34th Annual International Conference of the IEEE EMBS
San Diego, California USA, 28 August - 1 September, 2012

Network-based Enrichment Analysis of Gene Expression through Protein-Protein Interaction Data

Raimon Massanet-Vila[1,2,3], Francesc Fernández Albert[1,2,3], Pere Caminal[1,2,3] and Alexandre Perera[1,2,3]

[1]Dept. of Sistems Engineering, Automatics and Industrial Informatics. Technical University of Catalonia (UPC).
Pau Gargallo 5, 08028, Barcelona, Spain. (http://www.upc.edu).
Email: {raimon.massanet, francesc.fernandez.albert@upc.edu, pere.caminal, alexandre.perera}@upc.edu

[2]Biomedical Engineering Research Center (CREB), Barcelona, Spain. (http://www.creb.upc.es).

[3]CIBER-BBN in Bioengineering, Biomaterials and Nanomedicine, Spain. (http://www.ciber-bbn.es).

***Abstract*— High-throughput analysis of gene expression data is subject to technological and statistical issues that confuse the underlying expression-condition associations. In this contribution a network-based candidate gene prioritization strategy was applied to the enrichment of a publicly available gene expression dataset, focused on the study of the mechanosensitivity of genes exposed to altered pulmonary matrix stiffness. Results suggested that some genes which had not been taken into account in the original study could have an important role in the processes causing, or affected by, pulmonary fibrosis.**

I. INTRODUCTION

Gene expression microarrays measure the expression of thousands of genes simultaneously, which allows conducting hypothesis-free genome-wide studies of gene expression changes under varying experimental conditions. However, this lack of hypothesis causes that statistical tests are performed on a number of variables that can be several orders of magnitude higher than the sample size, causing that the results have a very low statistical power and low reproducibility [1]. Enriching gene expression data with other sources of independent data is a commonly used approach for prioritizing candidate genes, in an attempt to select for further studies genes with different sources of positive evidence [2]. In particular, protein-protein interaction (PPI) data has been widely used for gene expression enrichment [3]. Thus, network analysis has become a major topic in genomic and proteomic studies.

PPI networks are represented by graphs in which nodes represent proteins and edges represent binary protein interactions. In the last decades a large number of PPI databases have emerged. Figures 1 and 2 illustrate the wealth in PPI data that are publicly available to the researchers.

In this contribution we demonstrate a candidate gene prioritization strategy based on network analysis of PPI-enriched gene expression data. This methodology was applied to the analysis of a publicly available gene expression dataset [6]. That work was focused on finding genes that showed differential expression under induced pulmonary fibrosis.

*This work was supported by the Spanish Ministerio de Educación y Ciencia under the Ramón y Cajal Program, TEC2010-20886-C02-01, TEC2010-20886-C02-02 and the CIBER-BBN. CIBER-BBN in Bioengineering, Biomaterials and Nanomedicine is an initiative by the Instituto de Salud Carlos III.

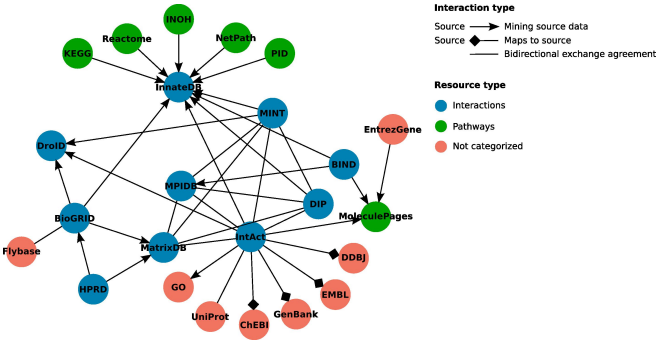

Fig. 1: Interaction databases and interchange of information among them. (Source: Klingström and Plewczynski 2011 [4]).

II. MATERIALS AND METHODS

The PPI data used was obtained from the Human Protein Reference Database (HPRD) in its version of 07/06, 2009 [7]. The data was downloaded and converted to network structure using the *igraph* R package [8].

The gene expression dataset used to demonstrate the methodology was obtained from the Gene Expression Omnibus (GEO), dataset GSE22011 by Liu *et al.* 2010 [6]. The focus of the work was to find mechanosensible genes that showed differential expression under different pulmonary tissue stiffness conditions. After performing a time-course-like gene expression analysis they found that the COX-2 protein, encoded by the PTGS2 gene, was a very significant gene that responded and contributed to changes in matrix stiffness.

In this work, a time-course-like gene expression analysis was performed, following the indications of Liu *et al.* 2010

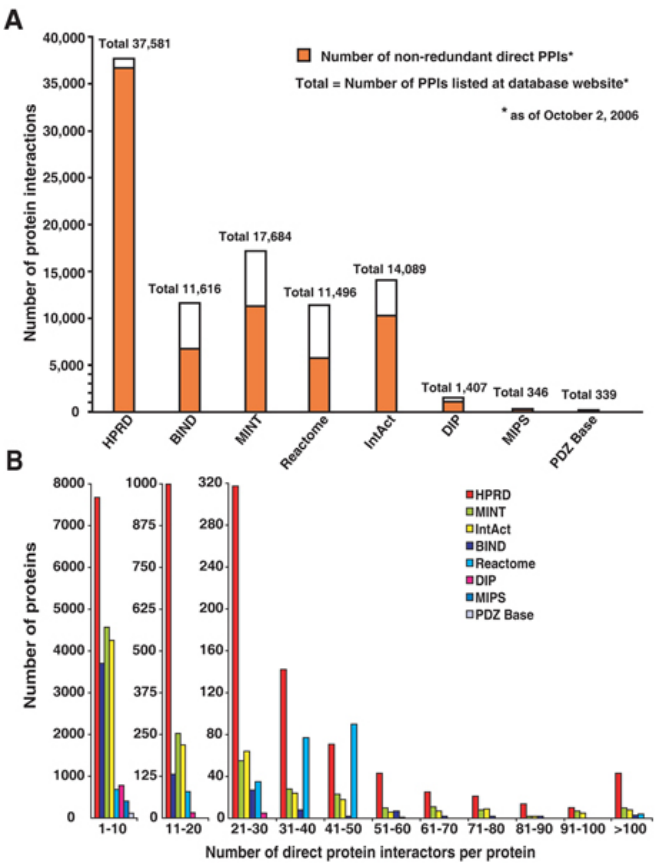

Fig. 2: A comparative analysis of different publicly available databases. (Source: Mathivanan *et. al* 2006 [5], reproduced with permission of the authors).

to obtain a set of candidate genes. Afterwards, a network-based analysis was performed in two steps:

1) **Construction of a local interaction environment.** This was done by calculating all shortest paths between all pairs of candidate genes. The interaction environment of the candidate genes was obtained by selecting only nodes and edges that belonged to any of the shortest paths.
2) **Prioritization of nodes according to different network topology measures.** Four different measures were computed for each node in the local interaction domain: degree, betweenness centrality, closeness centrality and clustering coefficient. The degree of a node (i.e. connectivity) is the number of edges that incide on that node. Betweenness centrality measures the centrality of a node regarding the number of times it is in the shortest path between any other two nodes of the graph. Closeness centrality measures the centrality of a node regarding the average distance of that node to any other node in the graph. The clustering coefficient measures to what extend the neighbors of a node have interactions among them, forming densely connected modules.

TABLE I: Gene connectivity

Gene symbol	Gene connectivity
TP53	21
ATXN1	12
PIK3R1	11
BMPR1B	11
CAV1	10
MATN2	**10**
S100A4	9
APP	9
BMP2	**9**
ACVR1	9
CDC42	9
TGFBR1	8
ABL1	8
SNAP25	8
TGFB1	8
IL7R	**8**
EP300	7
UBQLN4	7
HRAS	6
UNC119	6
BAT3	6
FYN	6
BRCA1	6
STAT3	6
PRKCA	5
TGFB2	5
CENPF	**5**

Top 25% of genes according to their connectivity. Genes belonging to the set of candidates genes are shown in bold.

III. RESULTS

Figure 3 shows the network corresponding to the local interaction environment of the set of candidate genes. It was formed by 108 genes and 221 interactions, and had a diameter of 7. Node connectivity ranged from 1 (ANLN) to 21 (TP53), and node betweenness centrality spanned from 0 (ANLN) to 1.33e3 (TP53).

Tables I and II show the 25% of genes with the highest values of connectivity and centrality, respectively. Surprisingly, PTGS2, the gene encoding for protein COX-2, is ranked 89 (82%) in the degree-based prioritization, 64 (59%) in the betweenness centrality-based prioritization, 25 (23%) in the closeness centrality-based prioritization and 69 (64%) in the clustering-based prioritization.

IV. DISCUSSION

The authors of the study from which the expression data was retrieved stated that the PTGS2 gene, that encodes for the cyclooxygenase 2 (COX-2) protein, was the most relevant result of their study. However, the basis on which COX-2 was chosen before other more significant candidates remains unclear. This fact illustrates the need for enrichment analysis, as, because of statistical limitations, the most relevant genes are not always the most significant. In that case the enrichment was probably performed based on a background knowledge which suggested that COX-2 was the candidate gene most likely to be related with the phenotype under study.

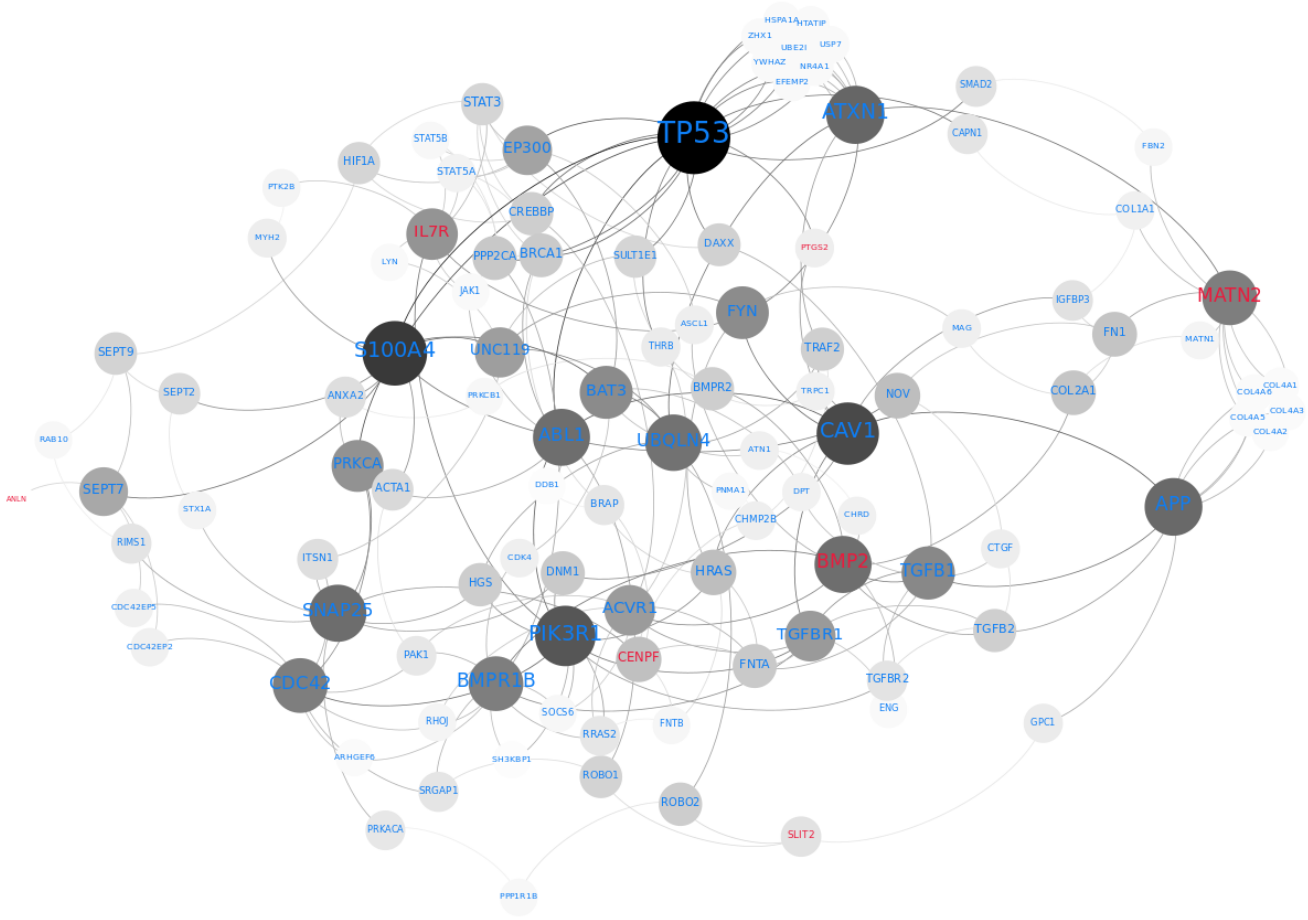

Fig. 3: The local interaction environment of the set of candidate genes. Candidate genes are labeled in red, whereas other nodes in the network are labeled in blue. Node and label size are proportional to gene connectivity. Node background color correlates with genes' betweenness centrality. It goes from white, for genes with low centrality, to black, for genes with high centrality.

Our automatic enrichment suggested a new candidate gene prioritization, based solely on topological properties of the resulting protein interaction network. Our results suggested that both TP53, CAV1 and PIK3R1 are very important genes in the biological processes related to the phenotype.

The TP53 gene encodes for the tumour protein 53, a vital protein that has a key role in regulating the cell cycle, preventing genome mutation and, thus, acting as a tumour suppressor protein. Defects in TP53 have been related to a variety of tumours and cancers, including lung cancer. In fact, more than 50% of human tumours contain mutations or deletions on TP53 [9].

CAV1 is the gene that encodes for caveolin-1. This protein is necessary for the formation of caveolae, which are regions of the cellular membrane that are responsible for a wide range of vital cellular processes like signal transduction, vesicular trafficking and tumour suppression [10]. The CAV1 gene is located in the chromosome 7, near a microsatellite that has been related to a variety of epithelial-based tumours, like breast cancer. CAV1 has a "natural" variant that is present in 16% of lung cancers. In addition, CAV1 seems to enhance cellular survival by the positive regulation of the PI3K/Akt signaling pathway, which could favour cell proliferation in abnormal cancer cells [11]. Expression of caveolin-1 seems to be directly correlated with cellular motility, which could explain that tumours with high expression of caveolin are more aggressive and present a higher rate of metastasis.

The PIK3R1 gene encodes for a regulatory protein that is involved in a wide variety of cellular processes and it has a central role in the PI3K/Akt signaling pathway, which is related to cellular proliferation, cellular motility and intracellular traffic, which are related to cancer. CAV1 positively regulates the PI3K/Akt signaling pathway, which could be the cause of the observed correlation between CAV1 and cell motility.

However, none of the genes described above was significant in the statistical test performed by Liu *et al.* 2010. Their good performance in the network topology-based prioritization could be a consequence of their importance in highly studied processes, which could introduce a bias in connectivity and, therefore, in topological measures [12]. The fact that all three genes have been related to cancer should

TABLE II: Gene betweenness centrality

Gene symbol	Betweenness centrality
TP53	1.33e+03
S100A4	7.16e+02
CAV1	6.19e+02
PIK3R1	5.47e+02
ATXN1	4.67e+02
APP	4.54e+02
SNAP25	4.35e+02
BMP2	**4.32e+02**
ABL1	4.31e+02
UBQLN4	4.16e+02
CDC42	3.66e+02
MATN2	**3.65e+02**
BMPR1B	3.64e+02
TGFB1	3.24e+02
BAT3	3.17e+02
FYN	3.11e+02
PRKCA	2.90e+02
IL7R	**2.81e+02**
TGFBR1	2.61e+02
ACVR1	2.58e+02
UNC119	2.48e+02
EP300	2.32e+02
SEPT7	2.16e+02
NOV	1.50e+02
HRAS	1.48e+02
CENPF	**1.45e+02**
FN1	1.34e+02

Top 25% of genes according to their betweenness centrality. Genes belonging to the set of candidates genes are shown in bold.

not be a surprise, since there seems to exist an increased risk of lung cancer for pulmonary fibrosis patients [13].

Two genes scored high in our network-based analysis, that had also scored high in the study by Liu *et al.* 2010: BMP2 and MATN2. The BMP2 gene encodes for the bone morphogenetic protein 2, which is a secreted protein found in lung tissues. It plays a fundamental role in bone and cartilage formation and is involved in the TGF-β signaling pathway, which is known to be altered by changes in matrix stiffness [6]. Furthermore, this protein seems to have an important role in the epithelial to mesenchymal transition (EMT), which could favour the metastatic behaviour of cancer cells. Defects in the receptor associated to BMP2 are the cause of several cancers, including breast and endometrial cancer. The protein encoded by the MATN2 gene is also a secreted protein that is hypothesized to have a role in the formation of the extracellular matrix filaments and could, therefore, be responsible of, or sensible to, changes in the matrix stiffness.

In conclusion, the proposed automatic, network-based prioritization strategy yielded potential candidates, that could have a role in the processes involving changes in matrix stiffness. The real causes of such changes still remain unknown, as unknown remain most of the mechanosensible genes that alter their expression because of the perceived change in extracellular stiffness. The reason why some idiopathic pulmonary fibrosis (IPF) patients develop lung cancer and other do not is probably closely related to the alteration of such mechanosensible genes. Elucidation of the cellular processes that lead to abnormal stiffness and/or discovery of the genes that change their expression under different stiffness conditions leading to lung cancer could be a major step towards the full understanding of the disease.

V. ACKNOWLEDGMENTS

This work was supported by the Spanish Ministerio de Educación y Ciencia under the Ramón y Cajal Program, TEC2010-20886-C02-01, TEC2010-20886-C02-02 and the CIBER-BBN. CIBER-BBN in Bioengineering, Biomaterials and Nanomedicine is an initiative by the Instituto de Salud Carlos III.

REFERENCES

[1] R. Nadon and J. Shoemaker, "Statistical issues with microarrays: processing and analysis," *Trends in Genetics*, vol. 18, no. 5, pp. 265–271, 5/1 2002.

[2] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov, "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 43, pp. 15 545–15 550, October 25 2005.

[3] L. Geistlinger, G. Csaba, R. Küffner, N. Mulder, and R. Zimmer, "From sets to graphs: towards a realistic enrichment analysis of transcriptomic systems," *Bioinformatics*, vol. 27, no. 13, pp. i366–i373, July 01 2011.

[4] T. Klingström and D. Plewczynski, "Protein-protein interaction and pathway databases, a graphical review," *Briefings in Bioinformatics*, vol. 12, no. 6, pp. 702–713, November 01 2011.

[5] S. Mathivanan, B. Periaswamy, T. Gandhi, K. Kandasamy, S. Suresh, R. Mohmood, Y. L. Ramachandra, and A. Pandey, "An evaluation of human protein-protein interaction data in the public domain," *BMC Bioinformatics*, vol. 7, p. S19, 2006.

[6] F. Liu, J. D. Mih, B. S. Shea, A. T. Kho, A. S. Sharif, A. M. Tager, and D. J. Tschumperlin, "Feedback amplification of fibrosis through matrix stiffening and cox-2 suppression," *The Journal of cell biology*, vol. 190, no. 4, pp. 693–706, August 23 2010.

[7] T. S. K. Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, L. Balakrishnan, A. Marimuthu, S. Banerjee, D. S. Somanathan, A. Sebastian, S. Rani, S. Ray, C. J. H. Kishore, S. Kanth, M. Ahmed, M. K. Kashyap, R. Mohmood, Y. L. Ramachandra, V. Krishna, B. A. Rahiman, S. Mohan, P. Ranganathan, S. Ramabadran, R. Chaerkady, and A. Pandey, "Human protein reference database–2009 update," *Nucleic acids research*, vol. 37, no. suppl_1, pp. D767–772, January 1 2009.

[8] G. Csardi and T. Nepusz, "The igraph software package for complex network research," *InterJournal. Complex Systems*, p. 1695, 2006. [Online]. Available: http://igraph.sf.net

[9] M. Hollstein, D. Sidransky, B. Vogelstein, and C. Harris, "p53 mutations in human cancers," *Science*, vol. 253, no. 5015, pp. 49–53, July 05 1991.

[10] T. Williams and M. Lisanti, "The caveolin proteins," *Genome biology*, vol. 5, no. 3, 2004.

[11] M. Shatz and M. Liscovitch, "Caveolin-1: A tumor-promoting role in human cancer," *Int J Radiat Biol*, vol. 84, no. 3, pp. 177–189, 01/01; 2012/03 2008.

[12] R. Massanet-Vila, P. Caminal, and A. Perera, "Graph theory-based measures as predictors of gene morbidity," in *32nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, August 31 - September 4, 2010 2010.

[13] J. Park, D. S. Kim, T. S. Shim, C.-M. Lim, Y. Koh, S. D. Lee, W. S. Kim, W. D. Kim, J. S. Lee, and K. S. Song, "Lung cancer in patients with idiopathic pulmonary fibrosis," *European Respiratory Journal*, vol. 17, no. 6, pp. 1216–1219, June 01 2001.