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

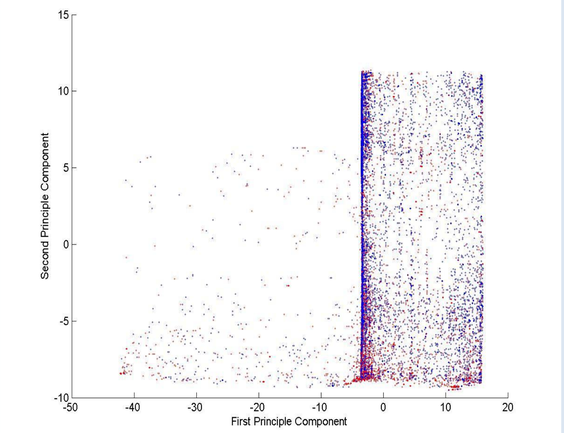
Ömer Nebil Yaveroğlu
Computer Engineering Department, METU
nebil@ceng.metu.edu.tr

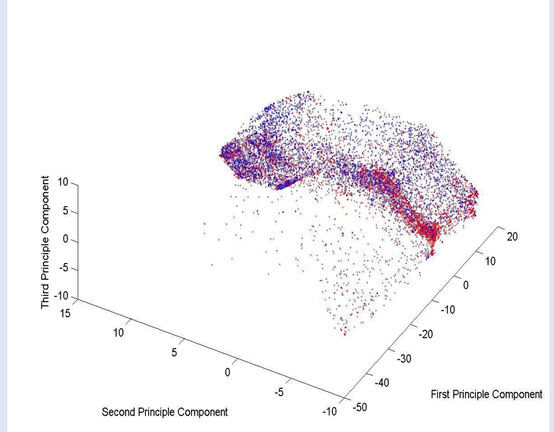
AIM

Performing protein-protein interaction prediction by looking at the existence of query proteins in different species using Principle Component Analysis, Multidimensional Scaling, K-means Clustering and Support Vector Machines

Principle Component Analysis

The dataset is projected onto the first two principle components forming a 2 dimensional plot representing the data. Also 3 dimensional plotting is performed by projection on the first three principle components.

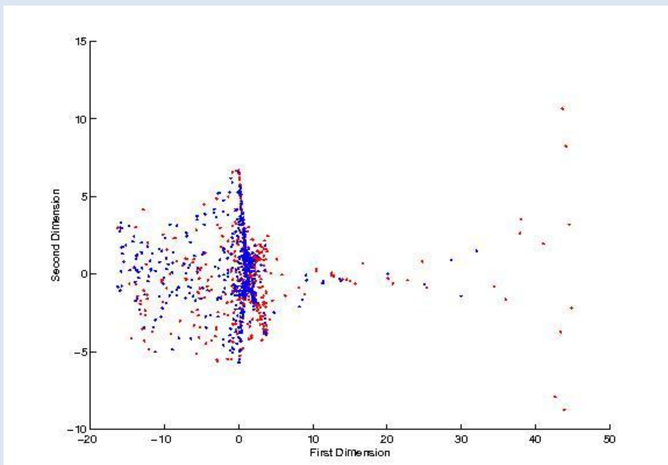




Blue: non-interacting protein pairs , Red : interacting protein pairs

Multidimensional Scaling

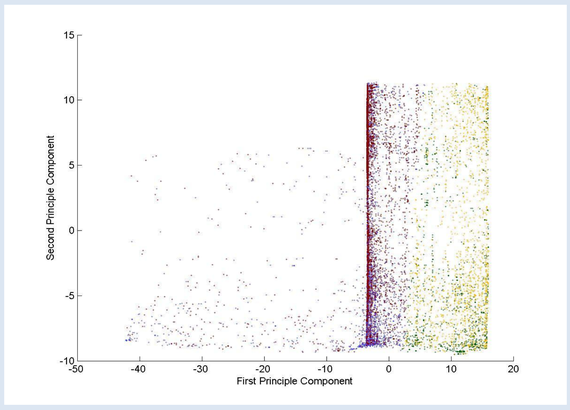
For finding non-linear patterns in data, multidimensional scaling is applied on the dataset. The number of instances have been reduced because of the computational complexity of the method.

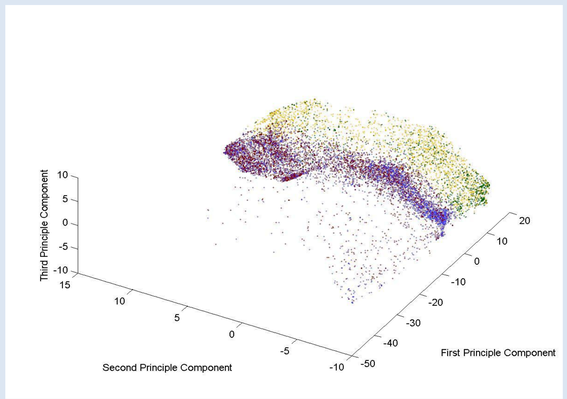


Blue: non-interacting protein pairs , Red : interacting protein pairs

K-means clustering

For finding out a direct separation of the data with an unsupervised clustering method, k-means clustering is applied. Euclidean distance is used for the computation of the distance matrix. 2 clusters created this way are compared to real labels of the data.

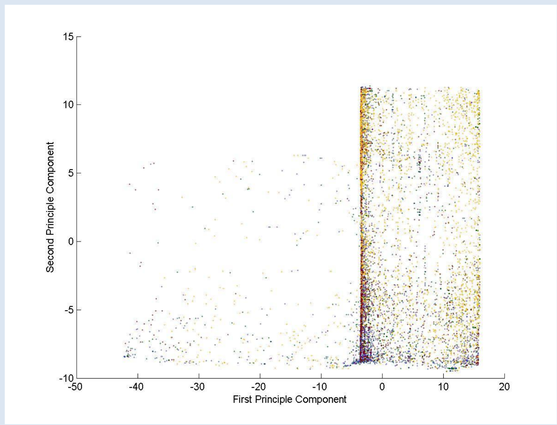


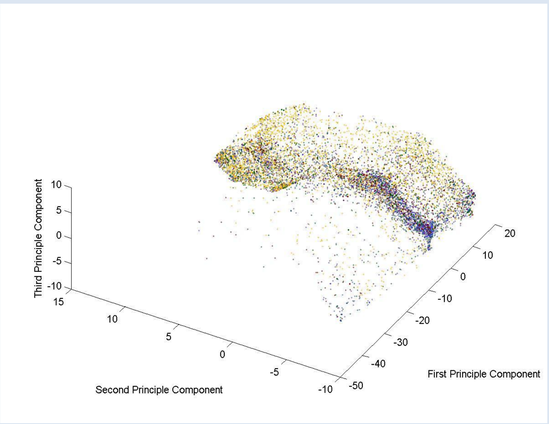


Blue: Interacting proteins in Cluster 1 , Green: Interacting proteins in Cluster 2
Red: Non-interacting proteins in Cluster 1 , Yellow : Non-interacting proteins in Cluster 2

Support Vector Machines

For performing supervised clustering with the labeled dataset, support vector machines are applied using the radial basis function as the kernel. In the best case, 64% prediction accuracy is reached with this kernel.





Blue: Interacting proteins in Cluster 1 , Green: Interacting proteins in Cluster 2
Red: Non-interacting proteins in Cluster 1 , Yellow : Non-interacting proteins in Cluster 2

CONCLUSION

Some predictions about protein interactions can be made with the application of mentioned methods. Statistical data analysis techniques show that some predictions can be made about the interaction status of a given protein pair by looking at the existence of the proteins in different species.