33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

# Analysis of incomplete gene expression dataset through protein-protein interaction information

Raimon Massanet-Vila [1,2,3], Teresa Padró [4,5], Anna Cardús [4,5],
Lina Badimón [4,5], Pere Caminal [1,2,3] and Alexandre Perera [1,2,3]

*Abstract*— **This paper shows a graph based method to analyze proteomic expression data. The method allows the prediction of the expression of genes not measured by the gene expression technology based on the local connectivity properties of the measured differentially expressed gene set. The prediction of the expression jointly with the stability of this prediction as a function of the variation of the initial expressed set is computed. The method is able to correctly predict one third of the proteins with independence of variations on the selection of the initial set. The algorithm is validated through a Matrix-Assisted Laser Desorption/Ionization Time of Flight Mass Spectrometer (MALDI-TOF) protein expression experiment aiming the study of the protein expression patterns and post-translational modifications in human endothelial vascular cells exposed to atherosclerotic levels of Low Density Lipoproteins (LDL).**

## I. INTRODUCTION

Most of the cellular processes and regulatory pathways of the cell are controlled by networks of interacting proteins. These networks determine how cells grow, divide, die, differentiate and communicate with other cells. Thus, failure in the docking of a pair of proteins, due to mutations in the generating gene or to post-translational modifications, can lead to malfunction of the corresponding process, impacting the pathway functionality and ultimately leading to disorder or disease.

The development of high-throughput technologies to discover new protein interactions has led to the need for creating and maintaining large protein-protein interaction (PPI) databases. PPI information can be structured as a network of interacting proteins. Such a network has a graph structure in which nodes correspond to proteins and edges correspond to interacting proteins. Different groups have contributed to the creation of a large number of databases aiming to settle a standard on the information content of each protein interaction, including interaction evidence, source, type of interaction, cross-references to ontologies and others (see [1] for a review of the most important).

The development of protein-protein interaction (PPI) databases and their increasing level of annotation have allowed the massive computational analysis of large interactions networks ([2]), which has also been focused on visualization methods and interactive query tools such as Osprey ([3]). Commercial applications for the analysis of PPI networks have also been promoted such as Ingenuity Pathway Analysis (IPA, from Ingenuity Systems, Redwood City, California, US). IPA and Osprey are usually regarded as gold standards with which new methods compare.

In general, the application of the computational analyses of these graphs is wide, from meta-analysis in genome wide association studies, selection of candidate genes in biochemistry and gene expression analysis. Many authors have contributed to the application of graph theory concepts to PPI networks in order to find topological differences between disease genes and non-disease genes, and generally, to the prediction genes related to phenotype ([4]). Some authors claim that they were able to accurately predict disease genes using features exclusively derived from PPI network topology [5].

Some applications have been developed for the visualization of combined PPI information and protein expression data [2]. Certainly, there is evidence that protein-protein interactions are related to mRNA coexpression [6], [7], [8]. Proteins that interact have expression profiles with a higher degree of correlation that what would be expected by chance. Interestingly, this correlation is not always positive; it can be negative, suggesting inhibition, cleavage or proteolysis. These results indicate that proteins that interact, or are in the same local interaction environment, tend to have correlated expression profiles. Furthermore, some methods that combine gene expression data and PPI information for candidate gene prioritization have been published [9]. The authors of these combined methods claim to obtain better results than using only PPI-based methods or expression-based methods.

Gene expression is commonly studied through DNA microarrays. This technology allows for the investigation of the transcriptional activity of several thousands of genes, typically measuring the amount of mRNA expressed by each coding sequence. However, there may exist a difference between the transcriptional profiles measured from the mRNA expression profiles and the actual protein level in cells. High-resolution two-dimensional gel electrophoresis (2DE) combined with sensitive state of the art mass spectrometry technology provide the possibility of identifying the proteomic

This work was supported by the Spanish Ministerio de Educación y Ciencia under the Ramón y Cajal Program and TEC2010-20886-C02-02, CIBER-BBN and CIBER-OBN.

[1] Department of ESAII. Technical University of Catalonia (UPC). Address: Pau Gargallo 5, 08028 Barcelona. Spain: raimon.massanet@upc.edu, pere.caminal@upc.edu, alexandre.perera@upc.edu

[2] Biomedical Engineering Research Center (CREB). Barcelona. Spain

[3] CIBER-BBN in Bioengineering, Biomedicine and Nanomaterials. Spain.

[4] Cardiovascular Research Center (CSIC-ICCC), IIB Sant Pau, Barcelona, Spain, acardus@csic-iccc.org, tpadro@csic-iccc.org, lbadimon@csic-iccc.org

[5] CIBER-OBN in Pathophysiology of Obesity and Nutrition. Spain

978-1-4244-4122-8/11/$26.00 ©2011 IEEE

pattern of the cells in a specific pathophysiologic status. 2DE methods separate proteins based on their isoelectric point in the first dimension and their mass in the second dimension.

2DE gels show some technology specific issues that open new opportunities for studying signal processing. Analysis of these expression images must include image warping and align correction, image fusion, spot detection, computation of consensus spot patterns and extraction of expression profiles jointly with spot identification [10]. The process is prone to some issues like missing values or undetected spots, weak spots, overlapping spots, resulting in some proteins that are in fact non-detectable by the technology [11]. In general, all protein expression technologies measure only a fraction of the complete set of proteins present.

This paper proposes the use of protein interaction networks for the enrichment of the information obtained by protein expression techniques such as 2D-Gel electrophoresis. The method mines protein relationships to guess non-measured proteins that could be differentially expressed on the process under study, based on the analysis of the local protein interaction network that shows most relationship with the differentially expressed subset.

## II. MATERIALS AND METHODS

The proposed method is based on a post-processing of the results of a protein expression experiment through the enrichment of the expression profiles with protein interaction information.

First, a set of *candidate proteins* is defined, which correspond to the set of proteins that has shown significant differential expression in the protein expression experiment. The methodology proposes a set of *new candidate proteins*, which contain proteins that were not identified in the protein expression experiment but the method considers to be relevant.

The algorithm starts by creating the network of proteins that connects all differentially expressed or candidate proteins. In that network, all shortest paths between pairs of candidate proteins are computed. These paths correspond to the minimal proteomic pathways between candidate proteins. Finally, a network that contains only this minimal pathways is created. See Fig. 1 for a graphical general description of the method. Because of the way the final network is constructed it contains proteins with a high probability of having a protein expression profile that is highly correlated with the set of candidate proteins. Each of these steps is explained in detail below.

First, a network that connects all candidate proteins is constructed. This network is built by iteratively adding interacting proteins to an initially disconnected network until it is connected. At iteration $I_0$, an initial network is created by adding the candidate proteins $P_0$. At each iteration $I_i$, all proteins that show known interaction with the proteins added in iteration $I_{i-1}$ are added to the network. This step continues until either the network is connected or a maximum number of iterations is reached. The maximum number of iterations defines our maximum *distance threshold*. This distance is a parameter of the algorithm that captures an observation scale through the proteome.

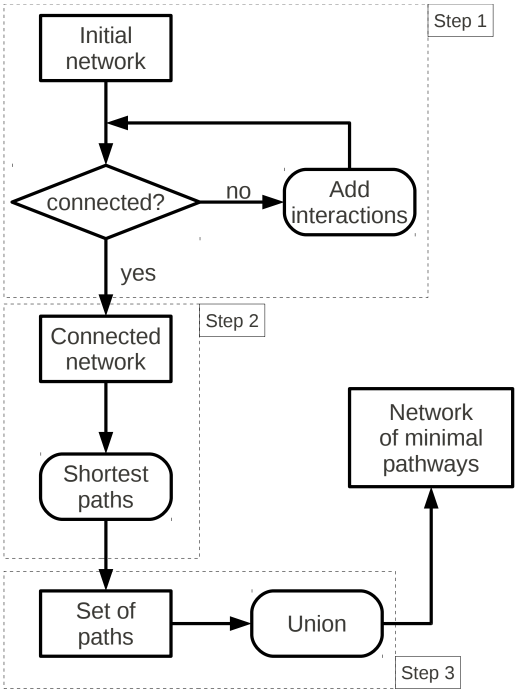

Fig. 1. Block diagram of proposed methodology.

The second step consists in calculating all shortest paths between all pairs of candidate proteins. This is done by using a modified version of Dijkstra's algorithm that finds all shortest paths between two nodes [12]. This is accomplished by keeping a predecessor tree structure, instead of a predecessor vector. In Dijkstra's algorithm a relaxation occurs when the current path is shorter than a previously calculated shortest path for node $v$. Then, the predecessor of $v$ is updated to be in accordance with the new shortest path. In the modified version, in addition to the normal relaxation, a new kind of relaxation occurs when the current path has an equal length than the previously computed shortest path. In this case a new predecessor is added as a child node in the predecessor structure for node $v$. Finally, minimal paths from node $u$ to node $v$ are reconstructed by navigating the predecessor tree. It is important to note that the set of minimal paths from $u$ to $v$ form a graph. These paths are the minimal proteomic pathways that connect candidate proteins.

In the third step the final network is generated by calculating the union of all minimal pathways. This graph consists of the union of all the graphs of minimal pathways between all pairs of candidate nodes. Since this network contains only minimal proteomic pathways that connect candidate proteins, it is in accordance with the literature that proteins in this network have a high probability of having expression profiles correlated with the expression of candidate genes ([6], [7], [8]). Thus, the method enriches the initial set of candidate proteins with other proteins that could have correlated expressions profiles and were not identified in the protein expression experiment. We call this network *minimal* *pathways network* (MPN).

### A. Validation

The quantitative validation of the method proposed was performed from two points of view. First, a *discovery rate* was calculated and used to assess the capability of the method to discover candidate proteins that were purposely removed from the initial set of candidates prior to the application of the method. Second, a *robustness index* was computed to assess the stability of the result under a change on the initial set of candidates.

A three fold cross-validation was performed to test the proposed method from two different points of view. The set of candidate proteins was randomly divided in a training set $S_t$ comprising two thirds of the candidates and a validation set $S_v$ containing the other third of the candidates. The training set $S_t$ was used as input for obtaining a set of new candidates: $S_c^{MPN}$ .

On the validation set, a *discovery ratio* was computed for each tested method $m$ as:

$$d_m = \frac{n_m}{|S_v|} \tag{1}$$

where $n_m$ is the number of proteins of $S_v$ that were found by the method $m$. This procedure was repeated 50 times and the samples of discovery ratios for the proposed method were compared using a Mann-Whitney test.

To assess the *robustness index* of the tested methods, 50 pairs of cross-validation runs were compared and a robustness index was calculated for each method $m$ as:

$$r_{ij} = \frac{\gamma(S_{t_i}, S_{t_j})}{\gamma(S^m_{c_i}, S^m_{c_j})} \tag{2}$$

where $\gamma$ is a function that measures the amount of change between the two sets.

## III. DATA

### A. PPI data

Protein-protein interaction data was obtained from the Human Protein Reference Database (HPRD) [13]. This database was downloaded, with permission of the owners, on May-03 2010. The data comprised 38 756 interactions among 9 630 proteins. A number of custom methods were built to transform this data to graph structure and to operate on graphs through *the R Language for Statistical Computing* [14].

### B. Protein expression data

The protein expression data were obtained from an *in vitro* cell culture study aimed to investigate the effect of high concentrations of human Low Density Lipoproteins (LDL) on the proteomic expression pattern of human endothelial cells. After cell extraction, proteins were separated by 2D gel electrophoresis, proteomic patterns were analyzed using a devoted software (PDQuest, BioRad) and proteins were identified with a Matrix-Assisted Laser Desorption/Ionization Time of Flight Mass Spectrometer (MALDI-TOF, GE-Healthcare).

A total of 85 proteins were identified in the first stage. A significant differential expression threshold was defined as a case/control ratio over 3/2 or below 2/3. Fig. 2 shows the distribution of the protein expression ratios. 39 proteins showed significant differential expression. Of those, 35 could be mapped to the HPRD database. This set of 35 were defined as *candidate proteins* used as input for the proposed methodology.

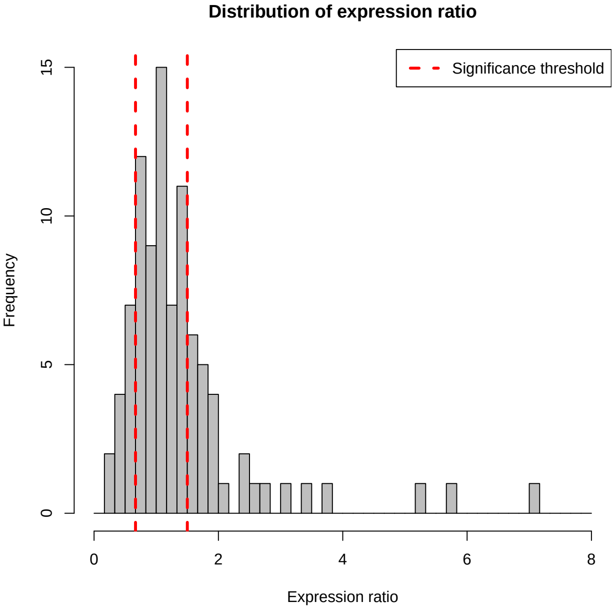

Fig. 2. Distribution of ratio case/control in the differential protein expression experiment data.

## IV. RESULTS

The network formed by the union of all minimal proteomic pathways between candidate proteins is shown in figure 4.

Considering all trials on the cross-validation, the mean discovery rate for the MPN method is 0.32 for the validation set, meaning that the method could find, in average, about one third of the true candidate proteins that were extracted from the initial set of candidates. Fig. 3 shows the distribution of the discovery rate for method MPN. To assess the significance of this distribution, it was compared against a null model consisting of paired random samples of proteins in the HPRD interactome through a paired Wilcoxon test. The test showed that the discovery rate of MPN is well above that expected by chance $pvalue = 2.98 \cdot 10^{-9}$.

The stability measured as (2) for the results of the MPN method is higher than the stability of the input candidate sets, in a Wilcoxon paired test ($pvalue < 2.2 \cdot 10^{-16}$). This means that the changes observed in the results yielded by the MPN method are actually less important than the changes induced in the input candidate set by the cross-validation process.

Both results demonstrate that PPI databases contain a sufficient number of annotations in order to predict missing protein expression from the 2D-Gel based identification

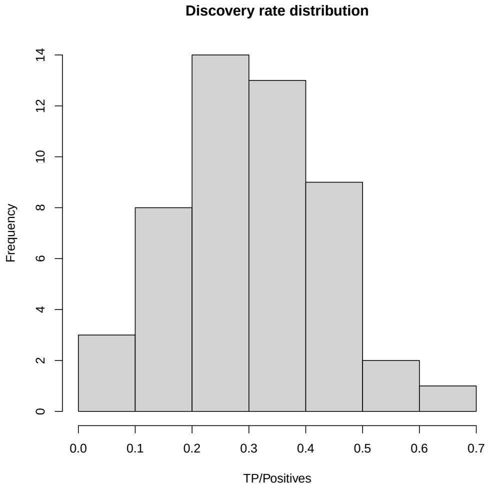

Fig. 3. Distribution of discovery rate in 50 random samples. Discovery rate is calculated as the amount of true positives found by the method divided by the total amount of positives.

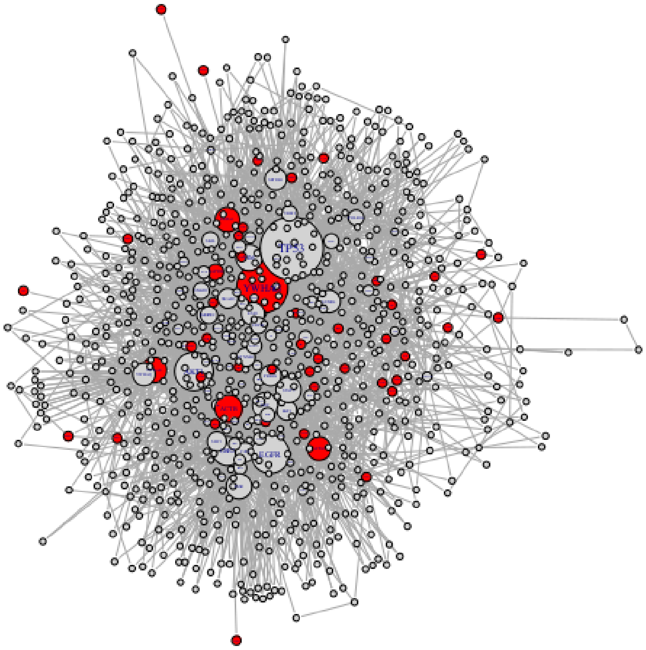

Fig. 4. The protein interaction network formed by the union of all shortest paths between differentially expressed proteins. The size of the nodes is proportional to their degree. The initial candidate proteins are marked in red.

process. The stability test showed that the method's results are sufficiently stable unregarding the initial candidate set. Although results have been computed for 2D-Gel MALDI-TOF experiments, the methodology can be applied to any gene expression methodology, including transcriptome analyses through DNA-Arrays and exome data in genome-wide studies.

## V. CONCLUSIONS AND FUTURE WORKS

This paper proposes a novel method for processing data from gene expression data for the prediction of missing or non-measured differential expressed proteins. The method has been validated through the 2D-Gel response of a cell line under LDL exposure. The method has been able predict over one third of the missing proteins that were differentially expressed in validation set and showed stability on the predicted differentially expressed set. This figures show a preliminary validation of the method. Future work will compare systematically this approach against similar methodologies found on some commercial packages like IPA.

## VI. ACKNOWLEDGEMENTS

CIBER-BBN in Bioengineering, Biomaterials and Nanomedicine is an initiative by ISCIII.
CIBER-OBN in Pathophysiology of Obesity and Nutrition.

## REFERENCES

[1] S. Mathivanan *et al.*, "An evaluation of human protein-protein interaction data in the public domain," *BMC Bioinformatics*, vol. 7, p. S19, 2006.

[2] J. P. Gonalves *et al.*, "Polar mapper: a computational tool for integrated visualization of protein interaction networks and mrna expression data," *Journal of The Royal Society Interface*, vol. 6, no. 39, pp. 881–896, October 06 2009.

[3] B. Breitkreutz *et al.*, "Osprey: a network visualization system," *Genome biology*, vol. 4, no. 3, 2003. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/12620107

[4] T. Ideker and R. Sharan, "Protein networks in disease," *Genome research*, vol. 18, no. 4, pp. 644–652, April 01 2008.

[5] J. Xu and Y. Li, "Discovering disease-genes by topological features in human proteinprotein interaction network," *Bioinformatics*, vol. 22, no. 22, pp. 2800–2805, November 15 2006.

[6] R. Jansen *et al.*, "Relating whole-genome expression data with protein-protein interactions," *Genome research*, vol. 12, no. 1, pp. 37–46, January 01 2002.

[7] N. Bhardwaj and H. Lu, "Correlation between gene expression profiles and proteinprotein interactions within and across genomes," *Bioinformatics*, vol. 21, no. 11, pp. 2730–2738, 2005.

[8] S. Tornow and H. W. Mewes, "Functional modules by relating protein interaction networks and gene expression," *Nucleic acids research*, vol. 31, no. 21, pp. 6283–6289, November 01 2003.

[9] X. Ma *et al.*, "Cgi: a new approach for prioritizing genes by combining gene expression and proteinprotein interaction data," *Bioinformatics*, vol. 23, no. 2, pp. 215–221, January 15 2007.

[10] M. Berth *et al.*, "The state of the art in the analysis of two-dimensional gel electrophoresis images," *Applied Microbiology and Biotechnology*, vol. 76, no. 6, pp. 1223–1243, 2007.

[11] R. Pedreschi *et al.*, "Treatment of missing values for multivariate statistical analysis of gel-based proteomics data," *Proteomics*, vol. 8, no. 7, pp. 1371–1383, 2008.

[12] E. W. Dijkstra, "A note on two problems in connexion with graphs," *Numerische Mathematik*, vol. 1, no. 1, pp. 269–271; 271, -12-01/ 1959. [Online]. Available: http://dx.doi.org/10.1007/BF01386390

[13] T. S. K. Prasad *et al.*, "Human protein reference database–2009 update," *Nucleic acids research*, vol. 37, no. suppl_1, pp. D767–772, January 1 2009.

[14] R Development Core Team, "R: A language and environment for statistical computing," 2009. [Online]. Available: http://www.R-project.org