ISSN : 0974 - 7435 Volume 8 Issue 1

BioTechnology
An Indian Journal

FULL PAPER

BTAIJ, 8(1), 2013 [114-119]

Feature dimension reduction algorithm based prediction method for protein- protein interaction

Tong Wang[1]*, Jihong Yan[3], Bicheng Ye[2], Jian Chen[1]

[1]Institute of Computer and Information, Shanghai Second Polytechnic University, Shanghai, 201209, (CHINA)

[2]School of Computer Science and Technology, Taiyuan University of Science and Technology, Taiyuan, 030024, (CHINA)

[3]Software Engineering Institute, East China Normal University, Shanghai, 200062, (CHINA)

E-mail: zjxywt@163.com

ABSTRACT

Protein-protein interaction is essential to cellular functions. Knowing the protein-protein interaction often provides useful clues for finding its biological function and interaction process with other molecules in a biological system. In this paper, we describe a simple, novel approach to improve the accuracy of predicting protein-protein interaction. Here, dimensionality reduction algorithm is introduced to predict the protein-protein interaction. Our jackknife test results indicate that it is very promising to use the dimensionality reduction approaches to cope with complicated problems in biological systems, such as predicting the protein-protein interaction.

KEYWORDS

Protein-protein interaction;
Feature dimension reduction;
PSSM.

INTRODUCTION

Protein–protein interactions (PPI) promise to reveal many aspects of the complex regulatory network underlying cellular function. PPI network is essential to understand the fundamental processes governing cell biology. Recently, studying PPI networks becomes possible due to advances in experimental high throughput genomics and proteomics technologies. A significant amount of experimental PPI network data for several organisms has already been generated and stored in various PPI interaction databases[1]. However, a majority of these PPI databases such as INTACT (http://www.ebi.ac.uk/intact), BIND (http://binddb.org), DIP (http://dip.doembi.ucla.edu) and MINT (http://mint.bio.uniroma2.it/mint) have been curated manually by domain experts and are far from comprehensive. Machine learning has been shown to have the potential to accelerate the mining and curation process of PPI knowledge[2].

Research in biology and biochemistry has lead to the discovery of various proteins with unknown function that seem to play an important role in biological processes. The accurate annotation of these proteins is often time consuming but can be aided by knowing the precise location of the protein's binding sites and/or its interacting partners. Since almost all proteins carry out their diverse functions by specific protein-protein interactions, the identification of these interacting partners is a wealth of knowledge towards understanding the biochemistry of a particular protein.

Currently the high throughput approach to identifying protein-protein interaction (PPI) is the yeast two-hybrid experiments[3,4]. Despite of its being high through-

put, a typical proteomic project can take over a year to complete and often with noisy or ambiguous data. This has motivated bioinformatics research to develop computational methods for predicting protein-protein interaction, which can then be quickly tested by coimmunoprecipitation or other related experiments. In[5,6], methods were developed for predicting the binding sites exploiting characteristics of the surface residues, whereas some methods focus on deriving sequence signatures from PPI and use these signatures for predicting PPI[7,8]. In a work by Ben-Hur and Noble, kernel methods were developed to predict protein-protein interaction using various sources of data.

Protein-protein interactions are central to all aspects of cellular function including for example gene regulation, immunological recognition and protein synthesis[9]. Hence, identification of binding sites between two interacting proteins is one of basic problems in the research of protein functions. Knowledge of the three-dimensional (3D) structure of the protein complex provides much valuable information on the protein binding site.

Several experimental methods such as X-ray crystallography and NMR can be used to obtain such information. However, they can not meet the requirements of proteomics-generated interaction data according to their current capability for providing such information. Therefore, computational methods are required to assist the identification of potential binding sites in proteins.

So far a number of computational methods[10] have been explored for the prediction of interaction sites in proteins based on the sequence information, 3D structure information or a combination of 3D structure and sequence information. Classification methods such as scoring functions[11], neural network[12], support vector machine (SVM)[13] and random forest[14] have been successful applied for predicting binding sites.

In The present study was initiated in an attempt to propose a completely different approach, the comprehensive comparative study of different DR methods in terms of their ability to predict protein-protein interaction. Moreover, protein sequences are represented by PSSM (Position-Specific Score Matrix)[15-18] which incorporate the evolution information. The result thus obtained is quite encouraging, indicating that the above approach can also be effectively used to deal with other complicated biological systems.

METHODS

Dataset

The experimental data in this study were derived from the dataset used by Liu et al.. This dataset contains 504 protein hetero chains. In addition, we also adopted their definition of surface residues and interface residues. According to this definition, the dataset contains surface residues, about 35.05% of which are interface residues.

Position-specific scoring matrix

In this study, a powerful sequence encoding scheme PSSM is introduced. It is useful to summarize the main definitions associated with this method here.

A protein sequence containing N amino acids can be represented by a 420-D (Dimensional) vector, i.e.,

$$\mathbf{P}_{\text{PSSM-420}} = \left[\overline{\text{A}}_1 \quad \overline{\text{A}}_2 \quad \cdots \quad \overline{\text{A}}_{20} \quad \text{S}_1 \quad \text{S}_2 \quad \cdots \quad \text{S}_{400}\right]^{\text{T}} \tag{1}$$

where the first 20 components are the average scores of every column in $\mathbf{P}_{PSSM}$ matrix. $\mathbf{P}_{PSSM}$ is shown as below:

$$\mathbf{P}_{PSSM} = \begin{pmatrix} \text{A}_{1\to1} & \text{A}_{1\to2} & \cdots & \text{A}_{1\to20} \\ \text{A}_{2\to1} & \text{A}_{2\to2} & \cdots & \text{A}_{2\to20} \\ \vdots & \vdots & \vdots & \vdots \\ \text{A}_{N\to1} & \text{A}_{N\to2} & \cdots & \text{A}_{N\to20} \end{pmatrix} \tag{2}$$

where $\text{A}_{i\to j}$ represents the score of amino acid residue at the i-th position of the protein sequence being substituted to the amino acid type $j(1 \le j \le 20)$ during evolution process. Here, the numerical codes 1, 2,…, 20 represent the 20 native amino acid types according to the alphabetical order of their single-residue codes. N denotes the length of the protein. In this study, $\mathbf{P}_{PSSM}$ is generated by carrying out PSI-BLAST. This process will search the Swiss-Prot database through three iterations for multiple sequence alignment against the protein $\mathbf{P}$. Every element in $\mathbf{P}_{PSSM}$ was scaled by a standardization procedure. The compo-

FULL PAPER

nents $S_1, S_2, \ldots, S_{400}$ in (1) are obtained by summing up all rows in the $\mathbf{P}_{PSSM}$, each of which corresponds to the same amino acid in the primary sequence $\mathbf{P}$. It means for each column in $\mathbf{P}_{PSSM}$, there are 20 values instead of N. Hence, we will have a vector of dimension 20×20 for a $\mathbf{P}_{PSSM}$.

PCA

Principal Components Analysis (PCA) constructs a low-dimensional representation of the data that describes as much of the variance in the data as possible. This is done by finding a linear basis of reduced dimensionality for the data, in which the amount of variance in the data is maximal[19].

In mathematical terms, PCA attempts to find a linear mapping M that maximizes $M^T \operatorname{cov}_{X-\bar{X}} M$, where $\operatorname{cov}_{X-\bar{X}}$ is the covariance matrix of the zero mean data X. It can be shown that linear mapping is formed by the d principal eigenvectors of the covariance matrix of the zero-mean data. Hence, PCA solves the eigenproblem

$$\underset{X-\bar{X}}{\mathrm{COV}} M = \lambda M \quad (3)$$

The eigenproblem is solved for the d principal eigenvalues λ. The low-dimensional data representations y_i of the datapoints x_i are computed by mapping them onto the linear basis M, i.e.,

$$Y = (X - \bar{X})M \quad (4)$$

LDA

Linear Discriminant Analysis (LDA) attempts to maximize the linear separability between datapoints belonging to different classes. In contrast to most other dimensionality reduction techniques, LDA is a supervised technique[19]. LDA finds a linear mapping M that maximizes the linear class separability in the low-dimensional representation of the data. The criteria that are used to formulate linear class separability in LDA are the within-class scatter S_W and the between-class scatter S_B, which are defined as:

$$S_W = \sum_c p_c \underset{X^c - \bar{X}^c}{\operatorname{cov}} \quad (5)$$

$$S_B = \underset{X-\bar{X}}{\operatorname{cov}} - S_W \quad (6)$$

where p_c is the class prior of class label c, $\operatorname{cov}_{X^c - \bar{X}^c}$ is the covariance matrix of the zero mean datapoints x_i assigned to class $c \in C$, and $\operatorname{cov}_{X-\bar{X}}$ is the covariance matrix of the zero mean data X. LDA optimizes the ratio between the within-class scatter S_W and the between-class scatter S_B in the low-dimensional representation of the data, by finding a linear mapping M that maximizes the so-called Fisher criterion

$$\phi(M) = \frac{\left| M^T S_B M \right|}{\left| M^T S_W M \right|} \quad (7)$$

This maximization can be performed by computing the d principal eigenvectors of $S_W^{-1} S_B$. The low-dimensional data representation Y of the datapoints in X can be computed by mapping them onto the linear basis M, i.e., $Y = (X - \bar{X})M$.

Kernel PCA

Kernel PCA (KPCA) is the reformulation of traditional linear PCA in a high-dimensional space that is constructed using a kernel function[19]. Kernel PCA computes the principal eigenvectors of the kernel matrix, rather than those of the covariance matrix.

The reformulation of traditional PCA in kernel space is straightforward, since a kernel matrix is similar to the inproduct of the datapoints in the high-dimensional space that is constructed using the kernel function. The application of PCA in kernel space provides Kernel PCA the property of constructing nonlinear mappings.

Kernel PCA computes the kernel matrix K of the datapoints x_i. The entries in the kernel matrix are defined by

$$k_{ij} = k(x_i, x_j) \quad (8)$$

where k is a kernel function. Subsequently, the kernel matrix K is centered using the following modification of the entries

$$k_{ij} = k_{ij} - \frac{1}{n}\sum_{l} k_{il} - \frac{1}{n}\sum_{l} k_{jl} + \frac{1}{n^2}\sum_{lm} k_{lm} \quad (9)$$

The centering operation corresponds to subtracting the mean of the features in traditional PCA. It makes sure that the features in the high-dimensional space defined by the kernel function are zeromean. Subsequently, the principal d eigenvectors v_i of the centered kernel matrix are computed. It can be shown that the eigenvectors of the covariance matrix α_i are scaled versions of the eigenvectors of the kernel matrix v_i

$$\alpha_i = \frac{1}{\sqrt{\lambda_i}} v_i \quad (10)$$

In order to obtain the low-dimensional data representation, the data is projected onto the eigenvectors of the covariance matrix α_i. The result of the projection is given by

$$Y\left\{\sum_j \alpha_1 k(x_j, x), \sum_j \alpha_2 k(x_j, x), ..., \sum_j \alpha_d k(x_j, x)\right\} \quad (11)$$

Kernel LDA

By introducing a kernel function which corresponds to the non-linear mapping, all the computation can conveniently be carried out in the input space. The problem can be finally solved as an eigen-decomposition problem like PCA, LDA and KPCA. From the theory of reproducing kernel we know that any solution $w \in F$ must lie in the span of all training samples in F. Let ϕ be a nonlinear mapping to some feature space F. F we need to maximize[20]

$$J(w) = \frac{w^T S_B^{\phi} w}{w^T S_W^{\phi} w} \quad (12)$$

where S_B is between-class scatter matrix and S_W is within-class scatter matrix. Therefore we can find an expansion for w of the form

$$w = \sum_{i=1}^{l} \alpha_i \phi(x_i) \quad (13)$$

Using the expansion Eq.13 and the definition of m_i^{ϕ} we write[20]

$$w^T m_i^{\phi} = \frac{1}{l_i}\sum_{j=1}^{l}\sum_{k=1}^{l_i} \alpha_j k(x_j, x_k^i) = \alpha^T \mathrm{M}_i \quad (14)$$

Where we defined $(M_i)_j := \frac{1}{l_i}\sum_{k=1}^{l_i} k(x_j, x_k^i)$ and replaced the dot products by the kernel function. Now consider the numerator of Eq.12. Be using the definition of S_B^{ϕ} and Eq.14 it can be rewritten as

$$w^T S_B^{\phi} w = \alpha^T M \alpha \quad (15)$$

where $M := (M_1 - M_2)(M_1 - M_2)^T$. Considering the denominator, using Eq.13, the definition of m_i^{ϕ} and a similar transformation as in Eq.15 we find:

$$w^T S_W^{\phi} w = \alpha^T N \alpha \quad (16)$$

Where we set $N := \sum_{j=1,2} K_j (I - 1_{l_j}) K_j^T$, K_j is a $l \times l_j$ matrix with $(K_j)_{nm} := k(x_n, x_m^j)$ (this is the kernel matrix for class j), I is the identity and 1_{l_j} the matrix with all entries $1/l_j$.

Combining Eq.15 and Eq.16 we can find linear discriminant in F by maximizing

$$J(\alpha) = \frac{\alpha^T M \alpha}{\alpha^T N \alpha}. \quad (17)$$

This problem can be solved (analogously to the algorithm in the input space) by finding the leading eigenvector of $N^{-1}M$. We will call this approach (nonlinear) Kernel LDA. The projection of a new pattern x onto w is given by

$$(w \cdot \phi(x)) = \sum_{i=1}^{l} \alpha_i k(x_i, x) \quad (18)$$

Thus, using Eq.18 we can map a protein sample into some high-dimensional feature space as desired.

EXPERIMENTAL RESULTS

The performance of four different DR methods from the perspective of identifying protein-protein interaction was compared. The accuracy of the low dimensional representations of the high dimensional data ob-

FULL PAPER

tained by the different DR methods was evaluated via KNN[21,22] algorithm. Accordingly, the jackknife test has been increasingly and widely adopted by investigators[23-26] to test the power of various predictors. Therefore, in this study, jackknife test was performed with the current approach in predicting the protein-protein interaction.

As shown in Table 1, the overall jackknife success rates obtained by DR methods in identifying the protein-protein interaction are higher than the ones obtained without using linear DR methods. Meantime, it indicates that supervised DR methods (LDA and KLDA) outperform unsupervised DR methods (PCA and KPCA) and the nonlinear DR methods (KPCA and KLDA) outperform linear DR methods (PCA and LDA). In summary, base on the observation, it is concluded that the overall jackknife success rate with KLDA is the highest relative to the other DR methods.

TABLE 1 : Success rates in identifying protein-protein interaction by the jackknife test

Method	Sequence encoding schemes	Test method (%)
		Jackknife
K-NN(K=1)	PSSM	81.80
PCA& K-NN(K=1)	PSSM	83.05
KPCA& K-NN(K=1)	PSSM	84.53
LDA& K-NN(K=1)	PSSM	88.70
KLDA& K-NN(K=1)	PSSM	92.38

CONCLUSIONS

In this paper, we compared the performance of four different DR methods from the perspective of discriminating protein-protein interaction. The results obtained are encouraging, which are higher than the ones obtained without DR methods. The application of DR approach to the prediction of protein-protein interaction is just an example to demonstrate its advantages. It has not escaped our notice that the DR approach can also be used to deal with many other complicated biological systems.

ACKNOWLEDGMENT

This work was supported by National Natural Science Foundation of China (Grant No. 61301249 and Grant No.61272036), Innovation Program of Shanghai Municipal Education Commission (Grant No.12YZ175) and The Soft Engineering of Key Subjects Construction in Shanghai Second Polytechnic University (Grant No. XXKZD1301).

REFERENCES

[1] D. J. Higham, M. Rasajski, and N. Przulj, "Fitting a geometric graph to a protein-protein interaction network," Bioinformatics, **24**, 1093-9 **(2008)**.

[2] I.Spasic, S.Ananiadou, J. McNaught, A.Kumar; Text mining and ontologies in biomedicine: making sense of raw text, Brief Bioinform, **6**, 239-51 **(2005)**.

[3] T.Ito et al.; A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. USA*, **98**, 4569-4574 **(2001)**.

[4] P.Ueta et al; A comprehensive analysis of proteinprotein interactions in *Saccharimyces cerevisiae*. *Nature*, **403**, 623-627 **(2000)**.

[5] J.L.Chung, W. Wang, P.Bourne; Exploiting Sequence and Structure Homologs to Identify Protein-Protein Binding Sites. *PROTEINS: Structure, Function, and Bioinformatics*, **62**, 630-640 **(2006)**.

[6] C.Yan, V.Honavar, D.Dobbs; Identifying protein-protein interaction sites from surface residues – A Support Vector Machine Approach. *Neural Computing Applications*., **13**, 123-129 **(2004)**.

[7] J.Fang et al. Discover protein sequence signatures from protein-protein interaction data. *BMC Bioinformatics*, **6**, 277 **(2005)**.

[8] S.Martin *et al*.; Predicting protein-protein interactions using signature products. *Bioinformatics*, **21**, 218-226 **(2005)**.

[9] J.F.Xia, X.M.Zhao, J.Song, D.H.Huang; APIS: accurate prediction of hot spots in protein interfaces by combining protrusion index with solvent accessibility. BMC bioinformatics, **11**, 174 **(2010)**.

[10] I.Ezkurdia et al; Progress and challenges in predicting protein-protein interaction sites. Briefings in Bioinformatics, **10(3)**, 233 **(2009)**.

[11] N.Burgoyne, R.Jackson; Predicting protein interaction sites: binding hot-spots in protein-protein and protein-ligand interfaces. Bioinformatics, **22(11)**, 1335 **(2006)**.

[12] H.Zhou, Y.Shan; Prediction of protein interaction sites from sequence profile and residue neighbor list. Proteins Structure Function and Genetics, **44(3)**,

336–343 **(2001)**.

[13] B.Wang et al.: Predicting protein interaction sites from residue spatial sequence profile and evolution rate. FEBS letters, **580(2)**, 380–384 **(2006)**.

[14] M.Iki, S.Tomi, K.Vlahoviek; Prediction of Protein–Protein Interaction Sites in Sequences and 3D Structures by Random Forests. PLoS Computational Biology, **5(1)**, **(2009)**.

[15] B.Liu et al.: Prediction of protein binding sites in protein structures using hidden Markov support vector machine. BMC bioinformatics, **10(1)**, 381 **(2009)**.

[16] K.C.Chou, H.B.Shen; MemType-2L: a web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. Biochemical and Biophysical Research Communications, **360(2)**, 339-345 **(2007)**.

[17] M.Kumar et al.; BhairPred: prediction of beta-hairpins in a protein from multiple alignment information using ANN and SVM techniques. Nucleic Acids Res, **33(Web Server issue)**, W154-9 **(2005)**.

[18] D.Xie et al.; LOCSVMPSI: a web server for subcellular localization of eukaryotic proteins using SVM and profile of PSI-BLAST. Nucleic Acids Res, **33(Web Server issue)**, W105-10 **(2005)**.

[19] L.J.P.V.D.Maaten, E.O.Postma, H.J.V.D.Herik; Dimensionality Reduction: A Comparative Review., **(2007)**.

[20] S.Mika, G.Ratsch, J.Weston, B.Scholkopf; Fisher discriminant analysis with kernels. in: Neural Networks for Signal Processing IX: Proceedings of the 1999 IEEE Signal Processing Society Workshop Madison, WI, US, 41–48 **(1999)**.

[21] T.Denoeux; A k-Nearest Neighbor Classification Rule Based on Dempster-ShaferTheory. IEEE Transactions on Systems, Man, and Cybernetics, **25(5)**, 804-813 **(1995)**.

[22] J.M.Keller, M.R.Gray; A fuzzy k-nearest neighbor algorithm. IEEE Transactions on Systems, Man, and Cybernetics, **15(4)**, 580-585 **(1985)**.

[23] X.Xiao et al.; Using complexity measure factor to predict protein subcellular location. Amino Acids, **28(1)**, 57-61 **(2005)**.

[24] B.Niu et al.; Predicting protein structural class with AdaBoost Learner. Protein and Peptide Letters, **13(5)**, 489-492 **(2006)**.

[25] J.Chen et al.; Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. Amino Acids, **33(3)**, 423-428 **(2007)**.

[26] D.Q.Liu et al.; Predicting secretory protein signal sequence cleavage sites by fusing the marks of global alignments. Amino Acids, **32(4)**, 493-6 **(2007)**.