A New Framework for Bridging the Gap from Protein-Protein Interactions to Biological Process Interactions

Christos Dimitrakopoulos[1], Andreas Dimitris Vlantis[1], Konstantinos Theofilatos[1], Spiros Likothanassis[1], and Seferina Mavroudi[1,2]

[1] Department of Computer Engineering and Informatics, University of Patras, Greece
[2] Department of Social Work, School of Sciences of Health and Care, Technological Educational Institute of Western Greece, Greece
{dimitrakop,vlantis,theofilk,likothan,mavroudi}@ceid.upatras.gr

Abstract. Proteins and their interactions have been proven to play a central role in many cellular processes and have been extensively studied so far. However of great importance, little work has been conducted for the identification of biological process interactions in the higher cellular level which could provide knowledge about the high level cellular functionalities and maybe enable researchers to explain mechanisms that lead to diseases. Existing computational approaches for predicting Biological Process interactions used PPI graphs of low quality and coverage but failed to utilize weighted PPI graphs to quantify the quality of the interactions. In the present paper, we propose a unified two-step framework to reach the goal of predicting biological process interactions. After conducting a comparative study we selected as a first step the EVOKALMA model as a very promising algorithm for robust PPI prediction and scoring. Then, in order to be able to handle weights, we combined it with a novel variation of an existing algorithm for predicting biological processes interactions. The overall methodology was applied for predicting biological processes interactions for Saccharomyces Cerevisiae and Homo Sapiens organisms, uncovering thousands of interactions for both organisms. Most of the linked processes come in agreement with the existing knowledge but many of them should be further studied.

Keywords: Protein-Protein Interactions, protein-protein interaction networks, Biological Process Interactions, EvoKalma Model, protein function, statistical enrichment.

1 Introduction

Among the numerous participants in molecular interactions, proteins are considered most important ones. In specific, proteins transmit regulatory signals throughout the cell, catalyze a huge number of chemical reactions, and are critical for the stability of numerous cellular structures. The total number of possible interactions within the cell is extremely large and the full identification of all true PPIs is a very challenging task. However, the identification of all true PPIs may contribute in understanding cellular functionality, designing more efficient medicines and uncovering the mechanisms that

L. Iliadis et al. (Eds.): AIAI 2014 Workshops, IFIP AICT 437, pp. 196–204, 2014.

lead to diseases. For these reasons, this problem has been extensively studied in the last decades and many experimental and computational techniques have been combined to solve it [1].

A Biological Process Network is a model designed to offer an insight at the interactions occurring among biological processes. The traditional approach for studying complex biological networks is based on the identification of interactions between genes and proteins. As a result, little is known about interactions of the higher order biological systems, such as biological processes. The knowledge derived by process interactions can be effectively used in protein function prediction, increasing both coverage and accuracy of predictions. Moreover, uncovering the interactions of biological processes can be a step towards understanding the cellular functions in a high level.

Existing methods for predicting interactions between biological processes take as inputs the PPI networks and functional annotations. In [2], processes are considered as interacting when more proteins annotated by them interact than expected by chance. Other methods also require gene expression information [3] to examine how interactions may change in different experimental conditions. In addition, this method incorporates weights at the protein interactions. In [4], the aim is to include as few inter-process links that successfully present as many gene interactions, to reduce complexity and allow further exploration in a greater detail. Despite the promising results of these approaches, their basic limitation is the utilization of PPI graphs of low quality and coverage as inputs. Moreover, most of them do not utilize confidence scores - weights for each PPI and others use confidence scores which only refer to the reference rate in the bibliography of a certain interaction.

In the present paper, we propose a computational framework for the prediction of PPIs, the prediction of a confidence score for each interaction which will reflect function, structural and sequential information and the prediction of interactions between biological processes. For the prediction and scoring of PPIs the method EVOKALMA was utilized [5]. The selection was only performed after the conduction of extended experiments in order to ensure the superior performance of the EVOKALMA model when compared with existing state of the art classification techniques. The experimental results confirmed that EVOKALMA method outperformed the other methodologies in all the examined metrics. For the prediction of biological process interactions we enhanced the methodology proposed in [2] in order to be able to handle weighted PPIs. In this way, the method accounted for the high false positive and false negative rates encountered in PPI datasets ensuring that only high-confidence interactions contribute significantly to the method's output while low-confidence interactions have only a low effect.

The proposed algorithmic framework was applied for Saccharomyces Cerevisiae and Homo Sapiens organisms and useful conclusions were made about the interactions of biological processes in these organisms.

The rest of the paper is organized as follows: In section 2 the datasets and the proposed algorithmic framework are described in detail. In section 3 the experimental results are presented and analyzed while in section 3 conclusions and proposals for future work are presented.

2 Datasets and Methods

2.1 Datasets

For the training and testing process of the methods for predicting PPIs 1000 positive interactions referred in HPRD [6] and 1000 negative protein interactions were selected. HPRD database is assumed to be highly reliable as it contains protein interactions that are supported by low and high throughput experimental evidence. The negative samples were created randomly from the unique identities of the whole set of proteins leaving out protein pairs which have been reported in iRefindex [7] as protein-protein interactions. For every protein pair in the deployed dataset 22 informative features were calculated including several co-expression features, function similarity features, sequence similarity, a homology based feature, domain-domain interaction feature and co-localization features. More details about the utilized features is available at [5]. All feature values are normalized in the range [0, 1] and missing are estimated using the kNN-impute methodology [8]. The trained EVOKALMA model (see section 2.2.1) was applied on an extended set of over 600000 protein pairs to predict most human PPIs and predicted 211367 PPIs. Moreover for every interaction a confidence score has been calculated. This confidence score indicates the strength of the interaction, its frequency and possibility to exist. These interactions alongside their weights (confidence scores) are stored in the HINT-KB.

The input graph datasets for predicting interactions among biological processes were: a) the protein interaction network for Saccharomyces Cerevisiae, obtained from Uniprot Database [9] (22969 interactions), b) the protein interaction network for Homo Sapiens, created from the interactions reported from iRefIndex database (115404 interactions) and, c) the weighted protein interaction network for H. Sapiens obtained from HINT-KB database (211367 interactions).

The input genetic interaction networks for S. Cerevisiae and H. Sapiens were created from BioGRID 3.2.98 database [10] and comprised 1606 genetic interactions for H. Sapiens and 145265 genetic interactions for S. Cerevisiae.

2.2 Methods

The overall methodology is consisted of a two step procedure. First, for each specific organism PPIs are predicted with computational prediction methods or retrieved through public available databases. Then, functional, structural and sequential information about the proteins are combined to predict a confidence score for these interactions. Finally, the constructed weighted PPI graphs are used as input for the biological process network prediction method to predict interactions between biological processes. All these methodologies are described in sections 2.2.1 -2.2.2.

2.2.1 EVOKALMA Model

The main idea of the proposed classification methodology, called (Evolutionary Kalman Mathematical Modelling - EvoKalMaModel) is to find a simple mathematical equation that governs the best classifier and enables the extraction of biological

knowledge. This method was based on previous methodologies proposed by our authoring group [11, 12] and it was initially published in [5]. It combines a state-of-the-art adaptive filtering technique named Kalman Filtering [13] with an adaptive genetic algorithm. The adaptive genetic algorithm is used to detect the optimal subset of terms in order to build the mathematical model for our predictor and then Extended Kalman Filters to compute its optimal parameters. The final model is in the form of a mathematical equation including a subset of the available mathematical terms and inputs. The evolutionary process is guided with a problem specific fitness function and utilizes an adaptive mutation rate to improve its convergence behavior.

2.2.2 Proposed Biological Process Network Method

Protein-protein interaction (PPI) networks are defined as graphs whose nodes represent proteins and the edges interactions between nodes/proteins. The weights of the edges (if they exist) represent the confidence scores of these interactions. Genetic interaction (GI) network is a graph whose nodes represent genes and the edges represent interactions between nodes/genes.

The definition of linked biological processes is based on an input undirected graph whose nodes represent proteins and edges represent an interaction. Two types of networks were used: physical level PPI networks and functional level GI networks. Only proteins annotated by at least one function (Gene Ontology function [14]) were considered. A link between process i and process j suggests that the proteins annotated by i tend to interact with proteins annotated by j more than expected by chance. The statistical enrichment of process j is calculated based on the set of proteins connected with proteins annotated by i, except those annotated by both i and j.

The formal definition is as follows:

Suppose N_i be the set of nodes annotated with process i (and possibly with other processes as well) and NB_i be the set of nodes not annotated with process i and connected with at least one node annotated with process i then:

$$\mathrm{NBi} = \{\mathrm{v}: < u, v > \in E, v \notin Ni, u \in Ni\} \quad (1)$$

Based on the above definitions, a process i is connected with a process j, if j is statistically enriched in NB_i set, when $\mathrm{P(i,j)} < threshold$ where:

$$\mathrm{P(i,j)} = \sum_{x=|\mathrm{NBi}\cap\mathrm{Nj}|}^{\min\{|\mathrm{Nj}|,|\mathrm{NBi}|\}} \frac{\binom{|\mathrm{Nj}|}{x}\binom{\mathrm{N}-|\mathrm{Nj}|}{|\mathrm{NBi}|-x}}{\binom{\mathrm{N}}{|\mathrm{NBi}|}} \quad (2)$$

For the aforementioned threshold a very strict value (0.001) was assigned (as proposed in [2]) in order to filter false positive results. N is defined as the sum of all proteins annotated with processes i and j, so P is the probability a link to connect a protein annotated with process i and a protein annotated with process j, among all possible links between proteins annotated with i or j.

The methodology so far, is similar to the methodology initially proposed in by [2]. The drawback of this approach is the assumption that interactions have a dual nature. However, recent approaches for constructing PPI graphs achieved to predict a confidence score for each interaction. Thus, in our approach for the quantity $|NB_i|$ in equation 2, we utilized the sum of the weights (confidence scores) that connects the nodes

not annotated with process i and the nodes annotated with processes i. Thus, connections of low confidence score will not be significant for the algorithm's output, while connections of high confidence score would be extremely significant.

To calculate semantic similarity of two biological processes the following formula was used:

$$\text{similarity}(t_i, t_j) = \frac{2*\log P(t_{\text{ancestor}})}{\log P(t_i)+\log P(t_j)} \quad (3)$$

where $P(t_i)$ is the probability of a gene to be annotated with t_i, that is the number of genes annotated with t_i divided by the total number of genes and $t_{ancestor}$ is the most specific common ancestor of t_i and t_j in Gene Ontology.

The pseudocode of the method is provided in Table 1:

Table 1. Pseudocode of BPN methodology

```
FOR processes i, j
    Find proteins annotated by these processes
    Exclude common proteins
    Find interactions among the remaining proteins
    Calculate P(i,j)
    IFP(i,j)<0.001 create link between i and j
REPEAT FOR EVERY i, j
```

3 Results and Discussion

For the problem of predicting and scoring PPIs many methods were applied and comparative results are provided in Table 2. The methodologies used for comparative reasons with the proposed EVOKALMA model, include the Naive Bayesian Classifier which is the algorithm utilized by most PPI databases that include computationally predicted PPIs [15, 16]. Moreover, two methodologies which have already provided encouraging results in predicting PPIs were used (Random Forests [17] and jGEPModel2.0 [18]) alongside with the hybrid combinations of Genetic Algorithms [19], Particle Swarm Optimization [20] and Differential Evolution [21] with Support Vector Machines (SVM) which have several applications in many fields.

Classifiers' performance is theoretically supposed to be more confident when more testing datasets are pooled. Therefore, our experimental setup was extended from one testing dataset to multiple testing datasets and performed double loop cross validation [22]. During the training 10-fold external and 9-fold internal cross validation were used. In particular, the external folds alter the subset of data used for testing the trained models (10 non-overlapping different subsets) after internal algorithms iterations (either heuristic iterations either parameter optimization steps), whereas the internal folds vary the subset of data used for validating the under training models (9 non-overlapping different subsets) during the algorithms internal iterations. The

optimal model of each implementation of the 9 internal folds was finally kept for every algorithm. Then, the average of the metrics in the external fold was calculated for the 10 optimal models for every algorithm and Table 2 presents these results.

Table 2. Comparative Results of PPI classification methods

ALGORITHM	ACCURACY	GEOMETRIC MEAN	SENSITIVITY	SPECIFICITY
Naive Bayesian Classifier	73.59%	74.84%	63.64%	88.00%
Random Forests	81.83%	81.82%	81.45%	82.20%
GA-SVM	79.19%	79.04%	74.34%	84.04%
PSO-SVM	81.64%	81.99%	82.08%	81.20%
DE-SVM	81.84%	81.84%	82.28%	81.40%
jGEPModel 2.0	82.67%	82.66%	83.28%	82.06%
EVOKALMA	**87.43%**	**87.40%**	**85.33%**	**89.51%**

From Table 2 it is clearly observed that EVOKALMA model outperformed significantly all the other deployed methodologies in all the utilized metrics. This strengthens our belief that EVOKALMA algorithm is the most appropriate solution for predicting and scoring PPIs.

The main concepts of the methodology are Interacting Processes and the method's output, the Biological Processes Network (BPN) which nodes are biological processes terms, as described by Gene Ontology database. This method, combined with experimental or computational data of genetic or protein interactions, led to the discovery of a variety of process interactions. Two types of networks were created: a) networks based on protein interactions and, b) networks based on genetic interactions. Some of the discovered connections are consistent with our biological knowledge, while others require further research. For example, the process "protein ubiquitination" (GO:0016567) was found to be PPI-linked with protein catabolism-related proteins and GI-linked with processes related with cell cycle.

A process a was defined as PPI-linked with a process b if the number of proteins which are annotated with b and interact with those annotated with a is greater than expected by chance. Based on the process interactions the BPNs were created. The network is directed, its nodes represent processes and a directed edge from process a to b dictates interaction from a to b. The network created for S. Cerevisiae contains 5285 edges and the network created for H. Sapiens contains 24758 edges.

Gene Ontology database uses various metrics for the semantic similarity of processes which are calculated by the distance of the processes on the hierarchical tree of Gene Ontology. These metrics provide values mostly between 0 and 1, where 0 means that the only common ancestor of two processes is the term GO:0008150 ("biological process"). Many of the PPI-linked processes share a high semantic similarity. While some others have low semantic similarity, it could be useful to be further studied.

As an example, the method discovered a link between “response to DNA damage stimulus” (GO:0006974) and “chromatin modification” (GO:0016568). Despite the fact that they have semantic similarity 0, chromatin and histone modification are utilized in the DNA damage response pathway [23].

Using the genetic interaction datasets, two more networks were created. Two genes are considered as interacting when their mutations show a combined effect which does not appear by either mutation alone. The network created for S. Cerevisiae contains 32967 edges and the network created for H. Sapiens contains 245 edges.

The resulting linked processes for the two organisms show great overlap, which is consistent with the fact that they share a large number of protein and genetic interactions. Unique linked processes also appear, such as the link between “hyperosmotic response” (GO:0006972) and “regulation of MAP kinase activity” (GO:0043405), which exists only in the S. Cerevisiae's network. The use of MAP kinase activity for the regulation of the hyperosmotic shock response, is a known mechanism of S. Cerevisiae.

Many of the proteins that genetically interact, also participated in PPI-linked processes. For S. Cerevisiae, the probability two proteins to participate in PPI-linked processes is 11% (only those processes with at least 50 genes participating were considered). If it is considered as known that these proteins interact genetically, the probability reaches 36%. Therefore, PPI-linked processes could be used by genetic interaction prediction algorithms to increase their performance. Also, two proteins are more likely to interact if they both belong to PPI-linked processes.

4 Conclusions and Future Work

In the present paper, we proposed a new holistic algorithmic framework to construct accurate biological process network. This framework is consisted of a PPI prediction and scoring algorithm which outperforms other existing methodologies and a variation of an existing algorithms for prediction biological process interactions which enables it to handle weighted PPI networks.

Utilizing the proposed methodology, various interactions between biological processes were predicted for both examined organisms, S. Cerevisiae and H. Sapiens. Some were consistent with previous knowledge, while others require further research. The method presented, focuses on the computational detection of links between processes, rather than their confirmation or biological interpretation. A challenge derived by this work is the need to review the structure of biological information, since many of the computed interactions show little semantic similarity based on the current structure of Gene Ontology database.

It was proven that PPI-linked processes could be used to enhance the performance of genetic interaction prediction algorithms. Identifying GI-linked processes cal lead to a new direction in generic interaction research. As a next step, the method could incorporate gene expression information, to study processes interactions at a treatment-control basis, as proposed in [3, 4].

Another interesting idea for future work, is to investigate which processes show different behavior in various conditions and how the interactions between them may change. A way to achieve that, is by incorporating gene expression information from different experimental conditions, such as infected and healthy tissue samples. Using microarray experiments, genes whose expression changes greatly (often determined by using simple statistical tests) can be defined as perturbed. A process is considered as perturbed when the genes annotated with this process are perturbed. A link between processes is perturbed when both incident processes are perturbed. A major drawback of this analysis is the report of all possible links between processes, which includes unperturbed links and possibly identical links. To overcome this constraint the utilized process links could be minimized by rewarding the inclusion of links representing many gene interactions.

References

1. Theofilatos, K., Dimitrakopoulos, C., Tsakalidis, A., Likothanassis, S., Papadimitriou, S., Mavroudi, S.: Computational Approaches for the Prediction of Protein-Protein Interactions: A Survey. Current Bioinformatics 6(4), 398–414 (2011)
2. Dotan-Cohen, D., Letovsky, S., Melkman, A.A., Kasif, S.: Biological Process Linkage-Networks. PLoS ONE 4(4), e5313 (2009), doi:10.1371/journal.pone.0005313
3. Lasher, C.D., Rajagopalan, P., Murali, T.M.: Discovering networks of perturbed biological processes in hepatocyte cultures. PLoS ONE 6(1), e15247 (2011)
4. Lasher, C., Rajagopalan, P., Murali, T.M.: Summarizing cellular responses as biological process networks. BMC Systems Biology (2013), http://dx.doi.org/10.1186/1752-0509-7-68
5. Theofilatos, K., Dimitrakopoulos, C., Likothanassis, S., Kleftogiannis, D., Moschopoulos, C., Alexakos, C., Papadimitriou, S., Mavroudi, S.: The Human Interactome Knowledge Base (HINT-KB): An integrative Human protein interaction database enriched with predicted protein protein interaction scores using a novel hybrid technique (Evolutionary Kalman Mathematical Modelling - EvoKalMaModel). Artificial Intelligence Review, 1–17 (2013), , doi: 10.1007/s10462-013-9409-8
6. Keshava Prasad, T.S., Goel, R., Kandasamy, K., et al.: Human Protein Reference Database-2009 update. Nucleic Acids Res. 37, D767–D772 (2009)
7. Razick, S., Magklaras, G., Donaldson, I.M.: iRefIndex: A consolidated protein interaction database with provenance. BMC Bioinformatics 9(1), 405 (2008)
8. Troyanskaya, O., Cantor, M., Sherlock, G., et al.: Missing value estimation methods for DNA microarrays. Bioinformatics 17(6), 520–525 (2001)
9. The UniProt Consortium: Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic Acids Res. 40, D71-D75 (2012)
10. Stark, C., Breitkreutz, B., Reguly, T., et al.: BioGRID: a general repository for interaction datasets. Nucleic Acids Res. 34, D535-D539 (2006)
11. Theofilatos, K.A., Dimitrakopoulos, C.M., Tsakalidis, A.K., et al.: A new hybrid method for predicting protein interactions using Genetic Algorithms and Extended Kalman Filters. In: Proceedings of the IEEE/EMBS Region 8 International Conference on Information Technology Applications in Biomedicine (ITAB). art. no. 5687765 (2010), doi : 10.1109/ITAB.2010.5687765

12. Dimitrakopoulos, C.M., Theofilatos, K.A., Georgopoulos, E.F., et al.: Efficient Computational Construction of Weighted Protein-Protein Interaction Networks Using Adaptive Filtering Techniques Combined with Natural-Selection Based Heuristic Algorithms. International Journal of Systems Biology and Biomedical Technologies (IJSBBT) 1(2), 20–34 (2011)
13. Welch, G., Bishop, G.: An Introduction to the Kalman Filter. University of North Carolina at Chapel Hill (1995)
14. Ashburner, M., Ball, C.A., Blake, J.A., et al.: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. 25, 25–29 (2000)
15. Scott, M., Barton, G.: Probabilistic prediction and ranking of human protein-protein interactions. BMC Bioinformatics 8, 239 (2007)
16. Zhang, Q., Petrey, D., Garzon, J., et al.: PrePPI: a structure-informed database of protein-protein interactions. Nucl. Acids Res (2012), doi:10.1093/nar/gks1231
17. Liu, Y., Kim, I., Zhao, H.: Protein interaction predictions from diverse sources. Drug Discov. Today 13, 409–416 (2008)
18. Theofilatos, K., Dimitrakopoulos, C., Antoniou, M., Georgopoulos, E., Papadimitriou, S., Likothanassis, S., Mavroudi, S.: Efficient Computational Prediction and Scoring of Human Protein-Protein Interactions Using a Novel Gene Expression Programming Methodology. In: Jayne, C., Yue, S., Iliadis, L. (eds.) EANN 2012. CCIS, vol. 311, pp. 472–481. Springer, Heidelberg (2012)
19. Holland, J.: Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence. MIT Press, Cambridge (1995)
20. Kennedy, J., Eberhart, R.: Particle swarm optimization. In: Proceedings of IEEE International Conference on Neural Networks, Piscataway, NJ, pp. 1942–1948 (1995)
21. Storn, R., Price, K.: Differential evolution - a simple and efficient heuristic for global optimization over continuous spaces. Journal of Global Optimization 11, 341–359 (1997)
22. Veenman, C.J., Tax, D.M.: LESS: a model-based classifier for sparse subspaces. IEEE Trans. Pattern Anal. Mach. Intell. 27(9), 1496–1500 (2005)
23. Unal, E., Arbel-Eden, A., Sattler, U., Shroff, R., et al.: DNA damage response pathway uses histone modification to assemble a double-strand break-specific cohesin domain. Mol. Cell. 16, 991–1002 (2003)