Prediction of protein-protein interactions through support vector machines

J. D. Arango Rodríguez, J. A. Jaramillo-Garzón and J. C. Arroyave-Ospina
Grupo de Automática, Electrónica y Ciencias Computacionales.
Instituto Tecnológico Metropolitano, Calle 73 No 76A-354, Medellín, Colombia
julianarango135888@correo.itm.edu.co, jorgejaramillo@itm.edu.co, johanaarroyave@itm.edu.co

Abstract

In this paper, a SVM-based method is implemented for the prediction of protein-protein interactions. This model is initially trained with a set of over 69.000 pairs of protein sequences based on documented positive interactions. Then, a cross-validation method is performed for estimating the accuracy of the system, showing acceptable performances in terms of sensitivity, specificity and geometric mean. The results are approximately balanced and the overall performance if around 70% *classified through a pairwise kernel and the parameters are set through an particle swarm optimization meta-heuristic and showing promising results for the field of bioinformatics.*

1. Introduction

Most proteins need to interacting with other proteins in order to perform their functions. Thus, information about their interactions can explain many cellular processes, then is very useful to find out the action mechanism of some diseases [1].

Protein-protein interactions (PPI) are physical contacts between a pair of proteins allowing them to perform a biochemical event that takes effect in several cellular processes. These interactions can influence the behavior of multiple cell signaling pathways and its prediction can provide useful information for discovering the unknown mechanisms in molecular events such as effects in cellular metabolism or cell trafficking [2]. Currently, there are some methods for the prediction of physical interactions that consider the evolution of the genes order across genomes, divergence of proteins across species, and the fusion of two separate proteins in a single protein. However, since the use of experimental methods can be expensive and time-consuming, there have been proposed several methods to predict protein-protein interactions from computational tools. These methods can be based on statistics, phylogenetic profiling and machine learning, among others. Such non experimental methods can provide good predictions based on previously reported experimental data[3, 4, 5].

There are some approaches for characterization of biochemical protein-protein interactions such as protein microarrays, which can detect interactions in vitro, but currently is only being applied to calmodulin and phospholipid binding proteins to domain interactions [6]. Also, mass spectrometry analysis of purified protein complexes is a powerful and sensitive tool but it has some drawbacks, for example: in the complex purification approach it is expressed just one of the interactors, this approach is more physiological because the analysis of PPI is done with direct and cooperative combinations, and finally, this can be more expensive than other methods [7]. On the other hand, computational prediction of protein-protein interactions has achieved increasing attention in the last years. Machine learning methods have been used to this purpose but they have reached low prediction performances, mostly because their free parameters are not properly sett[8].

In this paper it is used a support vector machine for the prediction of interactions between two proteins. Proteins pairs are classified through a pairwise kernel, and additionally the parameters are set through an particle swarm optimization meta-heuristic.

1.1. Background

1.2. Proteins

Proteins are polymers of amino acids, with each amino acid residue joined to its neighbor by a specific type of covalent bond. Twenty different amino acids are commonly found in proteins and these have a carboxyl group and an amino group bonded to the same carbon atom. They differ from each other in their side chains, or R groups, which vary in structure, size, and electric charge, and which influence the solubility of the amino acids in water. In addition to these 20 amino acids there are many less common ones [9].

Of all the molecules encountered in living organisms, proteins have the most diverse functions, like as cataly-

978-1-4673-9461-1/15/$31.00 ©2015 IEEE

sis, the catalytic proteins called the enzymes accelerate thousands of biochemical reactions in such processes as digestion, energy capture, and biosynthesis. The structural proteins often have very specialized properties in the cell. Proteins of movement are involved in all cell movements. Actin, tubulin, and other proteins comprise the cytoskeleton and other functions as defense, transport, stress response and regulation. The latter is related to binding a hormone molecule or a growth factor to cognate receptors on its target cell changes cellular function [10].

1.3. Physical methods

Initially every cellular function requires physical protein-protein interactions (PPIs) between cellular proteins and cellular functions are critically dependent on the correct assembly of proteins to become functional multiprotein complexes, where there is dynamic interchange of complex components in response to signals, from internal molecular cellular demands, or a cell environment. Also, the correct functioning of signaling pathways, transmitting signals from cell surface receptors via kinase networks to the nucleus, requires multiple sequential and transient interactions between upstream and downstream components of the particular pathway [11].

There are several techniques to know when exists a interaction between two proteins like as yeast two-hybrid, the modular basis of eukaryotic transcription factors provided the biological flexibility contributing to the development of the yeast two-hybrid technique. The base components of yeast two-hybrid hinge on transcription activator proteins (Young, 1998). The affinity purification coupled to mass spectrometry is based on important cellular processes, such as transcription, replication, and recombination, involve the action of DNA binding proteins, to study the biochemical properties of these transcription factors, it is necessary to purify the proteins to homogeneity. This would enable the factors to be characterized, facilitate the raising of antibodies, and ultimately provide a means for cloning the genes encoding these regulatory proteins [12]. And one of the most important methods is the co-inmunoprecipitation generated a novel genetic system to study these interactions by taking advantage of the properties of the GAL4 protein of the yeast Saccharomyces cerevisiae. This protein is a transcriptional activator required for the expression of genes encoding enzymes of galactose utilization. It consists of two separable and functionally essential domains: an N-terminal domain which binds to specific DNA sequences (UASG); and a C-terminal domain containing acidic regions, which is necessary to activate transcription. If both proteins form a complex through GAL4 activating region, the known protein could be interact with the other protein [13].

1.4. Non-physical methods

In other cases, is used the bioinformatic methods, among which are: Phylogenetic profiling that detects proteins that participate in a common structural complex or metabolic pathway. Proteins within these groups are defined as functionally linked. The underlying hypothesis is that functionally linked proteins evolve in a correlated fashion, and, therefore, they have homologs in the same subset of organisms, and the homologs its potentially to interact [14].

The methods based on similar phylogenetic trees, in this process whereby two or more species interact and influence genetic changes in one another. The process is also evident at the molecular level, where interacting proteins exhibit coordinated mutations to evolve at a similar rate, when a mutation exists changes occur within inter-protein contact sites or at regions implicated in the structural integrity of proteins, and this is taken into account to define the interaction [15].

The identification of structural patterns consider that the hot spots contribute dominantly to protein-protein interactions and has a significant energetic contribution to protein associations, the residue identity, size and charge, and the interactions it establishes with its neighboring residues should be crucial. This consists on the prediction of these hotspots of a one protein in order to analyze which is the most important to interact with a set of proteins [16]. And by las the classification methods are used to train a machine (classifier) to distinguish when exists a interaction between a pair of proteins. Each protein is represented as a vector and is necessary that the data have some features to the machine can distinguish between false or true [17].

1.5. Support vector machines

The support vector machine (SVMs) is a binary classifier, usually use for pattern recognition, and especially for the two-class classification problem [17]. This use critical features for the training group, these are much important to the accuracy in the SVM, if these are specific, the classification will be much better. For the classification with two-class problem is necessary assume a set of samples and consider definite features for a correct distinction. SVM implements so: It maps the input vectors into a high dimensional feature space and finally constructs an optimal separating hyperplane and which maximizes the margin, the distance between the hyperplane and the nearest data points of each class in the space [18].

The kernel function is based on information-geometric consideration of the structure of the Riemannian geometry induced by the kernel. The idea is to enlarge the spatial resolution around the boundary by a conformal transformation so that the separability of classes is increased and usually the kernel is then modified conformally in a data dependent way by using the information of the support vectors [19]. Aditionally, the SVM and other kernels methods derive their ability from the incorporation of prior knowledge via the kernel function and offers the application of diverse types of data.

2. Proposed methodology

Figure shows the overall proposed methodology. Each stage will be explained in the present section.

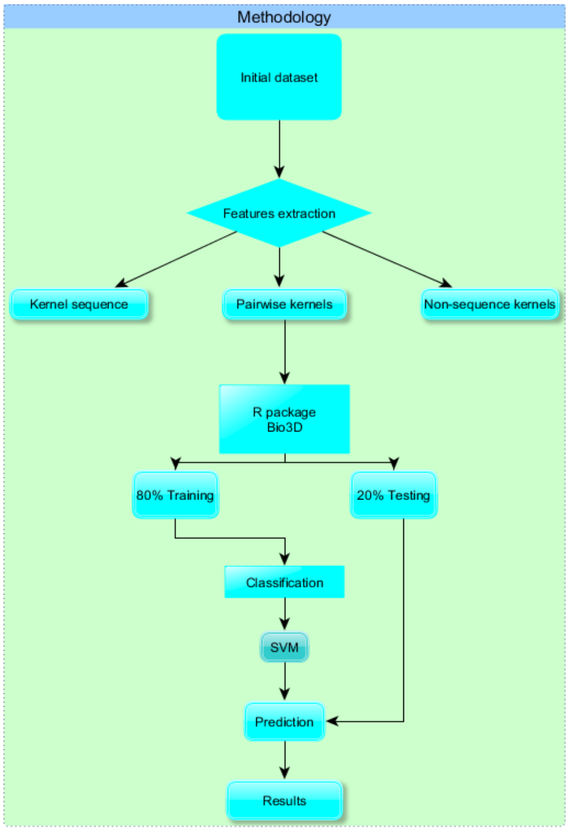

Figure 1. Flowchart of proposed methodology using learning methods.

2.1. Dataset

The dataset used in this paper was collected from just interactions from yeast proteinprotein interaction data collected in other research [20]. In this data contains several types of interactions like as: high-throughput yeast two-hybrid, correlated mRNA expression, genetic interaction (synthetic lethality), tandem affinity purification, high-throughput massspectrometric protein complex identification and other computational methods. Aditionally, some interactions was consider false positives because the interactions was classified mediating confidence scale, it depend of the method employed, and the low confidence is consider negative or unknown [8]. In this study, were taken the labels of the proteins that interact in the research of Mering for the respective characterization.

Then, with the initial dataset were obtained 69.723 positive interactions, whence it had to make more or less the same quantity of negative randomly interactions, for this case were used 69.770 this was done choosing a pair of proteins and making sure that these do not interact, and so on. All this is necessary to avoid a current drawback in the implementation of SVMs that is the class imbalance, because the classifiers generally perform poorly on imbalanced datasets[21].

2.2. Features extraction and kernel matrix construction

For the representation of a protein, we consider a several feature for pick out and test in the classification for the feature extraction we use the pairwise kernel that provides a similarity between pair of proteins (Specifically a pair of sequences). It is computed using R programming and software environment [22] with a package called Bio3D [23], this search a function that compare a pair of proteins to establish a similarity between a parameter by means a database. The model used by means of R with the package, is described in the equation 1 where X_1 is similar to X_1' and X_2 is similar to X_2' called pairwise kernel function [24]. The results is shown in order to relevant with the significant pairwise that is related to the possibility of a protein to interact with other. Each value obtained from this extraction is useful such as characteristic that resulting from kernel matrix, these values are important to predict the interaction between a pair of proteins because this will be the condition.

$$K((X_1,X_2),(X_1',X_2')) = K'(X_1,X_1')K'(X_2,X_2') + K'(X_1,X_2')K'(X_2,X_1') \quad (1)$$

This expression shows the way to construct a pairwise kernel to express the similarity between two pair of proteins in terms of similarities between individual proteins.

2.3. Classification

For the binary classification, it employs an algorithm which it is used when a protein pair can be interact or could not do, thus, these not form complex. It is necessary that the parameter settings should be tuned to get the best results in any case, and the kernel function should be included too. As shown in figure 1, 80% of the data is used for the training and more specifically for the classification, and 20 % is used to the prediction with the SVM that is the stage of testing. Finally, after testing, the obtained results are analyzed.

3. Results

Table 1 shows the specificity, sensitivity and geometric mean obtained for each fold in the cross-validation process.

Table 1. Detailed results in the cross-validation procedure

Fold	Sensitivity	Specificity	Geometric Mean
1	0.689	0.710	0.689
2	0.64	0.73	0.697
3	0.81	0.72	0.734
4	0.52	0.90	0.684
5	0.61	0.821	0.707
Mean	0.654 ± 0.107	0.776 ± 0.082	0.7022 ± 0.019

As can be observed, results are very consistent through all folds, with low standard deviations in sensitivity, specificity and geometric mean. Although the system is slightly more specific than sensitive, the results are approximately balanced and the overall performance if around 70%.

4. Conclusions

In this paper it is used a support vector machine to the prediction of interactions between two proteins, Proteins pairs are classified through a pairwise kernel, showing promising results for the field of bioinformatics. The performance achieved by the machine (Aproximately 70%) is considerable in terms of the features taken into account. Also, some methods based on sequence and attraction-repulsion models but uses other types of characterization.

As future work, the system is intended to analyze possible relationships between virus proteins and other protein sequences related with cellular proliferation or apoptosis, in order to establish its potential oncogenic capabilities. Alike, the results shows a similar performance like as methods focused in features while employed, considering that this method just uses one feature. This can lead to use two characteristics to improve the methods based on machine learning.

5. Acknowledgements

This work was funded by Instituto Tecnológico Metropolitano under the project ”Estudio de la regulación de la proliferación celular y apoptosis dependiente de la proteína NS5A del virus GB tipo C” with code P14225.

References

[1] P. Uetz, L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, J. R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart *et al.*, “A comprehensive analysis of protein–protein interactions in saccharomyces cerevisiae,” *Nature*, vol. 403, no. 6770, pp. 623–627, 2000.

[2] E. M. Marcotte, M. Pellegrini, H.-L. Ng, D. W. Rice, T. O. Yeates, and D. Eisenberg, “Detecting protein function and protein-protein interactions from genome sequences,” *Science*, vol. 285, no. 5428, pp. 751–753, 1999.

[3] F. Pazos and A. Valencia, “Similarity of phylogenetic trees as indicator of protein–protein interaction,” *Protein engineering*, vol. 14, no. 9, pp. 609–614, 2001.

[4] M. Deng, K. Zhang, S. Mehta, T. Chen, and F. Sun, “Prediction of protein function using protein-protein interaction data,” *Journal of Computational Biology*, vol. 10, no. 6, pp. 947–960, 2003.

[5] J. R. Bock and D. A. Gough, “Predicting protein–protein interactions from primary structure,” *Bioinformatics*, vol. 17, no. 5, pp. 455–460, 2001.

[6] H. Zhu, M. Bilgin, R. Bangham, D. Hall, A. Casamayor, P. Bertone, N. Lan, R. Jansen, S. Bidlingmaier, T. Houfek *et al.*, “Global analysis of protein activities using proteome chips,” *science*, vol. 293, no. 5537, pp. 2101–2105, 2001.

[7] G. Drewes and T. Bouwmeester, “Global approaches to protein–protein interactions,” *Current opinion in cell biology*, vol. 15, no. 2, pp. 199–205, 2003.

[8] S. M. Gomez, W. S. Noble, and A. Rzhetsky, “Learning to predict protein–protein interactions from protein sequences,” *Bioinformatics*, vol. 19, no. 15, pp. 1875–1881, 2003.

[9] D. L. Nelson, A. L. Lehninger, and M. M. Cox, *Lehninger principles of biochemistry*. Macmillan, 2008.

[10] S. Damodarn, “Amino acids, peptides and proteins in food chemistry,” 1996.

[11] J. Westermarck, J. Ivaska, and G. L. Corthals, “Identification of protein interactions involved in cellular signaling,” *Molecular & Cellular Proteomics*, vol. 12, no. 7, pp. 1752–1763, 2013.

[12] J. T. Kadonaga and R. Tjian, “Affinity purification of sequence-specific dna binding proteins,” *Proceedings of the National Academy of Sciences*, vol. 83, no. 16, pp. 5889–5893, 1986.

[13] S. Fields and O.-k. Song, “A novel genetic system to detect protein protein interactions,” 1989.

[14] M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates, "Assigning protein functions by comparative genome analysis: protein phylogenetic profiles," *Proceedings of the National Academy of Sciences*, vol. 96, no. 8, pp. 4285–4288, 1999.

[15] S.-H. Tan, Z. Zhang, and S.-K. Ng, "Advice: automated detection and validation of interaction by co-evolution," *Nucleic acids research*, vol. 32, no. suppl 2, pp. W69–W72, 2004.

[16] O. Keskin, B. Ma, and R. Nussinov, "Hot regions in protein–protein interactions: the organization and contribution of structurally conserved hot spot residues," *Journal of molecular biology*, vol. 345, no. 5, pp. 1281–1294, 2005.

[17] V. N. Vapnik and V. Vapnik, *Statistical learning theory*. Wiley New York, 1998, vol. 1.

[18] S. Hua and Z. Sun, "Support vector machine approach for protein subcellular localization prediction," *Bioinformatics*, vol. 17, no. 8, pp. 721–728, 2001.

[19] S.-i. Amari and S. Wu, "Improving support vector machine classifiers by modifying kernel functions," *Neural Networks*, vol. 12, no. 6, pp. 783–789, 1999.

[20] C. Von Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields, and P. Bork, "Comparative assessment of large-scale data sets of protein–protein interactions," *Nature*, vol. 417, no. 6887, pp. 399–403, 2002.

[21] R. Akbani, S. Kwek, and N. Japkowicz, "Applying support vector machines to imbalanced datasets," in *Machine Learning: ECML 2004*. Springer, 2004, pp. 39–50.

[22] R. Ihaka and R. Gentleman, "R: a language for data analysis and graphics," *Journal of computational and graphical statistics*, vol. 5, no. 3, pp. 299–314, 1996.

[23] B. J. Grant, A. P. Rodrigues, K. M. ElSawy, J. A. McCammon, and L. S. Caves, "Bio3d: an r package for the comparative analysis of protein structures," *Bioinformatics*, vol. 22, no. 21, pp. 2695–2696, 2006.

[24] A. Ben-Hur and W. S. Noble, "Kernel methods for predicting protein–protein interactions," *Bioinformatics*, vol. 21, no. suppl 1, pp. i38–i46, 2005.