International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:3, No:9, 2009

Selecting Negative Examples for Protein-Protein Interaction

Mohammad Shoyaib, M. Abdullah-Al-Wadud, and Oksam Chae

***Abstract*—**Proteomics is one of the largest areas of research for bioinformatics and medical science. An ambitious goal of proteomics is to elucidate the structure, interactions and functions of all proteins within cells and organisms. Predicting Protein-Protein Interaction (PPI) is one of the crucial and decisive problems in current research. Genomic data offer a great opportunity and at the same time a lot of challenges for the identification of these interactions. Many methods have already been proposed in this regard. In case of in-silico identification, most of the methods require both positive and negative examples of protein interaction and the perfection of these examples are very much crucial for the final prediction accuracy. Positive examples are relatively easy to obtain from well known databases. But the generation of negative examples is not a trivial task. Current PPI identification methods generate negative examples based on some assumptions, which are likely to affect their prediction accuracy. Hence, if more reliable negative examples are used, the PPI prediction methods may achieve even more accuracy. Focusing on this issue, a graph based negative example generation method is proposed, which is simple and more accurate than the existing approaches. An interaction graph of the protein sequences is created. The basic assumption is that the longer the shortest path between two protein-sequences in the interaction graph, the less is the possibility of their interaction. A well established PPI detection algorithm is employed with our negative examples and in most cases it increases the accuracy more than 10% in comparison with the negative pair selection method in that paper.

***Keywords*—**Interaction graph, Negative training data, Protein-Protein interaction, Support vector machine.

I. Introduction

PROTEIN-Protein Interaction (PPI) plays vital roles for many fundamental biological processes of living cells. Thus identifying these interactions are very much important for understanding the functions and physiological phenomenon of proteins for the discovery of novel medicines and protein based products with medical and industrial applications. Despite the high importance of recognizing the PPI, very little has been known so far, as the experimental approaches for PPI identification are very expensive and laborious. To address this tedious, labor-intensive and costly technique, researches are recently seeking for efficient computational methods to predict whether two proteins will interact or not. A better computational approach may easily substitute unnecessary experimental procedures, and thus save cost and increase the confidence of the experimental results.

In recent years, a number of computational methods for predicting Protein- Protein interaction based on different criteria for instance, amino acid sequence, structure etc., has been proposed [1-10, 32]. Most of these methods are composed of two phases: a training phase — to train the system/classifier with some known interacting as well as non-interacting protein pairs; and the testing phase – when given a pair of proteins, the system will predict whether they will interact or not. In the training phase, only positive examples (interactive protein pairs only) [15-17], or both positive and negative examples (Non-interactive protein pairs) [1-10, 32] may be used. Xiao et al. [8] argue that the usage of only positive examples for training phase may derive many false positive domain pairs, because these domain pairs may occur in the (unavailable) negative set with high frequency. High quality positive dataset and good algorithms may compensate here and achieve better results.

In the cases, where both positive and negative examples are used in training, there are two potential challenges. One is the generation of good positive examples – information about this may be gathered from the available renowned databases. The second one is the generation of negative examples, which is more difficult and also crucial for reliable prediction of PPI as negative examples are not available in general. There is also no benchmark of selecting dataset of non-interacting protein sequences. Moreover, most of the available well known databases do not give any information on non-interacting proteins. To validate the PPI prediction algorithms, researchers usually adopt their own techniques in generating negative examples based on some assumptions. However, these negative data are likely to affects their results greatly.

A good set of negative examples may increase the accuracy, even for the currently available methods for predicting PPI. In this paper, this issue is addressed and a general solution is proposed in this regard. A graph based solution for finding non-interacting protein sequences is proposed, which can be used for more successful training of the available techniques and significantly increase the accuracies of these methods.

Mohammad Shoyaib is with the Department of Computer Engineering, Kyung Hee University, 1 Seocheon, Giheung, Yongin, Gyonggi, Korea (e-mail: shoyaib@khu.ac.kr).

M. Abdullah-Al-Wadud is with the Department of Industrial and Management Engineering, Hankuk University of Foreign Studies, 89 Wangsan, Mohyun, Cheoin, Yongin, Gyonggi, Korea (e-mail: wadud@hufs.ac.kr).

Oksam Chae is with the Department of Computer Engineering, Kyung Hee University, 1 Seocheon, Yongin, Gyonggi, South Korea, 449-701 (corresponding author, phone: 82-31-201-2948; fax: 82-31-202-1723; e-mail: oschae@khu.ac.kr).

II. Literature Review

Most of the successful computational methods use both positive and negative examples for PPI prediction. For these methods, the collection of interacting protein pairs is relatively easier as because several well known and reliable databases are already published that are dedicated for PPI and they try to maintain high quality of interaction information. For instance., the Munich Information Center for Protein Sequence (MIPS) [19], the Database of Interacting Proteins (DIP)[20], the Biomolecular Interaction Network Database (BIND)[21], the Human Protein Reference Database (HPRD)[22], the Molecular Interaction database (MINT)[23], the Biological General Repository for Interaction Datasets (BioGRID) [24] etc. In spite of their best effort, information related to PPI is often incomplete and contradictory [25-27]. Further, [28] also mentions five challenging properties of genomic/proteomic data related to PPI and they are: (i) no reliable reference set, (ii) little overlap between different data sets, (iii) rarity of protein–protein interactions, (iv) missing data in protein annotations and experimental measurements and (v) noisy data. Despite of the aforementioned problems, good quality positive data is still achievable for the computational approach of PPI detection.

For Gold Standard Positives (GSP) data three criteria are mentioned in [29] and they are: A GSP should be as unbiased as possible, sampling all, or at least most, parts and processes of the cell [30], and a GSP must be of the highest reliability and reproducibility [31]. They [29] also give a nice demonstration for collecting good set of positive data.

The generation of Gold Standard Negatives (GSN) is more difficult, because the pairs tested and found not to interact are almost never reported [29]. But this is very much essential for successful training. The current methods for this can be divided into two broad categories. First, random pairing of protein to generate negative examples [5, 32-38] and the second, synthesizing negatives from proteins that are not co-localized i.e., proteins which are localized at different sub-cellular components [3, 8-12, 30] .

Ben-Hur et al. [32] advocate for a simple uniform random choice of non-interacting protein pairs from the set of all protein pairs, which are not known to interact. They state that this will yield an unbiased estimate of the true distribution. They also demonstrate that restricting negative examples to non-colocalized protein pairs leads to a biased estimate of the accuracy of a predictor of protein-protein interactions. Chen et al. [5] and Zaki et al. [6] also follow the same procedure for non – interacting protein pairs.

In the aforementioned cases, random pairs are selected from the set of all proteins but Shen et al. [1] consider those protein that appear in the positive data set only and a negative candidate pair is chosen in an exclusive way.

In [7], a slightly different procedure is found. They create negative controls by randomizing amino acids sequences sampled from DIP [20] and while doing so, they ensure the preservation of the following two things (1) amino acid composition and (2) di- and tri-peptide '*k*-let' frequencies [13-14], where $k > 1$.

Unlike the random pairing to generate non-interacting protein pairs as negative training data, Xiao et al. [8] generate biologically meaningful negative examples based on the proteins' biological information, namely, proteins from different cellular locations and functional activity. Further they argue that randomly generated negative dataset may contain unknown interacting protein pairs. Thus it may contaminate the training dataset and results may degrade. [3] and [10] also use negative examples from a list of proteins that are in separate sub cellular compartments. Similar approach is also followed by Rhodes et al. [9]. They identified GSN interaction set for HPRD [22] data set. Here GSN is a set of all protein pairs in which one protein is assigned the plasma membrane cellular component (1,426 proteins) and the other in nuclear cellular component (2,253 proteins).

GRIP [18] is a web based tool for generating gold standard dataset. It offers the both options (random or different localization of protein) to the user for the generation of negatives in case of MIPS [19] data source. That is, user can either choose a negative case consisting of a list of proteins obtained from different sub-cellular locations and complexes or can get a list of proteins randomly selected from sub-cellular locations and complexes. For the BioGRID [24] data source, GRIP defines 'negative' cases by the random selection of protein pairs only.

From the above discussion, it is observed that so far most of the PPI detection methods use either of the two methods for negatives. However, these methods may generate erroneous or biased estimate. To solve the problem, a graph based method is proposed which may hold good criteria of both while reducing the contamination and biasing tendency.

III. Proposed Method for Selecting Non-interacting Protein Pair

To describe the interactions among the proteins, let us first have a look at few related terminologies that are used in this paper.

Interaction Graph(IG): A graph G(V,E) is called an *Interaction Graph*, where V denotes the set of protein sequences and E denotes the set of edges denoting the pair of protein sequences that interact.

Let *A*, *B*, *C*, and *D* be four protein sequences, where the pair of interacting protein sequences are (*A*, *C*), (*A*, *D*), (*B*, *D*), (*C*, *D*). Let us also assume that no information is available about interaction between *A* and *B* as well as *B* and *C*. The IG for this scenario looks like the graph in Fig. 1.

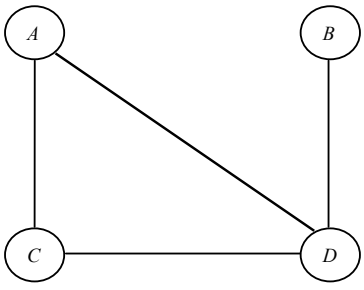

Fig. 1 An example Interaction Graph

Interaction Path: An Interaction Path is the shortest path between any two vertices in an IG. For example, the interaction path between A and B is (A, D, B).

Interaction Distance: Interaction distance $d(x, y)$ between two vertices x and y in an IG is defined as the length of an interaction path between them. In fact, $d(x, y)$ equals to the minimum number of edges need to traverse to move from x to y, or vice versa. For example, in Figure 1, $d(A,B) = 2$ and $d(A,C) = 1$.

As mentioned earlier, one widely used approach for selecting negative examples is to pick necessary number of pair of protein sequences randomly from all pair of sequences, for which no interaction data is currently at hand. This process of selection has a risk of selecting some unknown positive pair. To reduce such contamination in negative examples, a graph based approach is used. It can be said that the likelihood of interaction of two protein sequences decreases as the Interaction Distance between them increases in the corresponding IG. Hence, two protein sequences x and y can be considered as totally non-interacting if $d(x,y)=\infty$, i.e., they have no path in IG. Such pair of vertices can serve as negative dataset in the training data. In such cases, there is a high chance that many pair will represent non-colocalized proteins, similar to the method presented in [3, 8-12, 30]. However, such extreme set of non-interacting examples may demonstrate some biasness as because it does not include the non-interacting sequences that are co-localized. Hence, considering two protein sequences as non-interacting, when Interaction Distance between them is sufficiently large, may be a better approach.

Algorithm 1 Finding interaction paths.

Input parameters
M : Adjacency matrix of the IG
Output parameters
M_η : Final matrix
Auxiliary parameters
M_{Now} and M_{Prev} : Intermediate matrices.

Procedure

```
1.  Initialize M_Prev to M;
    /*Find the paths*/
2.  for l = 1 to η - 1 do
3.      if (l mod 2) = 1 then
4.          FindLongerPaths(M_Now, M_Prev, M);
5.      else
6.          FindLongerPaths(M_Prev, M_Now, M);
7.      end if
8.  end for
    /*Now copy the resulting matrix*/
9.  if (η mod 2) = 1 then
10.     Copy M_Prev to M_η ;
11. else
12.     Copy M_Now to M_η ;
13. end if
```

The protein sequences that are present in the positive dataset are first picked up and an IG is generated using them. Algorithm 1 then finds all pair (x, y) of protein sequences, where $d(x,y) \leq \eta$. This information actually provides us all possible the non-interacting pair (x, y) of protein sequences, where $d(x,y) > \eta$. Necessary training pairs are then randomly picked from this set of all possible pairs.

The adjacency matrix M has true values at the entries for the pairs (x, y) of protein sequences, where $d(x,y) = 1$ (i.e., the interacting pairs). Then, every iteration of the **for**-loop in Algorithm 1 marks entries for all pair (x, y) of protein sequences, where $d(x,y)=l+1$. The function *FindLongerPaths*(M_{Now}, M_{Prev}, M) actually performs these operations. It finds the paths in the IG having length l+1 with the help of all the paths with length less than or equal to l, which are already retrieved in previous steps. The algorithm for this function is presented in Algorithm 2.

Algorithm 2 Finding paths of length l+1.

Input parameters
M : Adjacency matrix of the IG.
M_{Prev} : The matrix holding the interaction paths whose lengths are less than or equal to l.
m : Total number of vertices in an IG.
Output parameters
M_{Now} : The matrix holding the interaction paths whose lengths are less than or equal to l+1.

Procedure

```
1.  for i = 0 to m-1 do
2.      for j = i+1 to m-1 do
3.          if M_Prev(i, j) = true then
            /*a path exists with length ≤ l*/
4.              M_Now (i, j) = true;
5.              M_Now (j, i) = true;
6.              Continue with the next j;
7.          end if
        /* assume no path exists*/
8.          M_Now (i, j) = false;
9.          M_Now (j, i) = false;
10.         for k = 0 to m-1 do
11.             if M_Prev(i, k) = true
                    and M_Prev(k, j) = true then
                /*a path exists with length ≤ l+1*/
12.                 M_Now (i, j) = true;
13.                 M_Now (j, i) = true;
14.                 Continue with the next j;
15.             end if
16.         end for
17.     end for
18. end for
```

After the execution of Algorithm 1, a *false* entry $M_\eta(i, j)$ represents that the corresponding nodes are more than η interaction distance apart, i.e., $d(i, j) > \eta$. Such pair of nodes are good candidates for non-interacting protein pairs. Selecting necessary number of pairs from them randomly also maintains

the uniformity in the set. Thus this method is capable of picking some unbiased and meaningful negative pairs for the training of the PPI prediction methods.

In our proposed methodology, PPI data is represented as interaction graph. As in [40], it is also considered that edge between two nodes represents evidence that they might share the same function. Now if the interaction distance (number of edges in the shortest path) between two nodes is increased, then the possibility of sharing same function between these two nodes will go down, and hence the possibility of interaction will also go down, which implies that the potential of taking these nodes as negative example will increase. By defining η, those pairs are taken that have a limited possibility of having same functions in them. This allows picking good non-interacting examples from similar sub-cellular locations. On the other hand, this way of choosing negatives also have a tendency of giving pair from different sub-cellular locations, i.e., the method does not ignore non-colocalized pair as well.

Ben-Hur et al.[32] advocate for random pairing for non-interacting proteins. However, they also agree that this simpler method has potential pitfalls as they may contaminate the negatives with positive examples. Our proposed method can avoid such contamination by filtering them out using Algorithm 1 before taking random pairs.

Thus the proposed approach is able to retrieve meaningful and useful negative examples. It generates negative examples that inherit good characteristics from both the random selection methods and the approaches of selecting non-colocalized proteins, while omitting the biases and ensuring better classification performance of PPI identifications. The experimental results also support this.

IV. Comparative Analysis and Discussion

A number of good algorithms are available for the computation of PPI as mentioned earlier. Among them, a well accepted methodology proposed by Shawn Martin *et al.* [2] is picked to show the strength of our negative examples. Before presenting the experimental results, the data sources and evaluation criteria used to explain the comparative performance analysis are described first.

A. Data Sources

Interacting protein sequence data are collected from DIP [20] which contains interaction information of different species. Thus interacting protein sequences of *E. coli*, *C. elegans*, *H. sapiens* (Human), *M. musculus* (house mouse) are collected from DIP. *H. polori* data provided by Shawn Martin *et al.* [2] are also used. For negative examples, our proposed method as well as the method of random generation of protein pairs among all possible pair that does not appear in the positive set is used. Each species data (both negative and positive) is divided into two disjoint sets for training and testing. 90% of data is used in training and the rest 10% is used for testing purpose.

B. Evaluation Criteria

There are several standard performance measures to evaluate the classification results. Classification Rate (CR) — percentage of interaction and non-interaction correctly classified as a whole, Recall (R) and Precision (P) are used, which are defined by Eq. 1, Eq. 2 and Eq. 3.

$$CR = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$R = \frac{TP}{TP + FN} \quad (2)$$

$$P = \frac{TP}{TP + FP} \quad (3)$$

C. Discussion

The algorithm proposed by Shawn Martin et al. [2] is employed for PPI identification. The programs that are provided by the authors on the web are used. In the case of negative data selection, negative examples are randomly picked as mentioned in [2]. Negative examples are also generated using the proposed approach. The results of the overall accuracy, Precision and Recall for all the species are presented in Table I and Table II.

TABLE I
CR (%) for Different Species

Species	CR using proposed negative examples	CR using random negative examples
H. sapiens	83.33	73.04
D. melanogaster	97.26	70.67
E. coli	88.91	81.82
S. cerevisiae	97.68	74.44
C. elegans,	88.81	73.64
h. pilori	93.84	82.19
M. musculus	68.00	57.89

TABLE II
Precision and Recall for Different Species

Species	Proposed negative examples		Random negative examples	
	Precision (%)	Recall (%)	Precision (%)	Recall (%)
H. sapiens	83.33	63.73	80.52	60.78
D.melanogaster	97.59	96.91	72.29	67.03
E. coli	92.98	84.17	86.41	75.52
S. cerevisiae	97.68	98.44	76.26	70.98
C. elegans,	95.65	81.32	80.45	62.45
h. pilori	97.73	89.58	83.82	79.17
M. musculus	73.68	56.00	57.89	44.00

The tables given above show better result in every case, which demonstrates the enhanced performance of the proposed negative examples generation methodology over the approach of randomly choosing negative examples. It can also be shown that the proposed approach works better than selecting non-colocalized pair too. In this connection, authors claim that our negative examples will increase accuracy and reliability for any good algorithm.

V. Conclusion

In this paper, a new method is introduced for negative data selection which helps to predict PPIs efficiently. Major advantage of this method is the logical selection of negative data set. The system is capable of utilizing all the possible interactions. The results reported here demonstrate that this method can reliably enhance the accuracy in predicting protein-protein interaction. Hence, it can be expected that, the promising results based on the proposed method for negative example generation will improve the performance of protein interaction prediction and thus protein interaction network also.

References

[1] Juwen Shen, Jian Zhang, Xiaomin Luo, Weiliang Zhu, Kunqian Yu, Kaixian Chen, Yixue Li, and Hualiang Jiang, "Predicting protein–protein interactions based only on sequences information", *PNAS*, vol. 104, no. 11,pp. 4337–4341, 2007.

[2] Shawn Martin, Diana Roe and Jean-Loup Faulon, "Predicting protein–protein interactions using signature products", Bioinformatics, Vol. 21 no. 2 2005, pp. 218–226

[3] Jin Wang, Chunhe Li, Erkang Wang and Xidi Wang, "Uncovering the rules for protein–protein interactions from yeast genomic data", PNAS, 2009, vol. 106, no. 10 , pp. 3752–3757.

[4] Xue-wen Chen and Mei Liu, "Prediction of Protein-Protein Interactions Using Random Decision Forest Framework", Bioinformatics, 21(24), pp. 4394–4400, 2005.

[5] Nazar Zaki, Safaai Deris and Hany Alashwal, "Protein-Protein Interaction Detection Based on Substring Sensitivity Measure", International Journal of Biological and Medical Sciences, 1:2 2006

[6] Joel R. Bock and David A. Gough," Predicting protein–protein interactions from primary structure", Vol. 17 no. 5 2001 pp. 455–460

[7] Xiao-Li Li, Soon-Heng Tan, See-Kiong Ng, "Improving domain-based protein interaction prediction using biologically-significant negative dataset", *International Journal of Data Mining and Bioinformatics*, Vol-1, No.2 pp. 138 – 149, 2006.

[8] Daniel R Rhodes,Scott A Tomlins, Sooryanarayana Varambally, Vasudeva Mahavisno, Terrence Barrette, Shanker Kalyana Sundaram, Debashis Ghosh, Akhilesh Pandey and Arul M Chinnaiyan, "Probabilistic model of the human protein-protein interaction network", Nature Biotechnology 23, 2005, pp. 951 - 959

[9] R. Jansen, H. Yu, D. Greenbaum, Y. Kluger, N.J. Krogan, S. Chung, A. Emili, M. Snyder, J.F. Greenblatt and M. Gerstein, "A Bayesian networks approach for predicting protein-protein interactions from genomic data", *Science,* 302: (5644), pp. 449-453, 2003.

[10] Lu LJ, Xia Y, Paccanaro A, Yu H and Gerstein M, "Assessing the limits of genomic data integration for predicting protein networks", Genome Res 2005, 15(7) pp. 945-953.

[11] Kumar,A., Agarwal,S., Heyman,John A., Matson S., Heidtman M., Piccirillo S., Umansky L., Drawid A., Jansen R., Liu, Y., Kei- Cheung H., Miller P., Gerstein M., Roeder G. S., and Snyder M., "Subcellular localization of the yeast proteome", Genes Dev., 16, 2002, pp. 707–719.

[12] E. Coward, "Shufflet: shuffling sequences while conserving the k-let counts", *Bioinformatics*, 15, pp. 1058–1059.

[13] D. Kandel, Y. Mathias, R. Unger and P. Winkler, "Shuffling biological sequences", *Discrete Appl. Math*., 71, pp. 171–185, 1996.

[14] M. Deng, F. Sun, S. Metha and T. Chen, "Inferring domain-domain interactions from protein-protein interactions", *Genome Research*, Vol. 12, pp.1540–1548, 2002.

[15] S.K. Ng, Z. Zhang, and S.H. Tan, "Integrative approach for computationally inferring protein domain interactions", *Bioinformatics*, Vol. 19, pp.923–929, 2003.

[16] Wan, K.K. and Jong, P., "Large scale statistical prediction of protein-protein interaction by potentially interacting domain (pid) pair",, Genome Informatics, Vol. 13, 2002, pp.45–50.

[17] Fiona Browne, Haiying Wang, Huiru Zheng and Francisco Azuaje, "GRIP: A web-based system for constructing Gold Standard datasets for protein-protein interaction prediction", Source Code for Biology and Medicine 2009, 4:2

[18] P. Pagel, S. Kovac, M. Oesterheld, B. Brauner, I. Dunger-Kaltenbach, G. Frishman, C. Montrone, P. Mark, V. Stümpflen, H.W. Mewes, A. Ruepp and D. Frishman, "The MIPS mammalian protein-protein interaction database", Bioinformatics, 21, pp. 832–834,2005.

[19] L . Salwinski, C.S. Miller, A.J. Smith, F.K. Pettit, J.U. Bowie and D. Eisenberg, "The Database of Interacting Proteins: 2004 update", *Nucleic Acids Res*, 32 Database issue:D449-51, 2004.

[20] Bader, G.D., Betel, D. and Hogue, C.W., "BIND: the Biomolecular Interaction Network Database", Nucleic Acids Res. 31, 2003, pp. 248–250.

[21] Mishra, G.R. et al., "Human protein reference database; 2006 update", Nucleic Acids Res. 34, D411–D414, Network Database. Nucleic Acids Res. 31, 2003, 248–250.

[22] A. Chatr-aryamontri et al. "MINT: the Molecular INTeraction database", *Nucleic Acids Res*. 35, D572–D574, 2007.

[23] T. Reguly et al., "Comprehensive curation and analysis of global interaction networks in Saccharomyces cerevisiae", *J. Biol.*, 5, 11, 2006.

[24] C. von Mering et al., "Comparative assessment of large-scale data sets of protein-protein interactions", Nature, 417, 2002, pp. 399-403.

[25] A. M. Deane, L. Salwinski, I. Xenarios, D. Eisenberg, *Mol. Cell. Proteomics 1*, 349, 2002.

[26] A.M. Edwards, B. Kus, R. Jansen, D. Greenbaum, J. Greenblatt and M. Gerstein, "Bridging structural biology and genomics: assessing protein interaction data with known complexes", *Trends Genet* 18, pp. 529-536, 2002.

[27] Jingkai Yu and Farshad Fotouhi, "Computational Approaches for Predicting Protein–Protein Interactions: A Survey", J Med Sys 30(1), 2006, pp. 39–44.

[28] Michael E Cusick, Haiyuan Yu, Alex Smolyar, Kavitha Venkatesan,Anne-Ruxandra Carvunis, Nicolas Simonis, Jean-François Rual, Heather Borick,Pascal Braun, Matija Dreze, Jean Vandenhaute, Mary Galli, Junshi Yazaki,David E Hill1, Joseph R Ecker, Frederick P Roth and Marc Vidal, "Literature-curated protein interaction datasets", *Nature Methods*, VOL.6 NO.1, JANUARY 2009.

[29] Jansen, R. and Gerstein, M., "Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction", Curr. Opin. Microbiol. 7, 2004, pp. 535–545.

[30] P. Braun et al., "An experimentally derived confidence score for binary protein-protein interactions", *Nat. Methods* 6, pp. 91–97, 2008.

[31] Ben-Hur A and Noble S, "Choosing negative examples for the prediction of protein-protein interactions", BMC Bioinformatics, 2006, 7:S2.

[32] S.M. Gomez, W.S. Noble and A. Rzhetsky, "Learning to predict proteinprotein interactions", *Bioinformatics*, 19:1875-1881, 2003.

[33] Ben-Hur A and Noble WS, "Kernel methods for predicting proteinprotein interactions", Bioinformatics, 2005, 21(suppl 1):i38-i46.

[34] Zhang LV, Wong S, King O and Roth F, "Predicting co-complexed protein pairs using genomic and proteomic data integration", BMC Bioinformatics, 2004, 5:38-53.

[35] Qi Y, Klein-Seetharaman J and Bar-Joseph Z, "Random Forest Similarity for Protein-Protein Interaction Prediction from Multiple Sources", Proceedings of the Pacific Symposium on Biocomputing 2005.

[36] Han, D., Kim, H., Jang, W. and Lee, S., "Domain combination based protein-protein interaction possibility ranking method",, IEEE Fourth Symposium on Bioinformatics and Bioengineering, 2004, pp.434–441.

[37] Han, D., Kim, H., Seo, J. and Jang, W. , "Domain combination based probabilistic framework for protein-protein interaction predication", Genome Informatics, Vol. 14, 2003, pp.250–259.

[38] Iakes Ezkurdia, Lisa Bartoli, Piero Fariselli, Rita Casadio, Alfonso Valencia and Michael L. Tress, "Progress and challenges in predicting protein-protein interaction sites", *Briefings In Bioinformatics*. vol 10. no 3., Advance Access publication April 3, 2009

[39] Stanley Letovsky and Simon Kasif, "Predicting protein function from protein/protein interaction data: a probabilistic approach", *Bioinformatics*, Vol. 19 Suppl. 1, pp. i197–i204, 2003.