Predicting Interaction Status of a Pair of Proteins Using Statistical Data Analysis Methods

Ömer Nebil Yaveroğlu
Computer Engineering Department, Middle East Technical University
nebil@ceng.metu.edu.tr

Abstract

In this paper, four methods are proposed to predict whether two given proteins are interacting or not. The four methods applied for prediction of interaction status of proteins are Principle Component Analysis, Multidimensional Scaling, K-means Clustering and Support Vector Machines. These methods are applied on a dataset constructed by checking the existence of proteins of Saccharomyces Cerevisiae organism in different species. Several results obtained by applying these methods show that it is possible to make some predictions about protein interactions by the application of statistical data analysis methods. The most obvious classification results were obtained by the application of Support Vector Machines.

1. Introduction

Identification of protein-protein interactions (PPIs) is important for understanding protein functions and biological processes in a cell. There are many methods proposed for identifying the interaction of protein pairs. Most of these methods use protein properties such as protein sequences [1], primary structures of proteins for this prediction [2].

In this paper, a new method based on existence of proteins in different species is proposed for protein-protein interaction prediction. A dataset is constructed by comparing the protein pairs and species. A score is given for each pair of species and protein pairs according to the existence of the proteins in the considered specie. Applying statistical data analysis methods on the dataset constructed this way, I tried to predict whether a given pair of proteins are interacting or not. For this purpose, I have separately applied principle component analysis, multidimensional scaling, k-means clustering and support vector machines. The dataset applied consists of labeled data so the results obtained can be easily criticized in terms of accuracy. The highest accuracy obtained is the result of support vector machine application with 64.0026% accuracy.

2. Materials and Methods

2.1. Data set construction

The PPI data collected from *Saccharomyces Cerevisiae* is used in order to construct the dataset. This PPI data includes 11698 protein pairs. 5849 of these protein pairs interact and the rest 5849 protein pairs don’t interact. For the construction of the used dataset, an existence check is made between the proteins in protein pairs and species. If the dataset is considered as an m by n matrix, the rows of the matrix represent the protein-protein pairs and the columns represent different species. If the proteins in a row both exist in the specie specifying the column, the cell value at the intersection of the related column and cell takes the value 2. In a similar way, if both proteins don’t exist in the related specie, that cell value becomes 0. If one of the proteins exists in the related specie but the other is not, the related cell takes the value -1. This scoring mechanism is constructed depending on the amount of effect of protein existence in species on the protein interactions. If one of the proteins exists and the other doesn’t exist in a given specie, this shows that there is a small amount of probability on the existence of a protein-protein interaction on the given protein pair. So this should be penalized in order to reduce its effect on the prediction. 450 species are checked this way for each of the protein-protein pairs. Constructing the dataset with this scoring mechanism, the dataset becomes 11698 by 450 matrix which shows the comparison of species and protein-protein interactions with the values

2, 0 and -1. This constructed dataset is used for all of the methods mentioned in this paper.

2.2. Application of principle component analysis

The first effort spent to discriminate the interacting and non-interacting protein pairs is done by applying principle component analysis (PCA) on the dataset. PCA is a way of identifying patterns in data, and expressing the data in such a way as to highlight their similarities and differences [3]. It is also possible to reduce the dimensions of the data which makes it possible to scatter high dimensional data in two or three dimensions. Both two dimensional and three dimensional PCA are performed on the data in this study using Statistics Toolbox of Matlab. The aim was to see whether the interacting and non-interacting protein pairs can be visualized as separated after the application of PCA and plotting the projected data on to a coordinate system to allow visualization of the dataset.

2.3. Application of multidimensional scaling

Multidimensional scaling (MDS) is a set of statistical techniques used for information visualization for exploring similarities and dissimilarities in data [4]. It can also be used for data dimension reduction and it is a useful method for discovering non-linear patterns in data. Aim of applying MDS was similar to PCA with the idea of finding out non-linear properties of the data. Application of MDS on the dataset is performed by using Statistical Toolbox of Matlab. During the construction of the distance matrix of the dataset, Euclidean distance is used. Then MDS is performed using squared stress criterion. Memory requirements of Matlab required dataset size reduction during the application even if I have used the multi-cluster Nar Machine of our department. So MDS is performed over 1000 protein-protein pairs extracted from the data set in order to be able to get a result.

2.4. Application of k-means clustering

K-means clustering is an unsupervised clustering technique which tries to optimize a given criterion function. This clustering technique directly looks for a division of n objects into k groups. This method is applied to our dataset in order to check if it is possible to divide the dataset into two without any label information. I have performed k-means clustering on the data using Statistics Toolbox of Matlab. After performing k-means clustering on the data, I have performed PCA projection with a coloring scheme depending on the resulting cluster that the sample is put by k-means clustering and the real label of the data. As a result, a projection consisting of samples of 4 different colors is formed. It is possible to visualize some outliers by this way. Also by counting the number of samples in each of the color groups, it is possible to see whether performed clustering separates the interacting and non-interacting proteins well.

2.5. Application of support vector machines

Support Vector Machine (SVM) is a useful technique for informed data classification [5]. Main usage of SVM consists of training and testing phases. In the training phase, a model of the data is created from the given training data. In the testing phase, the given testing data is classified depending on the model created in the training phase. Several different models can be created using different kernel functions in the training phase. The library named Libsvm provides tools for performing classification using SVM. A Python code which automates the process of SVM application is also available in this library package. This code tries to find optimal parameters for SVM application using radial basis function (RBF) as the kernel function and returns an accuracy result on the clustering that is created by the SVM classification. Using this code and giving training and testing datasets of sizes 250, 500, 750 and 1000, I have performed SVM classification on data. Also projecting the classified data as performed on the k-means clustering, a visualization of the clustering performed is possible.

3. Results and Discussion

3.1. Principle component analysis results

Principle component analysis is a basic statistical analysis method that can be applied for recognizing some patterns in the dataset. Projecting the data on the first two principle components generates a plot of the dataset as in Supplementary Material (SM) Figure 1. During the generation of the given plot, the samples which represent the interacting proteins are scattered red and the non-interacting samples are colored blue. As can be seen from Figure 1, there is no obvious separation of interacting and non-interacting proteins. But the non-interacting protein pairs are more spread over the coordinate system while the interacting proteins are more grouped. Also it is more likely to see interaction proteins at the lower values of the second principle component.

When the data is projected onto the first three principle components, a three dimensional plot as in SM Figure 2 is created. In this plot, the pattern in the data is more obvious because of the addition of the third dimension. Although there is not a 100% separation between the interacting and non-interacting proteins, there seems a pretty clear grouping of data. The interacting proteins are dense especially on the area at the intersection of 0 at the first principle component and [-10,0] at the second principle component. Also the non-interacting proteins are more grouped in the area at the intersection of [-10,0] on the first principle component and [5,10] on the second principle component. Even with these groupings, it is not easy to directly say the class of interaction of a given protein-protein pair.

3.2. Multidimensional scaling results

Usage of MDS is crucial for finding out non-linear patterns in data. For this aim, performing MDS on dataset using Euclidean distance as the distance metric seems a solution to find out some non-linear properties of data. Using squared stress as the criterion of MDS, it was possible for me to get some results. But in general MDS is a procedure including huge amount of computations. For this reason, many memory problems occurred and no results were produced. Usage of the departments' multi cluster machine, Nar, couldn't produce any results with the whole dataset either. Because of this problem, data size reduction was necessary in order to get some results. With a reduction to 1000 instances of data, the results in SM Figure 3 are produced. In this figure, the interacting protein pairs are colored red and non-interacting protein pairs are colored blue for illustration purposes.

As can be seen from the given figure, there is a grouping of non-interacting protein pairs around the value 0 of the first dimension. On the other hand, interacting protein pairs don't show a strong clustering. Even if it is possible to see a clear grouping of non-interacting protein pairs, it is not possible to use this method for classification purposes because of the huge amount of memory requirements of the method. The method is not practical for the purposes of protein-protein interaction prediction.

3.3. K-means clustering results

K-means clustering is used for our aim to find out if it is possible to classify a protein pair directly without any prior knowledge about the dataset. Performing binary classification using k-means clustering is an easy task and computational manageable. With the application of k-means clustering, all 11698 protein pairs are classified into two clusters. In order to perform the best separation between the clusters, I have replicated the clustering 15 times and took the clustering which gives the best separation between the clusters. Comparing the performed clustering with the labels of the data, it was possible to find some accuracy results of the classification performed by k-means clustering. In Table 1, there is a comparison of the clusters created and the labels of data.

Table 1: Comparison of k-means clustering results with the labels of the protein pairs

	Interacting	Non-interacting
Cluster 1	4672	4076
Cluster 2	1153	1797

Table 1 show that k-means clustering doesn't precisely separate the interacting and non-interacting proteins. It creates a huge cluster as Cluster 1 and a smaller cluster as Cluster 2. Because of this unbalanced separation, most of the protein pairs are classified in cluster 1. As a result, the interacting and non-interacting proteins are not well separated. It would be a meaningful classification if the two opposite corners of this table had more data than the other two opposite corners.

Even with this situation, it is possible to come up with some results from k-means clustering by visualizing the data. For this aim, usage of PCA and projecting the data on the first two principle components generated SM Figure 4. Also a projection on the first three principle components performed in the same way is available as SM Figure 5. Coloring scheme of these projections can be found in Table 2.

Table 2: Coloring scheme of K-means clustering projections

	Interacting	Non-interacting
Cluster 1	Blue	Red
Cluster 2	Green	Yellow

With this coloring scheme, it is easier to recognize the clustering in the data. As can be seen from SM Figure 4, there exist some clustering especially around the (-3,-5) for blue, (13,-5) for green, (15, 7) for yellow and (-3, 7) for red. It is also possible to say that red instances form a grouping as a line as can be seen from SM Figure 4. It is easier to see these clusters in SM Figure 5 in three dimensions. This can be used to make predictions when predicting the interaction status of a given protein pair. Given a protein pair, by projecting it onto the first three principle components, more accurate predictions can be made by looking at its distance to these clusters. For instance, if the given sample is placed on (-2,-7) on the first two principle components then it would be more accurate to say that the proteins in this pair interact. This doesn't guarantee 100% prediction accuracy but it gives some intuition about which group that this protein pair may belong to.

To get a more statistical measure of validity on the k-means clustering performed on the data set, I have performed validation on the dataset using Dunn's Index and Davies Bouldin Validation Techniques. I have used Cluster Validation Toolbox of Matlab for this purpose achieving validity values 0.4697 in Davies Bouldin Validation and infinite for Dunn's Index Validation. The reason of getting infinite result in Dunn's Index Validation may be a bug on the implementation of validation technique. But Davies Bouldin validation gives returns a small value showing that the clustering performed is not a strongly separated one.

3.4. Support vector machine results

Using an informed learning technique to classify protein pairs as interacting or non-interacting is another option that can be considered to make predictions about a protein pairs' situation. In this manner, with the usage of Libsvm library, performing SVM classification is another method that I have tried. In order to create the training and testing datasets, I have divided the dataset into two parts. To get more accurate results, different sizes of training and testing data are created. During the creation of the datasets, same amount of interacting and non-interacting protein pairs are put into the dataset. After creating the training dataset, the non-included protein pairs are included in the testing dataset. Giving these datasets of different sizes to a script in Libsvm package, some classifications are performed returning the highest accuracy of classification by comparing the classification results with the labels of the data. In Table 3, you can see the accuracy results returned for different sizes of training and testing data size.

As can be seen from Table 3, a classification accuracy of 64.0026% can be achieved in the best case. This amount of accuracy shows some improvement over random assignment of classification values for protein pairs. But this improvement is not a huge improvement since there are only two classes, interacting and non-interacting. Even with a random class determination, 50% accuracy can be acquired with a basic probability calculation. But this result is important for showing that with a more suitable statistical analysis method; it may be possible to make predictions based on the dataset constructed by the existence of protein structures in different species.

Table 3: Accuracy results of SVM classification for different sizes of training and testing data

Training Dataset Size	Testing Dataset Size	Accuracy of Classification	Number of Correctly Classified instances
500	11198	62.0736%	6951
1000	10698	64.0026%	6847
1500	10198	61.2179%	6243
2000	9698	61.3529%	5950

Visualization of the clustered data is also performed with the method used in k-means clustering. The testing data is projected on the first two principal components using a coloring scheme as in Table 2. SM Figure 6 represents this two dimensional projection. In fact, the results of k-means clustering give the same intuition about the interacting and non interacting proteins. But SVM may produce more accurate results because of its supervised nature. In SM Figure 7, it is possible to see the projection of the data on the first three principle components. SM Figure 7 gives more clear visualization of the data.

4. Conclusion

In this article, I have tried some statistical data analysis methods to find out a separation of interacting and non-interacting proteins of *Saccharomyces Cerevisiae*. The methods applied for this aim are Principle Component Analysis, Multidimensional Scaling, K-means Clustering and Support Vector Machines. Constructing a dataset with an existence check of proteins in different species, a different insight is introduced to protein-protein interaction prediction. The results of applying these methods on the dataset constructed showed that it is possible to make some predictions based on the relation of proteins and the species. There have been some patterns formed by the application of the methods even if they do not claim high accuracies of prediction. An accuracy of 64.0026% is obtained with the use of SVM even if it is a supervised classification method. The accuracy may be increased with the use of more advanced classification methods. Changing the scoring mechanism of the dataset construction algorithm may also result with better classification results. Also k-means clustering with higher number of cluster numbers may give better clustered data. Overall, such a classification method would give a simpler prediction model which can be used in protein-protein interaction prediction.

References

[1] Yanzhi Guo, Lezheng Yu, Zhining Wen, Menglong Li. Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences. *Nucleic Acids Research Vol. 36, No. 9, 2008*

[2] Joel R. Bock, David A. Gough. Predicting protein–protein interactions from primary structure. *Bioinformatics Vol. 17 no. 5 2001*.

[3] Lindsay I Smith. A tutorial on Principal Components Analysis. 2002

[4] http://en.wikipedia.org/wiki/Multidimensional_scaling. Multidimensional Scaling. Wikipedia

[5] Chih-Wei Hsu, Chih-Chung Chang, Chih-Jen Lin. A Practical Guide to Support Vector Classification. 2008