$
    $T_{k+t} = 0$

DiffMatrix is the different between the sum of the normalized matrices of the all Tensor product of the interacting pairs and the sum normalized matrices of the all Tensor product of the non interacting pairs

foreach interacting pairs a and b
  $A_k = freqVec^a \otimes freqVec^b$
  $B_k = A_k / A_k$ (normalizing using element by element division)
  $\text{PositiveSum} = \sum_{k=1}^{n} B_k$

foreach non-interacting pairs c and d
  $C_k = freqVec^c \otimes freqVec^d$
  $D_k = C_k / C_k$ (normalizing using element by element division)
  $\text{NegativeSum} = \sum_{k=1}^{n} D_k$

DiffMatrix= PositiveSum- NegativeSum

V. Feature Selection and Support Vector Machines

The previous set of features can be finalized by using one of the feature selections techniques. Signal-to-Noise (S2N) is a fast and reliable feature selection technique. It ranks the features with the ratio of the "signal" (the difference between the mean values of the two classes), and the "noise" (the within class standard deviation). This criterion is similar to the Pearson correlation coefficient, the Ttest criterion and the Fisher criterion. The S2N top ranking features can be selected by using the following formula [19].

$$S2N = \frac{|\mu_+ - \mu_-|}{\sigma_+ + \sigma_-} \quad (1)$$

The last step can be used to predict the PPIs is applying one of the machine learning methods on the finalized set of features. Support Vector Machines (SVM) has been successfully applied to a wide range of pattern recognition and classification problems. SVM can be used to find a hyperplane which divides the data into two classes: the first class is denoted by +1's and the second class is denoted by 1's. The hyperplane is the set of points X satisfying:

$$W^T . X - b = 0 \quad (2)$$

Where *W* is the weights and *X* is a vector. The optimal hyperplanes can be chosen by using soft margin method which splits the classes with maximum margin and minimum error. The following constraint must be added to prevent the points falling into the margin:

$W^T x_i + b \geq 1$ for all x_i of the first class.
$W^T x_i + b \leq -1$ for all x_i of the second class.

For non-linear cases, the data must be mapped into a richer feature space. SVMs use an implicit mapping Φ of the input data into a high-dimensional feature space defined by a kernel function. A general kernel equation is:

$$k(x_1, x_2) = (t + x_1 . x_2)^d e^{(-h\|x_1 - x_2\|^2)} \quad (3)$$

There are many possible kernel functions such as:
RBF Kernel: *h=c, t=0, d=0.*
Linear Kernel: *h=0, t=0* and *d=1*.
Polynomial (homogeneous): *h=0, t=* β and *d=m.*
Polynomial (inhomogeneous): *h=0, t=* 1 and *d=m.*
In this paper a polynomial radius base function (PRBF) is implemented where *d=3, t=1* and *h=0.07*. Thus the Non-linear form is:

$$\min_{W,b,\xi} \frac{1}{2} W^T W + C \sum_{i=1}^{l} \xi_i$$

Subject to

$$y_i (W^T \Phi (x_i) + b) \geq 1 - \xi_i$$
$$\xi_i \geq 0 \quad (4)$$

TABLE I
COMPARISON BETWEEN FOUR SETS OF FEATURES, THREE ENCODING STRATEGIES USING SVM.

Encoding strategies	Features	Training		Validation		Testing	
		BER	AUC	BER	AUC	BER	AUC
Polarity	**Largest**	0.32	98.90	7.12	94.76	34.41	73.10
	Smallest	0.60	100.00	6.51	94.12	31.20	76.12
	Random	0.12	100.00	2.00	96.86	29.18	78.71
	Concatenation	0.31	100.00	1.14	97.33	30.82	77.55
Structure	**Largest**	0.20	99.31	4.31	93.23	29.30	79.40
	Smallest	0.09	100.00	4.23	94.01	29.35	80.91
	Random	0.03	100.00	1.12	99.27	28.81	83.74
	Concatenation	0.15	100.00	1.14	98.43	27.10	81.16
Chemical properties	**Largest**	0.22	99.80	1.10	98.12	25.30	86.13
	Smallest	0.17	100.00	0.83	99.11	21.75	88.18
	Random	0.00	100.00	0.22	100.00	13.31	91.42
	Concatenation	0.12	100.00	0.62	99.40	23.50	88.58

TABLE II
COMPARISON BETWEEN FOUR SIZES OF THE AMINO ACIDS SUBSEQUENCES.

Subsequence Size (m)	Vector size	Indices (v)	Training		Testing		Time(s)
			BER	AUC	BER	AUC	
Two	49	2401	0.37	99.21	24.31	78.01	1283
Three	343	6000	0.00	100.00	13.31	91.42	2675
Four	2401	7000	0.30	100.00	24.11	87.72	2820
Five	16807	8000	0.11	99.75	22.10	87.93	3098

VI. EXPERIMENTAL RESULTS

To compare the prediction accuracy of the suggested techniques, *Drosophila melanogaster* (fruit fly) protein-protein interaction dataset is collected from http://bioinformatics.org.au (under Tools and Data/Databases and Datasets) [20]. The dataset contains 30878 pairs, the pairs were matched by using various methods such as: affinity-chromatography, two hybrid pooling approach and immuno-precipitation. Another 30000 random protein pairs are generated to be considered as non interacting proteins. 40000 pairs are used to create the indices and 20000 pairs are used for training, validation and testing. Matlab 8.0a and CLOP package are used to implement and to compare the state-of-art prediction methods (CLOP Package http://clopinet.com/CLOP/). Two measurements are used: Balance Error Rate (BER) and Area Under Carve (AUC). A k-folding scheme with k=10 is applied to all data. Table 1 compares four sets of features, three of them are introduced in the previous section and the fourth is a simple concatenation between two frequency vectors. On the other hand, three encoding strategies are compared, which are based on polarity, structure and chemical properties. Subsequence with size m=3, v=6000 (the number of the generated indices) and S2N=4000 (the number of the selected features) are used in all experiments in Table 1. It can be observed that the best result can be obtained by using random indices with chemical properties encoding strategy, where BER =13.31 and AUC = 91.41.

Table 2 compares four sizes of the amino acids subsequences, the chemical properties encoding strategy, random indices and S2N=4000 are applied to all experiments. It can be notice that in the case m=2, the maximum length of the frequency vector is 49 and the Tensor product of two vectors is 49*49=2401, thus the number of the generated indices can not exceed 2401. The number of the other indices is selected to be 6000, 7000 and 8000. However it is clear that the best prediction rate is by using m=3.

Table 3 compares between the state of the art of the prediction methods, the chemical properties encoding strategy, random indices and S2N=4000 are applied to all experiments. Six methods are tested by using AUC and BER, the methods are: Random Forest, NeuralNet, LinearSVM, Kridge, NaiveBayes and NonLinearSVM. The best result can be obtain by using NonLinearSVM, therefore, we can conclude that, NonLinearSVM outperforms the state-of-art prediction methods.

TABLE III
COMPARISON BETWEEN THE STATE OF THE ART OF THE PREDICTION METHODS

Method	Testing	
	BER	AUC
Random Forest	33.56	69.30
NeuralNet	30.32	73.17
LinearSVM	37.92	63.12
Kridge	34.93	69.50
NaiveBayes	36.81	64.04
NonLinearSVM	13.31	91.42

VII. Conclusion

Protein-protein interaction is important for tasks ranging from metabolic analysis to drug discovery. With the huge volume of the protein sequences that are stored in the databanks, it is highly demanded to develop a fast and accurate method based on protein sequences to predict protein-protein interaction. In this paper three encoding strategies, four different sets of features and six machine learning methods are implemented and compared. Dataset of 60,000 fruit fly PPIs and non-PPIs are used, the results indicate that using random indices with chemical properties and SVM is superior to the other methods. The next step will be to generate more sets of features, use new encoding strategies, and apply it to other organisms such as human PPIs task where the relatively small number of positive set is a major obstacle.

References

[1] H. Chua, W. Hugo, G. Liu**,** X. Li**,** L. Wong and S. Ng, "A probabilistic graph-theoretic approach to integrate multiple predictions for the protein-protein subnetwork prediction challenge,*" Annals of the New York Academy of Sciences*, vol. 1158, pp 224-233, 2009.

[2] X. Ren and J. Xia, "Prediction of Protein-Protein Interaction Sites by Using Autocorrelation Descriptor and Support Vector Machine," *Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence, Lecture Notes in Computer Science*, vol. 6216, pp. 76-82, 2010.

[3] L. Salwinski, C.S. Miller, A.J. Smith, F.K. Pettit, J.U. Bowie and D. Eisenberg, "The Database of Interacting Proteins," *NAR* vol. **32,**(Database issue), D449-51, 2004.

[4] P. Pagel, S. Kovac, M. Oesterheld, B. Brauner, I. Dunger-Kaltenbach, G. Frishman, C. Montrone, P. Mark, V. Stümpflen, H.W. Mewes, A. Ruepp and D. Frishman, "The MIPS mammalian protein-protein interaction database," *Bioinformatics* vol. 21, no. 6, pp. 832-834; 2005.

[5] L. J. Jensen, M. Kuhn, M. Stark, S. Chaffron, C. Creevey, J. Muller, T. Doerks, P. Julien, A. Roth, M. Simonovic, P. Bork and cC. von Mering, "STRING 8-a global view on proteins and their functional interactions in 630 organisms," *Nucleic Acids Res* vol. 37 Database: D412-D416, 2009.

[6] R. Jansen, H. Yu, D. Greenbaum, Y. Kluger, N. Krogan, S. Chung, A. Emili, M. Snyder, J. Greenblatt and M. Gerstein, "A Bayesian networks approach for predicting protein-protein interactions from genomic data," *Science*, vol 302, pp. 449-453, 2003.

[7] V. Zhang, S. Wong, O. King and F. Roth, Predicting, "co-complexed protein pairs using genomic and proteomic data integration," *BMC Bioinformatics*, vol. 5, no. 1, 38, 2004.

[8] Y. Qi, J. Klein-Seetharaman and Z. Bar-Joseph,"Random forest similarity for protein-protein interaction prediction from multiple sources," *Pac Symp Biocomput*, pp. 531–542, 2005.

[9] Y. Qi, J. Klein-Seetharaman and Z. Bar-Joseph, "A mixture of feature experts approach for protein-protein interaction prediction,", *BMC Bioinformatics*, vol. 8 (S10)**:**S6, 2007 [Online]. Available: http://www.biomedcentral.com/1471-2105/8/S10/S6.

[10] M. Li, L. Lin, X. Wang and T. Liu, "Protein-protein interaction site prediction based on conditional random fields," *Bioinformatics*, vol. 23, no. 5, pp. 597-604, 2007.

[11] J. Espadaler, O. Romero-Isart, R. Jackson and B. Oliva**, "**Prediction of protein-protein interactions using distant conservation of sequence patterns and structure relationships," *Bioinformatics*, vol 21, no.16, pp. 3360-3368, 2005.

[12] B. Wang, L. Sheng Ge, D. Huang and H. Wong, " prediction of protein–protein interacting sites by combining SVM algorithm with Bayesian methods, " Proceedings of the Third International Conference on Natural Computation, vol. 02, pp. 329-333, 2007.

[13] Y. Wang, J. Wang, Z. Yang and N. Deng, " prediction of protein-protein interaction based only on coding sequences," *The Third International Symposium on Optimization and Systems Biology* (OSB'09), pp. 151–158, September 20–22, 2009.

[14] A. Bakar, J. Taheri and A. Zomaya, "Fuzzy systems modeling for protein-protein interaction prediction in *Saccharomyces cerevisie*," 18th World IMACS / MODSIM Congress, Cairns, Australia July 13-17, 2009.

[15] O. N. Yaveroglu and T. Can, "Predicting Protein-Protein Interactions from Protein Sequences Using Phylogenetic Profiles," *in Proceedings of the International Conference on Bioinformatics, Computational and Systems Biology (ICBCSB'09), Singapore, World Academy of Science, Engineering and Technology*, vol. 56 pp. 241-247. June 2009.

[16] J. W. Shen, J. Zhang, X. Luo, W. Zhu, K. Yu, K. Chen, Y. Li and H. Jiang, "Predicting protein-protein interactions based only on sequences information**,"** *Proc Natl Acad Sci USA*, vol. 104, no. 11, pp 4337-4341, 2007.

[17] K.C. Timberlake, " *The chemistry of life*," in Chemistry, 5th Edition, Haper-Collins Publishers Inc, NY, 1992.

[18] J. Koolman, K.H. Rohm, *Colour Atlas of Biochemistry*, Thieme, Stuttgart, 1996.

[19] E. Al-Daoud, "Integration of Support Vector Machine and Bayesian Neural Network for Data Mining and Classification," *World Academy of Science, Engineering and Technology* vol. 64 pp. 202 207, 2010.

[20] C.J.Shin, S.Wong, M.J. Davis and M.A. Ragan, "Protein-protein interaction as a predictor of subcellular location," *BMC Systems Biology* vol. 3, no. 28, 2009.