INFORMATION SYSTEMS IN MANAGEMENT

Information Systems in Management (2016) Vol. 5 (1) 109−118

APPLICATION OF SPARSE LINEAR DISCRIMINANT ANALYSIS FOR PREDICTION OF PROTEIN-PROTEIN INTERACTIONS

KATARZYNA STĄPOR, PIOTR FABIAN

Silesian Technical University, Faculty of Automatic Control, Electronics and Computer Science

To understand the complex cellular mechanisms involved in a biological system, it is necessary to study protein-protein interactions (PPIs) at the molecular level, in which prediction of PPIs plays a significant role. In this paper we propose a new classification approach based on the sparse discriminant analysis [10] to predict obligate (permanent) and non-obligate (transient) protein-protein interactions. The sparse discriminant analysis [10] circumvents the limitations of the classical discriminant analysis [4, 9] in the high dimensional low sample size settings by incorporating inherently the feature selection into the optimization procedure. To characterize properties of protein interaction, we proposed to use the binding free energies. The performance of our proposed classifier is 75% ± 5%.

Keywords: sparse discriminant analysis, feature selection, protein-protein interaction

1. Introduction

Proteins are large molecules that constitute the bulk of the cellular machinery of any living organism or biological system. Regulation of biochemical pathways, signaling cascades and transduction, cellular motion, gene regulation, forming a protein complex, modifying or carrying another protein are some of the essential biological processes in living cells performed by protein-protein interactions (PPIs) [5]. As a consequence, to understand the complex cellular mechanisms involved in

a biological system, it is necessary to study the nature of these interactions at the molecular level, in which prediction of PPIs plays a significant role.

PPIs have been investigated in various ways, involving both experimental (in vivo or in vitro) and computational (in silico) approaches [2, 8]. Experimental approaches tend to be costly, labor intensive and suffer from noise. Therefore, using computational approaches for prediction of PPIs is a good choice for many reasons.

There are different types of protein-protein interactions that provide different levels of information on different biological processes [5]. For example, based on the affinity and stability, PPIs can be divided into: 1) non-obligate complexes: binding components (proteins) can form stable structures and cannot exist in vivo independently, 2) obligate complexes: components do not form stable functional structures on their own and can be stable in vivo independently. Based on the duration and life time of the interactions, there are transient complexes (temporarily in vivo) and permanent ones (interactions are stable and irreversible). In general, all obligate complexes are permanent. Except from some examples, all non-obligate interactions can be considered as transient.

Although interfaces have been the main subject of study to predict protein-protein interactions, an accuracy of 70% has been independently achieved by several different groups [7, 8, 11, 12]. These approaches have been carried out by analyzing a wide range of parameters, including solvation energies, amino acid composition, conservation, electrostatic energies, and hydrophobicity and different classification strategies. Up to this moment, the best results (**78%**) were obtained in [7] by using contact and binding free energies as features and the discriminant analysis [4, 9] combined with the initial selection of features to cope with the limitations of the discriminant analysis [4, 9] in the high-dimensional, low-sample size (HDLSS) settings (i.e. when the number of features is greater than the sample size). But there are two main weak points in the work [7]. First, the initial feature selection method causes that some important information is lost. Second, the Authors in [7] did not provide the method for the estimation of variance of their classifier. So, we do not know what is the error rate of their result 78%.

In this paper, we propose the new classification approach based on the sparse discriminant analysis [10] to predict obligate (permanent) and non-obligate (transient) protein-protein interactions. The sparse discriminant analysis [10] circumvents the limitations of discriminant analysis in the HDLSS by incorporating inherently the feature selection into the optimization procedure. As a results, the new method [10] finds the sparse projection directions. To characterize properties of protein interaction, we proposed to use the binding free energies. The performance of our proposed classifier is 75% ± 5%.

In this study we use discriminant analysis for the predictive purposes only (*predictive discriminant analysis,* PDA), i.e. to predict group membership given a number of continuous variables. The study for explaining group separation or

group differences in terms of variable importance which is the aim of the *descriptive discriminant analysis* (DDA) will be the subject of our future research in which the correlation structure will be examined. We also plan to compare it with other variable importance methods like for example linear ordering.

There is an important distinction between DDA and PDA. In DDA, adding, of variables to a statistical analysis does not take away from effect size, and often increases uncorrected effect sizes. However, in PDA, fewer variables can yield greater classification accuracy, whereas in DDA fewer variables cannot yield greater discrimination. Thus, good features selected for PDA are those giving the best prediction performance.

This paper is organized as follows. Section 2 shortly presents the classical discriminant analysis as well as its sparse version. The proposed classification method for protein-protein interaction is described in Section 3 while the results of the conducted experiments with this method – in section 4. Section 5 comprises the conclusions.

2. Fisher and Sparse regularized linear discriminant analyses

Fisher Linear Discriminant analysis (FLDA) [4, 9] is a multivariate technique which is concerned with the search for a linear transformation that reduces the dimension of a given *p*-dimensional statistical model to *q* ($q < p$) dimensions, while maximally preserving the discriminatory information for the several classes within the model.

Formally, suppose that there are *k* classes and let $x_{ij}\ \ j = 1,...,n_i$ be vectors of observations from the *i*-th class, $i = 1,\ldots, k$. Set $n = n_1 + ... + n_k$ and let $X_{n\times p} = (x_{11},...,x_{1n_1},...,x_{k1},...,x_{kn_k})^T$, where *p* is a dimensionality of an input space. FLDA determines a linear mapping *L*, i.e. a $q \times p$ matrix *A*, that maximizes the so-called Fisher criterion J_F (1):

$$J_F(A) = tr((AS_W A^T)^{-1}(AS_B A^T)) \tag{1}$$

where $S_B = \sum_{i=1}^{k} p_i (m_i - \bar{m})(m_i - \bar{m})^T$ and $S_W = \sum_{i=1}^{k} p_i S_i$ are the between-class and the average within-class scatter matrix, respectively;

$S_i = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (x_{ij} - m_i)(x_{ij} - m_i)^T$ is the within-class covariance matrix of class *i*, m_i is the mean vector of class *i*, p_i is it's *a priori* probability, and $\bar{m} = \sum_{i=1}^{k} p_i m_i$ is the overall mean. FLDA maximizes the ratio of between-class scatter to average within-class scatter in the lower-dimensional space. Optimizing (1) comes down to

determining an eigenvalue decomposition of $S_W^{-1}S_B$, and taking the rows of A to equal the q eigenvectors corresponding to the q largest eigenvalues. There are no more than $\min(p, k-1)$ eigenvectors corresponding to nonzero eigenvalues.

In the high-dimensional, low-sample size (HDLSS) settings, the within-class covariance matrix S_W is singular and the classical FLDA breaks down. Several extensions have been proposed to overcome this problem but all of them possess the data pilling problem [6]. To ameliorate this problem, some sparse version of LDA have been proposed.

In our approach, to circumvent this problem, we adapt the sparse linear discriminant approach (slda) from [10] that incorporates feature selection in FLDA. The term "sparse" means that the discriminant vectors have only a small number of nonzero components. The underlying assumption is that, among the large number of variables there are many irrelevant or redundant variables for the purpose of classification. This method is based on the connection of FLDA and a generalized eigenvalue problem, stated formally by the following theorem [10].

T h e o r e m

Suppose S_w is a positive definite matrix and denote its Cholesky decomposition as $S_w = R_w^T R_w$ (R_w is an upper triangular matrix). Let H_b be $k \times p$ matrix, $V_1, ..., V_q$ ($q \le \min(p, k-1)$) denote the eigenvectors of $S_W^{-1}S_B$ corresponding to the q largest eigenvalues $\lambda_1 \ge ... \ge \lambda_q$, $A = [\alpha_1, ..., \alpha_q]$, $B = [\beta_1, ..., \beta_q]$. For $\lambda > 0$ let $\hat{A}, \hat{B}$ be the solution to the following problem (2):

$$\min_{A,B} \sum_{i=1}^{k} \left\| R_w^{-T} H_{b,i} - AB^T H_{b,i} \right\|^2 + \lambda \sum_{j=1}^{q} \beta_j^T (S_w) \beta_j \text{ subject to } A^T A = I_{p \times q}, \qquad (2)$$

where:

$H_{b,i} = \sqrt{n_i}(\overline{x}_i - \overline{x})^T$ is the *i*-th row of the matrix

$H_b = \left(\sqrt{n_1}(\overline{x}_1 - \overline{x}), ..., \sqrt{n_k}(\overline{x}_k - \overline{x}) \right)^T$, e^{n_i} is a vector of ones with length n_i.

Then $\hat{\beta}_j$, $j = 1, ..., q$, span the same linear space as V_j, $j = 1, ..., q$.

The following method of regularization is applied in [10] to circumvent the singularity problem and to obtain the sparse linear discriminants: i.e. the first q sparse discriminant directions $\beta_1, ..., \beta_q$ are defined as the solutions to the following optimization problem (3):

$$\min_{A,B} \sum_{i=1}^{k} \left\| R_w^{-T} H_{b,i} - AB^T H_{b,i} \right\|^2 + \lambda \sum_{j=1}^{q} \beta_j^T \left(S_w + \gamma \frac{tr(S_w)}{p} I \right) \beta_j + \sum_{j=1}^{q} \lambda_{1,j} \left\| \beta_j \right\|_1 \qquad (3)$$

subject to $A^T A = I_{p\times q}$, where $B = [\beta_1, ..., \beta_q]$, $\|\beta_j\|_1$ is the 1-norm of the vector β_j, the same λ is used for all *q* directions, different $\lambda_{1,j}$'s are allowed to penalize different discriminant directions.

According to the theorem stated above, the solution of the optimization problem (2) is independent of the value of λ, but this does not necessarily imply that the solution of the regularized problem (3) is also independent of λ. However, our empirical study suggests that the solution is very stable when λ varies in a wide range, for example in (0.01, 10000).

We can use K-fold cross validation (CV) [9] to select the optimal parameters $\lambda_{1,j}$, but when the dimension of the input data is very large, the numerical algorithm becomes time consuming and we can let $\lambda_{1,1} = ... = \lambda_{1,q}$. The tuning parameter γ controls the strength of the regularization of the matrix S_w, the large values will bias too much S_w towards identity matrix (high degree of regularization). In our empirical studies, we find that the results are not sensitive to the choice of γ if a small value that is less than 0.1 is used, in our studies we set $\gamma = 0.05$. More careful studies of choice of γ are left for future research.

The above problem can be numerically solved by alternating optimization over *A* and *B* [10] and the resulting algorithm is summarized below.

Regularized sparse LDA (rSLDA) algorithm (based on [10])

1. Form the matrices from the input data:

$$H_w = X - \begin{pmatrix} e^{n_1}(\bar{x}_1)^T \\ \dots \\ e^{n_K}(\bar{x}_K)^T \end{pmatrix}$$

$$H_b = \left(\sqrt{n_1}(\bar{x}_1 - \bar{x}), ..., \sqrt{n_k}(\bar{x}_k - \bar{x})\right)^T$$

2. Compute upper triangular matrix R_w from the Cholesky decomposition of:

$\left(S_w + \gamma\frac{tr(S_w)}{p}I\right)$ such that $\left(S_w + \gamma\frac{tr(S_w)}{p}I\right) = R_w^T R_w$

3. Solve the *q* independent optimization problems

$$\min_{\beta_j} \beta_j^T(\tilde{W}^T\tilde{W})\beta_j - 2\tilde{y}^T\tilde{W}\beta_j + \lambda_1\|\beta_j\|_1 \quad j = 1,\dots,q$$

where $\tilde{W}_{(n+p)\times p} = \begin{pmatrix} H_b \\ \sqrt{\lambda} \cdot R_w \end{pmatrix}$ $\tilde{y}_{(n+p)\times 1} = \begin{pmatrix} H_b R_w^{-1} \alpha_j \\ 0 \end{pmatrix}$

4. Compute SVD:

$$R_w^{-T}\left(H_B^T H_B\right) B = UDV^T \text{ and let } A = UV^T$$

5. Repeat steps 3 and 4 until converges.

3. Protein-protein interaction classification method

To characterize properties of protein interaction, we proposed to use the binding free energies. These were computed using *FastContact* [3], which obtains their fast estimates. FastContact delivers the electrostatic energy, solvation free energy, and the top 20 maximum and minimum values for:

1) residues contributing to the binding free energy,
2) ligand residues contributing to the solvation free energy,
3) ligand residues contributing to the electrostatic energy,
4) receptor residues contributing to the solvation free energy,
5) receptor residues contributing to the electrostatic energy,
6) receptor-ligand residue solvation constants,
7) receptor-ligand residue electrostatic constants.

Thus, all these values and the total solvation and electrostatic energy values compose a total of 282 features characterizing interaction.

To create a dataset for classification, we used the pre-classified dataset from previous study [7] containing 62 transient and 75 obligate complexes as two different classes for classification. Each complex is listed in the form of chains for ligand and receptor respectively. The relevant data about the structure of each complex was obtained from the Protein Data Bank (PDB) [1] and then obtaining the 282 features by invoking FastContact.

Due to the fact that the number of features (282) is greater than the number of samples in a dataset (137), we have HDLSS setting, so we apply sparse regularized linear discriminant analysis for the calculation of discriminant directions, i.e. the algorithm sparse rLDA described above.

For the classification of the samples in the new discriminant space, we applied the *nearest mean classifier* [4, 9] as the classification algorithm. The nearest mean (centroid, prototype) classifier assigns to new observations the label of the class of training samples whose mean is closest to the observation.

4. Experimental results

In our experiments we have used the dataset of 137 protein complexes described in [11]. 75 samples in this dataset belong to the first class (i.e. "obligate interactions") and 62 samples to the second class (i.e. "non-obligate interactions"). This dataset is randomly divided into a "training set" and "testing set" in a ratio of 4:1.

As we have only two classes ($k = 2$), there is only one discriminant direction β_1 ($q = 1$). Using all variables in constructing the discriminant vector β_1 might cause the overfitting of the training data, resulting in high testing error rate. Moreover it is computationally demanding, so sparsification would be a good choice.

Denote the number of significant variables involved in specifying the discriminant direction β_1 (i.e. giving the best prediction), to be *m*. To find these most significant variables we have performed the experiment with varying values of *m*. For a given value of *m*, only the *m* maximum values of the coordinates of the vector β_1 (so called *beta* values) are left, the rest is zeroed.

Fig. 2 shows the components of vector β_1 obtained by the rSLDA algorithm in one of experiments converted to the absolute values and sorted in the ascending order.

We leave only *m* biggest values, zeroing all others. We keep track of indices of these biggest values and modify the original β_1 leaving only *m* biggest values. These values are used to cast the original 282-dimensional vector onto a one-dimensional space. The projection of the samples from the protein dataset uses only these *m* non-zero coefficients.

Then, classification is performed in such new discriminant space by the nearest mean (centroid) classifier. The classification performance is measured on the separate test set.

The results are shown in Fig. 1. We can observe that the error rate of the nearest mean classifier grows rapidly and then decreases with the rise of *m*, up to 28 (error = ~25% ± 5% measured on the testing set). Then, for bigger values of *m*, almost a constant error rate was observed.

From the plot it is clear that if we specify m=28 as the number of component variables in discriminant vector β_1 – sparse LDA algorithm can discriminate the two classes fairly well (the classifier performance = **~75% ± 5%**) (where 5 is the confidence interval).

These 28 input features ("selected" by the rSLDA algorithm) are the most significant for classification (i.e. giving the best classification performance). These are the following from the full set of 282 features (corresponding to the ascending order of the absolute value of the coefficients composing vector β_1):

202 198 281 200 48 42 243 203 47 133 128 121 161 160
157 132 49 156 46 134 241 131 155 158 127 119 135 41

Among these 28 features – 13 are from the receptor residues contributing to the desolvation free energy, but these are not from the beginning of the above list. It can be observed that in each of the 7 groups of energetic features – only features with extreme (min or max) contribution to the energy are always selected. The features from the beginning of the list are those from the receptor residues contributing to the electrostatics energy. One may conclude that electrostatic energy is the most important in the prediction of obligate/non-obligate protein-protein interactions. Electrostatic energy involves a long-range interaction and occur between charged atoms of two interacting proteins.

Thus, the rSLDA algorithm does suggest which constituents are the most important in the classification of interactions.

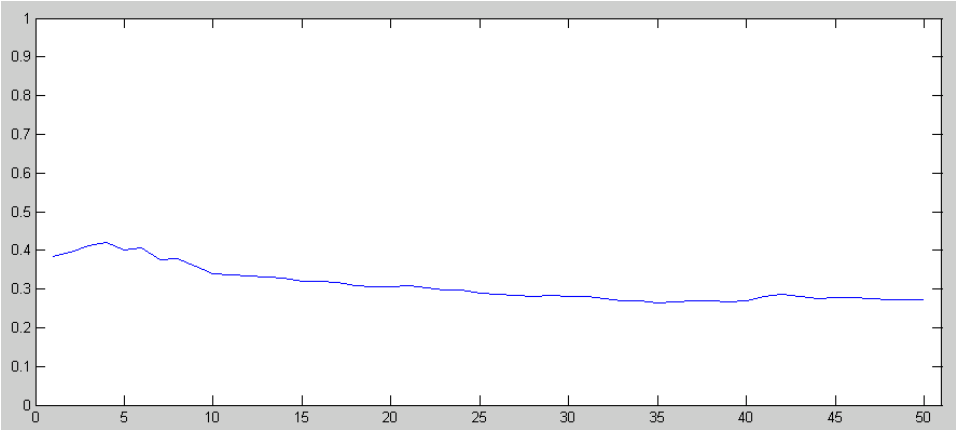

Figure 1. The average classification error rate as a function of the number of variables using nearest centroid method on the projected data – the local minimum is at 28

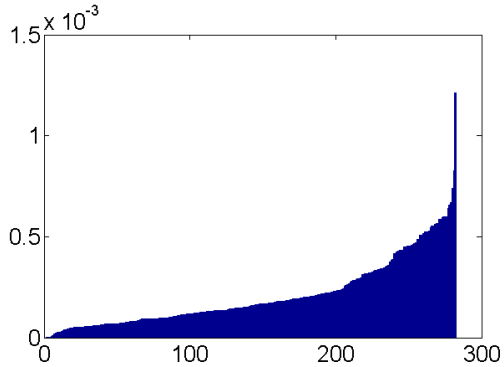

Figure 2. Components of β obtained by the rSLDA algorithm in one of experiments converted to absolute values and sorted in ascending order (description in text)

5. Conclusion

We have proposed a classification approach for obligate/non-obligate (transient) protein-protein complexes. We have used regularized version of sparse linear discriminant analysis algorithm [10] for feature extraction as well as for input variable selection. To discriminate between two types of protein interactions: obligate and non-obligate, we have used the "energetic features". These are based on the binding free energy defined as the sum of the desolvation and electrostatic energies. These were computed effectively using the package FastContact [3]. The results on the protein-protein interactions dataset showed that using only 28 from 282 input variables enables the classification of the mentioned two types of interactions with the performance of $\mathbf{75\% \pm 5\%}$. Among the most important features are those from residues contributing to the electrostatic energy.

The hypothesis on the importance of the electrostatic energy in the prediction of obligate/non-obligate protein-protein interactions should be confirmed by the additional experiments on bigger protein datasets. This will be the subject of our future research.

REFERENCES

[1] Berman H. et al. (2000) *The Protein Data Bank.* Nucleid Acid Research 28, 235-242.

[2] Bordner A., Abagyan R. (2005) *Statistical analysis and prediction of protein-protein interfaces*. Proteins 60 (3), 353-366.

[3] Camacho C., Zhang C. (2005) *FastContact: rapid estimate of contact and binding free energies*. Bioinformatics 21 (10), 2534-2536.

[4] Fukunaga K. (1990) *Introduction to statistical pattern recognition.* New York: Academic Press.

[5] Jones S., Thornton J.M. (1996) *Principles of protein-protein interactions.* Proc. Natl. Acad. Sci. USA 93(1), 13-20.

[6] Marron J. et al. (2007). *Distance-weighted discrimination.* Journal of American Statistical Association, 102, 1267-1273.

[7] Rueda L. et al. (2010) *Biological protein-protein interaction prediction using binding free energies and linear dimensionality reduction.* In: Dijkstra T., et al. (eds): PRIB 2010, LNBI 6282, 383-394, Springer Berlin.

[8] Skrabanek L. et al (2008) *Computational prediction of protein-protein interactions*. Molecular Biotechnology, 38(1), 1-17.

[9] Stąpor K. (2011) *Classification methods in computer vision*. PWN Warszawa (in Polish).

[10] Qiao Z., Zhou L., Huang J. (2009) *Sparse linear discriminant analysis with applications to high dimensional low sample size data.* IAENG Int. Journal of Applied Mathematics, 39, 1.

[11] Zhou H., Shan Y. (2001) *Prediction of protein-protein interaction sites from sequence profile and residue neighbor list.* Proteins 44(3), 336-343.

[12] Zhu H., et al. (2006) *NoxClass: prediction of protein-protein interaction types*. BMC Bioinformatics 7 (27).