PREDICTION OF PROTEIN-PROTEIN INTERACTIONS USING STATISTICAL DATA ANALYSIS METHODS

ABSTRACT

In this paper, four methods are proposed to predict whether two given proteins interact 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 phylogenetic profiles of *Saccharomyces Cerevisiae* (baker's yeast) protein pairs. The phylogenetic profile dataset is constructed from protein sequence information alone by checking the existence of homologs of *Saccharomyces Cerevisiae* proteins in different species. Several results obtained by applying these methods show that it is possible to make accurate predictions about protein interactions using sequence information alone by the application of statistical data analysis methods. The best classification results are obtained by 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 predicting the interaction of protein pairs. Some of these methods use protein sequence information alone [3, 1].

In this paper, a new method based on phylogenetic profiles of proteins is proposed for protein-protein interaction prediction. The phylogenetic profile dataset is constructed by looking for homologs of a protein in a number of species. A score is given for each protein sequence pair indicating the level of sequence similarity. Protein pairs that exhibit sequence similarity above a given score threshold are deemed homologs. By applying statistical data analysis methods on the dataset of phylogenetic profiles, we try to predict whether a given pair of proteins are interacting or not. For this purpose, we have separately applied principle component analysis, multidimensional scaling, k-means clustering, and support vector machines. We evaluate the accuracy of these methods on a benchmark interaction dataset of *S. cerevisiae* where the interacting and non-interacting pairs are known. The highest accuracy obtained is the result of support vector machine application with 64.0026% accuracy.

2. MATERIALS AND METHODS

2.1 Dataset Construction

The PPI data [3] collected from *Saccharomyces Cerevisiae* organism 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 do not interact. The interacting proteins are taken from the core subset of the Database of Interacting Proteins (DIP) [6] and non-interacting proteins are formed by pairing proteins in different subcellular localizations.

For the construction of the phlogenetic profile dataset used in this study, the homologs of each protein is searched in a set of 450 fully sequenced genomes. For a single protein, the result is a binary vector in which the existence of homologs are indicated by a value of 1 and a value of 0 indicates that there is no homolog of the protein in the corresponding organism. Two phylogenetic profiles of length n are combined to get a combined profile of length n using the following procedure. If both the proteins have a homolog in the same organism, the column value corresponding to that organism is set to 2 in the combined profile. In a similar way, if both proteins do not have a homolog in an organism, the column corresponding to that organism is set to 0. If one of the proteins has a homolog and the other does not, then the corresponding column is set to -1. The general idea of this scoring mechanism is based on considering the existence of homologs of a protein pair in similar organisms as an evidence for functional similarity and interaction. If one of the proteins exists and the other does not exist in a given organism, this shows that there is a small possibility of an 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. We use 450 fully sequenced organisms to check for homology. After constructing the phylogenetic profile dataset with this scoring mechanism, the dataset becomes a 11698 by 450 matrix which shows the combined phylogenetic profiles of protein-protein pairs 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 to discriminate the interacting and non-interacting protein pairs is 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 [5]. It is also possible to reduce the dimensions of the data which can be used in scattering 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 of this approach is to see whether the interacting and non-interacting protein pairs can be visualized as two separated clusters after the application of PCA and plotting the projected data on 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. It can also be used for data dimension reduction and it is a useful method for discovering non-linear patterns in data. Our aim of applying MDS is similar to PCA along with the idea of finding out non-linear properties of the data. Application of MDS on the dataset is performed by using the 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 of MDS even if multi cluster HPC machine of Middle East Technical University is used. Therefore MDS is performed over 1000 protein-protein pairs randomly extracted from the dataset in order to get a result with the current computational resources.

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. It is applied to our dataset in order to check if it is possible to divide the dataset into two without using any label information, namely interacting (cluster 1) and non-interacting (cluster 2) clusters. We have performed k-means clustering on the data using Statistics Toolbox of Matlab. After performing k-means clustering on the data, we have performed PCA projection. In this projection, a coloring schema consisting of four colors is used. The colors used in this schema represent interacting protein pairs clustered in cluster 1, interacting protein pairs clustered in cluster 2, non-interacting protein pairs clustered in cluster 1 and non-interacting protein pairs clustered in cluster 2. It is possible to visualize some outliers in the dataset by this way. Also by counting the number of samples in each of the color groups, it is possible to see whether k-means 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 supervised data classification [4]. 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 [2] 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 formed by the SVM classification. Giving training datasets of sizes 500, 1000, 1500 and 2000 to this code, we have performed SVM classification on data. Projecting this classified data as performed on the k-means clustering, a visualization of the clustering performed by SVM is created.

3. RESULTS AND DISCUSSION

3.1 Principle Component Analysis Results

Principle component analysis is a basic statistical data 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 Figure 1. During the generation of the given plot, the samples which represent the interacting proteins are colored in red and the non-interacting samples are colored in 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 spread over the coordinate system while the interacting proteins are more grouped. Also it is more likely to see interacting 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 Figure 2 is created. In this plot, the pattern in the data is more obvious by means of separation because of the addition of the third dimension. Although we cannot observe a 100% accurate separation between the interacting and non-interacting proteins, there seems a pretty clear grouping of the 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 [0,10] on the first principle component and [5,10] on the second principle component. Even with these groupings, it is not easy to directly state whether a given protein-protein pair is an interacting pair.

3.2 Multidimensional Scaling Results

Usage of MDS is crucial for finding out non-linear patterns in the data. We performed MDS on our dataset using Euclidean distance as the distance metric. Using squared stress as the criterion function of MDS, it was possible to get promising results. But in general, MDS is a procedure including huge amount of computations. For this reason, memory problems occurred and no results were produced for the whole dataset. Usage of high performance computing (HPC) cluster machine could not produce any results with the whole dataset neither. Because of this problem, dataset size reduction was necessary in order to get some results. After a dataset size reduction to 1000 random instances of the data, the results in Figure 3 are produced. In this figure, the interacting protein pairs are colored in red and non-interacting protein pairs are colored in blue for illustration purposes.

As can be seen from Figure 3, there exists a grouping of non-interacting protein pairs around the value 0 of the first dimension. On the other hand, interacting protein pairs do not show a strong clustering. It is possible to see a clear grouping of non-interacting protein pairs; however, it is not possible to use this method for classification purposes. The reason is that, as stated before, this method has huge amount of mem-

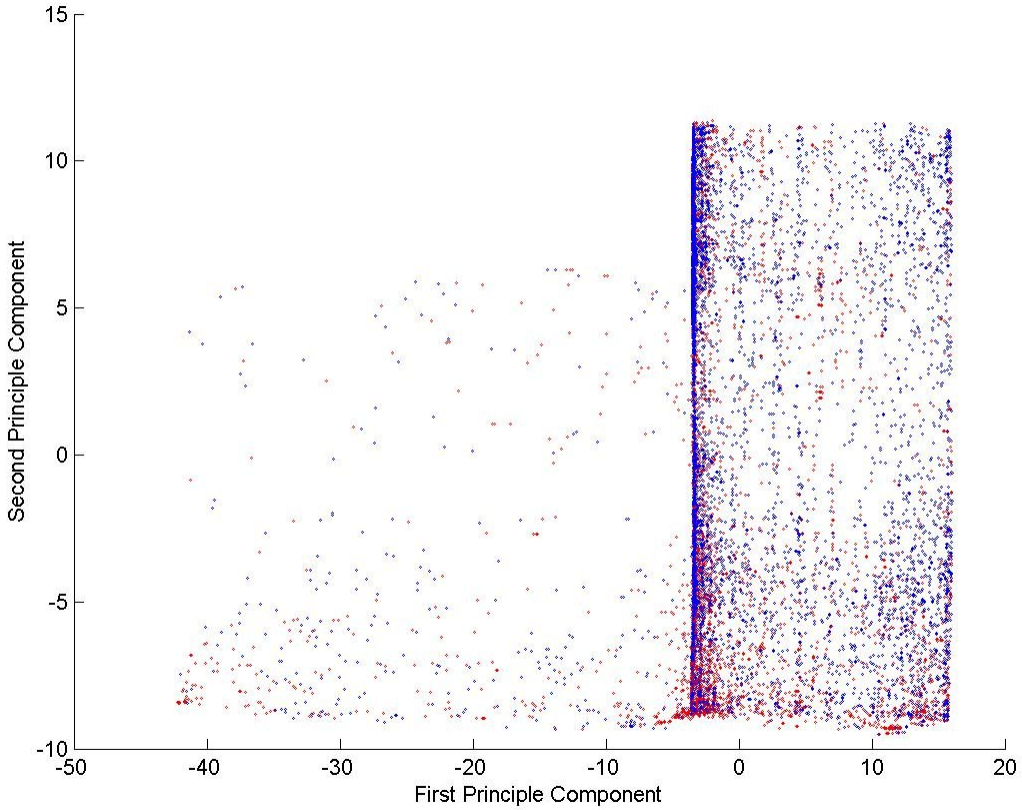

Figure 1: The projection of data on the first two principle components

ory requirements and as a result the usage of this method is not practical.

3.3 K-means Clustering Results

K-means clustering is used for the aim of finding out whether it is possible to classify a protein pair without any prior knowledge about the label of the data. Performing binary classification using k-means clustering is an easy task and computationally manageable. With the application of k-means clustering, all 11698 protein pairs are classified into two clusters. In order to find the best separation between the clusters, we have replicated the clustering 15 times and selected the clustering which gives the best separation between the clusters. Comparing the performed clustering with the labels of the data, it was possible to compute 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 shows that k-means clustering does not precisely separate the interacting and non-interacting proteins. It creates a huge cluster with 8748 instances as Cluster 1 and a smaller cluster with 2950 instances as Cluster 2. The clustering performed is unbalanced. Moreover, the interacting and non-interacting proteins are not well separated, most of them being classified in Cluster 1. It would be a meaningful classification if the two opposite corners of this table had higher values than the other two opposite corners. In other words, if both clusters were individually well matched with one of the labels, this would be an evidence to the accuracy of the method. But the results show that the k-means method does not provide us any useful information about PPI classification.

Although the classification accuracy of k-means clustering is low, it is still possible to come up with meaningful results from k-means clustering by visualizing the data. For this purpose, performing PCA projection of the k-means clustered data on the first two principle components generated Figure 4. Also a projection on the first three principle components performed in the same way is available in Figure 5. Coloring schema used in these projections can be found in Table 2. As can be seen from Figure 5, visualization of k-means clustering produces similar results to PCA projection. The difference of k-means clustering is that the dataset is represented in four clusters instead of two. If the clusters formed by false positive and false negative instances can be identified, it is possible to make better predictions compared to the direct application of PCA.

To get a more significant statistical measure of validity on the k-means clustering performed on the dataset, we have

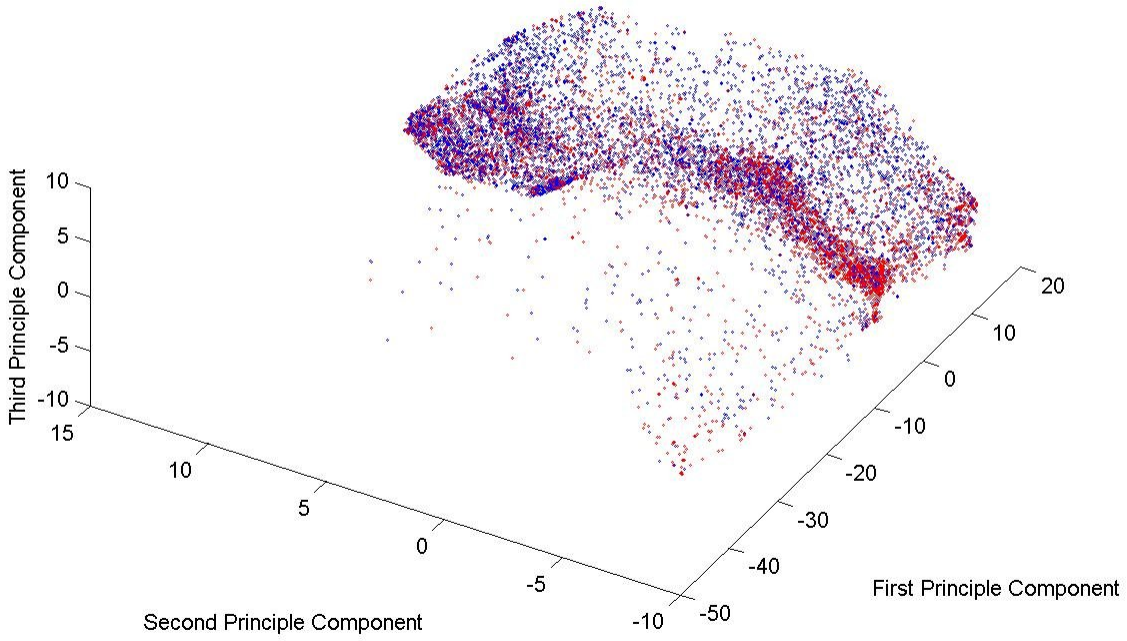

Figure 2: The projection of data on the first three principle components

Table 2: Coloring scheme of K-means clustering projections

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

performed validation on the dataset using Dunn's Index and Davies Bouldin Validation Techniques. We 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 the validation technique. But Davies Bouldin validation returns a small value showing that the clustering performed is not a strongly separated one.

3.4 Support Vector Machine Results

Using a supervised learning technique to classify protein pairs as interacting or non-interacting is another option that can be considered to make predictions about interaction status of a protein pair. In this manner, with the usage of Libsvm library, performing SVM classification is another method that we have investigated. In order to create the training and testing datasets, we have divided the initial dataset into two parts randomly. To get more accurate results, we have tried different sizes of training and testing data. During the creation of the datasets, same amount of interacting and non-interacting protein pairs are included in 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, SVM classification is performed returning the highest classification accuracy. The classification accuracy is obtained by comparing the classification results with the real labels of the data. In Table 3, the accuracy results obtained for different sizes of training and testing data sizes can be found.

As can be seen from Table 3, the highest accuracy result that our tests achieved is 64.0026%. With a random class determination, 50% accuracy can be acquired by a basic probability calculation. There are only two categories that a protein pair can be, either interacting or non-interacting. As a result, given a protein pair, there is 50% chance that you can make a true prediction randomly. Therefore accuracy obtained by applying SVM shows some improvement over random class determination. The improvement amount may seem small but this result is significant. It is an evidence for showing that it is possible to make protein-protein interaction predictions based on the phylogenetic profile dataset constructed by protein sequence information alone.

Visualization of the clustered data is also performed using a similar PCA-based visualization as in k-means clustering. The testing data is projected on the first two principal components using the coloring schema as in Table 2. Figure 6 depicts this two dimensional projection. In fact, the results of k-means clustering is similar to the results of SVM classification. But SVM produces more accurate results because

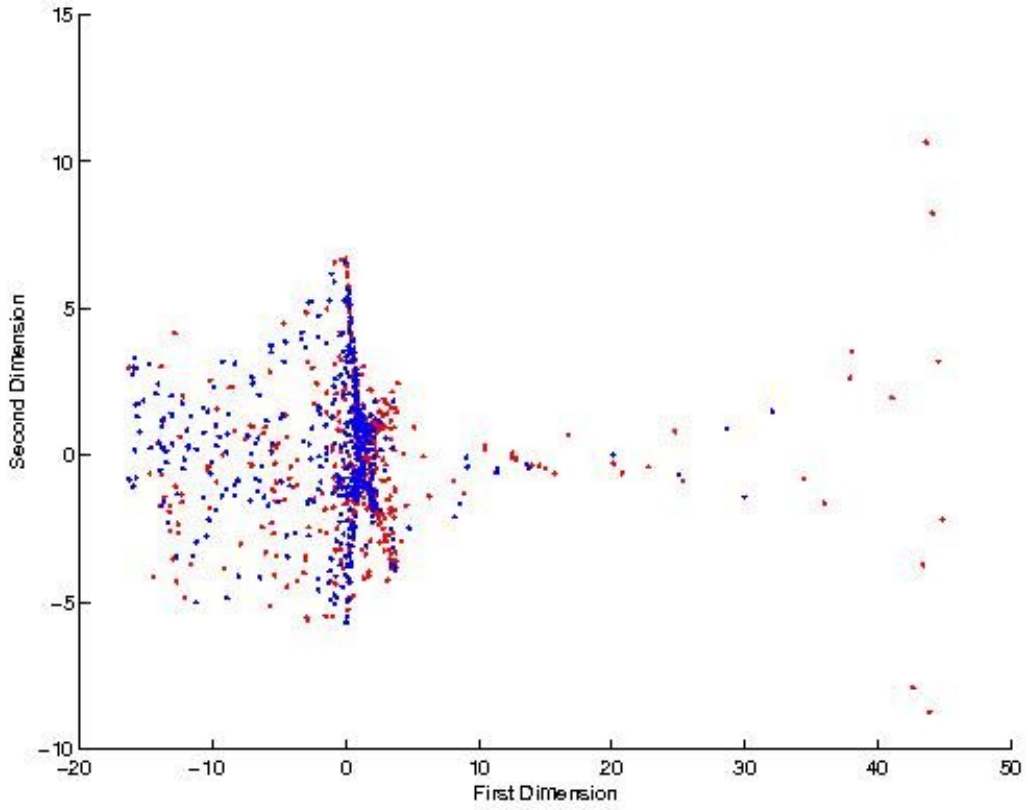

Figure 3: The projection of a subset of dataset using Multidimensional Scaling

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

of its supervised nature. In Figure 7, it is possible to see the projection of the data on the first three principle components. Figure 7 shows that the false positive and false negative instances are reduced considerably compared to k-means clustering.

4. CONCLUSIONS

In this study, we have tried some statistical data analysis methods to find out a discrimination of interacting and non-interacting proteins of *Saccharomyces Cerevisiae*. The methods applied for this purpose were 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 the protein-protein interaction prediction. The results of applying these methods on the dataset constructed shows that it is possible to make some predictions based on the phylogenetic profiles of proteins. Some patterns have been found by the application of these methods even if they do not claim high accuracies of prediction. Among the applied methods, highest accuracy of 64.0026% is obtained with the use of SVM. The accuracy can be increased with the use of more advanced classification methods. Improving the scoring mechanism of the dataset construction algorithm may also result with better classification results. Also k-means clustering with higher number of clusters may produce better clustered data. As an overall result, a simple prediction model which can be used in protein-protein interaction prediction is constructed in this study.

REFERENCES

[1] Joel R. Bock, David A. Gough, "Predicting protein-protein interactions from primary structure" *Bioinformatics*, vol. 17, no. 5, 2001.

[2] Chih-Chung Chang, Chih-Jen Lin, "LIBSVM: a Library for Support Vector Machines", 2003.

[3] Yanzhi Guo, Lezheng Yu, Zhining Wen, Menglong Li, "Using support vector machine combined with auto

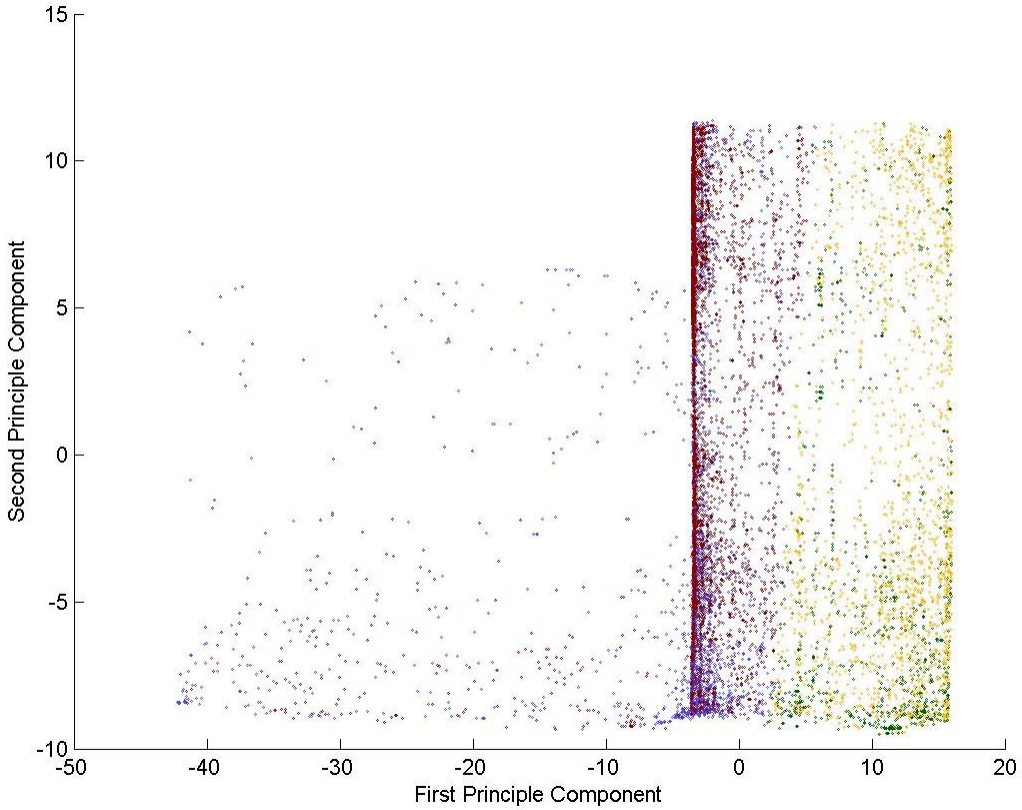

Figure 4: The projection of the data onto the first two principle components with a coloring depending on the k-means clustering results and the labels of the data

covariance to predict protein-protein interactions from protein sequences" *Nucleic Acids Research*, vol. 36, no. 9, 2008.

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

[5] Lindsay I. Smith, "A tutorial on Principal Components Analysis", 2002.

[6] I. Xenarios, L. Salwinski, X. J. Duan, P. Higney, S. M. Kim, D. Eisenberg, "DIP:the database of interacting proteins. A research tool for studying cellular networks of protein interactions." *Nucleic Acids Research*, vol. 30, pages: 303-305, 2002.

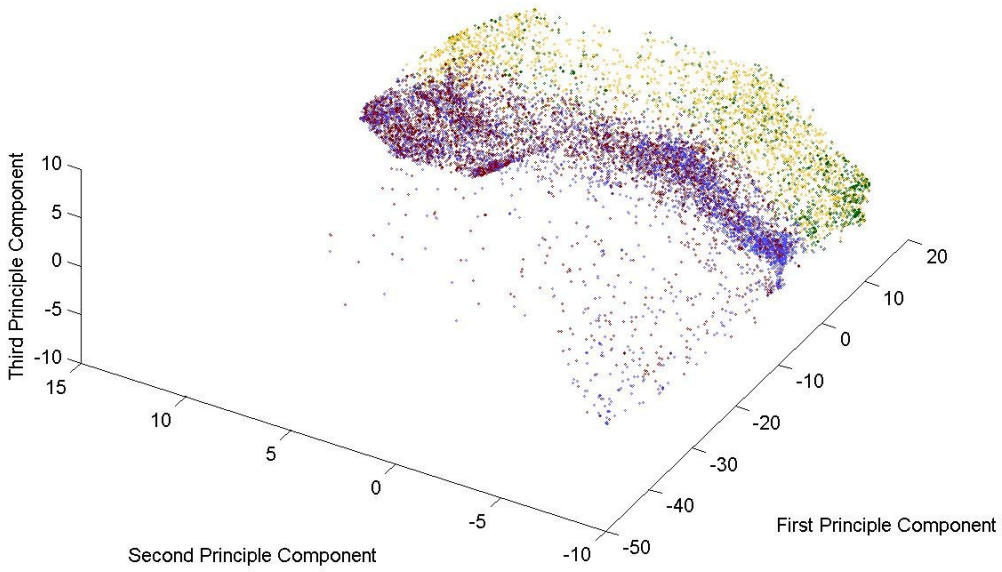

Figure 5: The projection of the data onto the first three principle components with a coloring depending on the k-means clustering results and the labels of the data

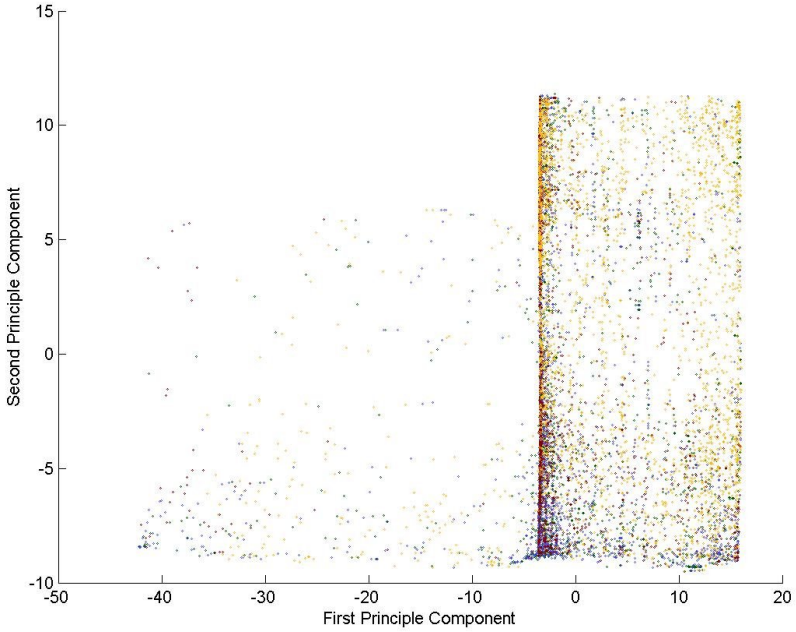

Figure 6: The projection of the data onto the first two principle components with a coloring depending on the classification results performed by support vector machines and the labels of the data

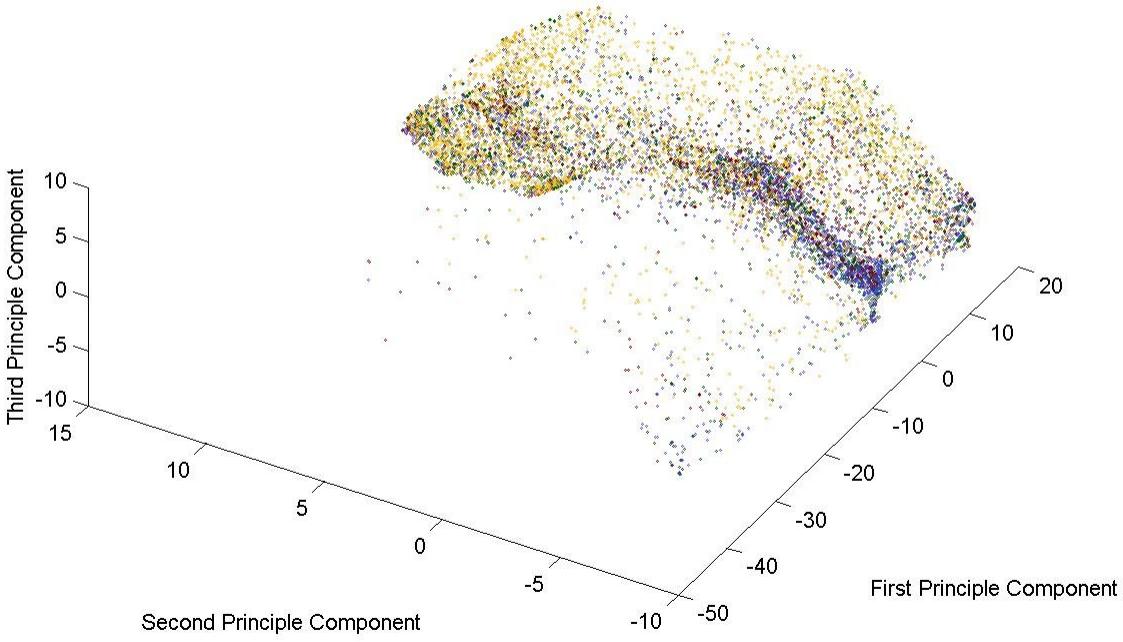

Figure 7: The projection of the data onto the first three principle components with a coloring depending on the classification results performed by support vector machines and the labels of the data