A Hybrid Methodology for Pattern Recognition in Signaling Cervical Cancer Pathways

David Escarcega[1], Fernando Ramos[1], Ana Espinosa[2], and Jaime Berumen[2]

[1] ITESM, Computer Science Department, Morelos, México
daescarcega@gmail.com, fernando.ramos@itesm.mx
[2] Hospital General de México, Unidad de Medicina Genómica, Ciudad de México, México
anaesga@hotmail.com, jaimeberumen@hotmail.com

Abstract. Cervical Cancer (CC) is the result of the infection of high risk Human Papilloma Viruses. mRNA microarray expression data provides biologists with evidences of cellular compensatory gene expression mechanisms in the CC progression. Pattern recognition of signalling pathways through expression data can reveal interesting insights for the understanding of CC. Consequently, gene expression data should be submitted to different pre-processing tasks. In this paper we propose a methodology based on the integration of expression data and signalling pathways as a needed phase for the pattern recognition within signaling CC pathways. Our results provide a top-down interpretation approach where biologists interact with the recognized patterns inside signalling pathways.

1 Introduction

Cervical Cancer (CC) is one of the most widespread cancers in women worldwide [1]. Cervical carcinogenesis is caused by an infection of high-risk Human Papilloma Viruses (hrHPV) [2]. After hrHPV infection and CC progression other transformation events occur within the cell, for instance, deregulation of genes expression levels and alteration of cellular processes either metabolic or signaling cascades [3].

Based on the integration of signaling pathways and high-throughput gene expression data, biologists seek to find modified or unchanged cellular processes related with cervical carcinogenesis or CC progression. Signaling pathways regulate the reception of external biochemical information, that will affect processes inside the cell, or intracellular interchange information; assemble of cascade events within the cell and finally, activate of cellular response to internal or external stimuli. Meanwhile, thousands of genes transcription levels can be measured using a single microarray [4], either to prove or propose novel hypothesis of complex diseases, as CC, by providing gene expression profiles. Gene expression profiles allow individual comparison of genes expression between populations or extrapolation of genes state [5]. Expression profiles integrated with signaling pathways eases the process for inferring the inner state of the cellular mechanisms by providing biologists with a big picture of expression compensation of

J.A. Carrasco-Ochoa et al. (Eds.): MCPR 2010, LNCS 6256, pp. 301–310, 2010.

genes, probably related with CC progression. Microarray data should be normalized before subsequent analysis could be accomplished. Normalization is the removal of technical noise, generated by experimental protocol, leaving expression profiles intact [6]. Once expression data is normalized, different workflows to infer internal cellular behavior could be followed.

Clustering of gene expression data is a common workflow to infer unknown genes function, find new disease subclasses, and primarily, data reduction and visualization. Clustering approaches group genes with similar expression level by measuring closeness in a quantitative way [7]. Clustering methods focus on quantitative data are considered 'unsupervised' methods [8], meaning that no gene functionality or previous phenotypic is considered for gene classification. Clustering approaches provide an overall picture of data variation. Classified expression profiles can be enriched with ontology data or cellular context, i.e. Gene Ontology [9]. An alternative method that has acquired an increased attention from genomic and computational scientists is to use pathway contexts to infer cellular processes alterations [10]. The pathway context provides biologists with a functional perspective, visualization of cellular processes and the impact of genes expression variations in such processes [11]. Furthermore, pathway analysis goes beyond the genes list interpretation of expression levels by considering cellular interactions associated with a phenotype [12]. Pathways could also be stored and enriched by biologists' expertise interactions or inserting new data provided by metabolomics or proteomics experimentation [13]. Based on the implementation of the methodology of data integration proposed in this work, the results will contribute to facilitate the interpretation of CC gene expression data and the inference of hypothesis formulation made by biologist interactions. The integration of recognized patterns into signaling pathways, represented by Petri nets, simplifies as well the interaction with biologists for the enrichment process of signaling pathways.

In this work an introduction is presented in section1. The remainder of the paper is organized as follow; section 2 exposes relevant works related with data bases of signaling pathways and gene expression data; section 3 provides an introduction of computational models and signaling pathways; section 4 describes our methodology and in section 5 we expose our conclusion and future work.

2 Databases of Signaling Pathways and Gene Expression Data

Nowadays, available pathway databases contain organized gene regulation relationships mapped into metabolic or signaling pathways, for instance, KEGG [16], BioCarta [14] and MetaCyc [15]. Signaling pathway databases are mostly used as inert diagrams of signaling pathways, as KEGG or Biocarta. Nevertheless, some databases also provide XML or SQL interfaces of pathways data, as KGML which provides an XML abstraction of the KEGG pathway database [16]. Other efforts to collect gene regulation data, on a large scale, are based on using text mining approaches, iHOP [17].

KEGG provides a curated reference to study and analyze metabolic and signaling pathways, including different cellular processes [16]. KEGG offers metabolic and non-metabolic pathways. KGML data lacks of the details provided by pathways diagrams, for instance, some relations between proteins are not included in the KGML data. The KEGG pathway database is widely used and different approaches have been proposed to integrate the KEGG knowledge base into pathway modeling. Heiner and Koch [18], modeled apoptosis, from KEGG apoptosis diagrams, and provided a qualitative Petri Net model, enabling the confirmation of known properties as well as new insights of intrinsic and extrinsic apoptotic pathways. Other tools, as KEGGraph [19] converts KGML data into graphs, capturing the topology of KEGG diagrams; KEGGanim [20] is a visualization tool that integrates pathways and microarray expression data but lacks of interaction with biologists to enrich the signaling pathways. Cell Illustrator [21] has a connection to the KGML repository; however it is limited to the metabolism pathway acquisition. Alternatively, KEGG converter [22] is an online tool that emphasizes the conversion of KGML data into executable SBML models. In this work, we work with EIP and CP non-metabolic pathways from the KEGG database; we complement and integrate KGML data and gene expression data for pattern recognition within signaling transduction cascades. We emphasize the interaction and enrichment of results through biologists' expertise.

3 Computational Models and Developments

Different computational models could be frameworks for experimental interpretation, as expression microarrays. Notice that, acquired models could be validated, improved and enriched with accurate interpretation made by biologists. To achieve this goal several computational and formal models have been proposed, for instance, Boolean networks [23], Bayesian networks [24], graph interaction networks [25] and Petri nets [26].

Petri nets (PNs), proposed by Carl Petri, are bipartite graph representation of processes useful both for visualization and computational analysis of dynamic systems. PNs are a directed-bipartite graph with two types of nodes: places and transitions. Reddy et al. apply PNs to represent biochemical reactions networks [27]. PNs graph-structure enables biologist to track processes and the interactions among their elements. Places represent static elements of the system and transitions correspond to interactions between elements of the system. Transitions are a powerful tool representing interactions that could result in relevant semantic significances of the processes involved in the system. A formal overview of PNs related with biological systems is exposed in [28].

Different extensions of standard PNs have been proposed to model signaling pathways: coloured petri nets have been applied for modelling EGF signaling pathway [29]; stochastic petri nets captures uncertainty related within pathways [30]; finally, Matsuno et al proposed an extension of PNs to model continuous and discrete behaviours in a system [31]. In this work, we use PNs as a tool that facilitates the interaction of biologists with the modeled system.

4 Methodology

In this section, we describe the proposed methodology to integrate cervical cancer expression data and KGML data; recognition of patterns in signaling pathways, and lastly, recognized patterns to be enriched with biologists' interactions by modifying Petri net models. Fig. 1 synthesizes the tasks we propose.

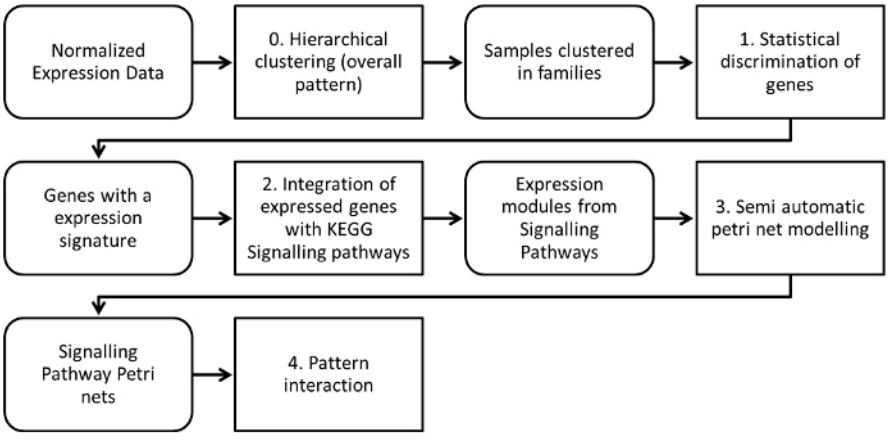

Fig. 1. This figure shows the methodology we propose. Rounded rectangles stand for input or output data, initial data is a list of normalized gene expression data and through methodology is transformed into a Petri net model. Normal rectangles represent a task or process that transforms input data.

4.1 Hierarchical Clustering

In this paper, we applied our methodology to a dataset of thirty nine cases and twelve controls. A case represents a sample of CC tissue; a control represents normal tissue. All samples were analyzed with the Affymetrix HG-Focus gene expression microarray. Each microarray represents over 8,500 genes from the NCBI RefSeq [32]. The dataset was obtained by the Unidad de Medicina Genómica team of the Hospital General de México.

Several algorithms to normalize expression data have been developed. In this work, initial input data was normalized with FlexaArray, which is a statistical program for expression microarray processing [33]. We applied a robust multi-array average (RMA) algorithm [34]. A matrix of expression data is the output of the RMA accomplishment.

With initial input data, our first question to answer is weather controls and cases have different expression profiles. Therefore, we performed an unsupervised clustering; using the R hierarchical clustering tools [35]. First, a Pearson test to measure the correlation and dependence between samples, and secondly, we use the Spearman correlation to group genes, dendogram with genes clustering not

shown.The first clustering aims to express the certainty that gene expression profiles are well-differentiated between cases and controls. In fact, four clusters of CC cases expression profiles are close in distance. Nevertheless, our first clustering is focus on quantitative data and provides biologists with a global reference of data and no evidence of cellular processes if provided. In the following section, we try to answer our next question, which genes with an expression level are important to each CC case in comparison with controls.

4.2 Statistical Discrimination of Genes

So far, hierarchical clustering delivers a differentiation between cases and controls; and four clusters of CC cases. In order to assign a significant over or sub expression level to each gene we use a z-score. Z-scores are assigned to each gene by grouping a CC cluster with the control group, using the matrix shown in figure 2. Then, z-score is calculated for each gene to assign them over, normal or sub expression values [36]. Z-scores are calculated by subtracting the total average gene intensity, within a cervical cancer group and control group, from the raw intensity data for each gene, and dividing that result by the standard deviation (SD) of all of the measured intensities, according to the formula:

$$z - score = (g_x - mean_{g_1} \dots g_n)/SD_{g_1} \dots g_n \tag{1}$$

Then, z-score is calculated for each gene to assign them over, normal or sub expression values. Genes with a z-score value under -1.96 are considered to be under expressed with respect to the media and genes with a z-score over 1.96 are considered to be over expressed with respect to the media. Z-score provides a discriminant by assigning an expression level to each gene per CC case within a cluster obtained in hierarchical clustering. A z-score with a value of 1.96, either negative or positive, represents a significant value of 0.05 for a gene to be up or down regulated. The statistical discrimination outputs a list of over and under expressed genes, which now will be integrated with KGML data to identify a signaling pathway context for each gene involved in a signal transduction process.

A	CCc_1	CCc_2	...	CCc_n
g_1	$e_{1,1}$	$e_{1,2}$	...	$e_{1,n}$
g_2	$e_{2,1}$			
...				
g_m	$e_{m,1}$	$e_{m,2}$	...	$e_{m,n}$

C	c_1	c_2	...	c_n
g_1	$e_{1,1}$	$e_{1,2}$	...	$e_{1,n}$
g_2	$e_{2,1}$			
...				
g_m	$e_{m,1}$	$e_{m,2}$	...	$e_{m,n}$

Fig. 2. In matrix A will be represented the first cluster obtained in hierarchical clustering, where, each row, g_{m} , represent a gene; each column represent a cervical cancer case, CCc_{n}; and each cell is gene expression value, $e_{\mathrm{m,n}}$. In matrix C, each row g_{m} represents a gene; each column, C_{n}, represents a control sample; and each cell is gene expression value, $e_{\mathrm{m,n}}$.

4.3 Integration of Relevant Genes and KEGG Signaling Pathways

In this section, we describe the integration of over and sub expression genes and KGML data to find each gene context within signaling pathways. As mentioned before, we work with environmental information processing (EIP) and cellular processes (CP) signaling pathways from the KEGG database.

At this point of the methodology, two subtasks must be achieved to integrate expression data and signaling pathways. First, the KGML data files are downloaded, directly from the KEGG ftp, subsequently; each KGML file is parsed to extract information. A local database, named KGMLD, is created to store information of pathways, genes of each pathway and relations between genes.

And secondly, each gene, from the microarray, with an over or sub express z-score is associated with a gene expression level by gene name matching from the KGMLD. Context for genes, with an expression level, is accomplished by searching genes that interact directly to the gene with a significant expression level.

Integration of signaling pathways and expression data is presented to biologists as shown in figure 3. Figure 3 depicts the process of integration of significant genes and a signaling pathway context, it is exemplified using a segment from the MAPK signaling pathway: 3A) first, a set of genes G, with a significant expression level, is presented to biologists; 3B) then, a set of adjacent genes N, where each gene gi from G is adjacent to one or more genes from N; 3C) finally, a set of relations, R. Relations associates each gene gi with adjacent genes belonging to the set N. Each gene, gi, could be connected with one or more genes of N.

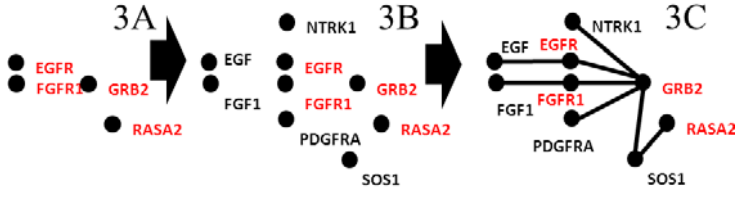

Fig. 3. This figure shows the steps to be accomplished in the integration of significant gene expression data into signaling pathway context. Vertices or genes with a name in red represent genes with a significant expression level, 3A. Genes with a name in black represent adjacent genes, 3B. Finally, relationships between both sets of genes take place by linking them, 3C.

In the subsequent task, the set of recognized graphs, GG, is integrated with the complete signaling pathway. In this example, only three graphs are displayed nonetheless the Petri net model will incorporate the complete set of graphs. Probably, not all these genes are biologically interesting. Nevertheless, we recognized a substructure inside the signaling pathway and are presented to be interpreted.

4.4 Semi-automatic Petri Net Modeling

The full context of the integration of expression data and signaling pathways is achieved in this step. As previous steps, the following subtasks are essential to achieve the final model. Firstly, a petri net model is created from KGML data, stored in the KGMLD; secondly, as mentioned, KGML data contains broken relations or missing elements, we manually incorporated missing elements based on KEGG diagram of the signaling pathway and saved in the KGMLD; finally, the set of gene graphs obtained are displayed in the proper signaling pathway petri net model. Figure 4 shows a segment of the MAPK signaling pathway with recognized expression graphs, for lack of space we present a representative segment of the MAPK signalling pathway.

As shown in Figure 4, blue places represent adjacent genes of those denoting a significant expression level and whose variation in expression could impact part of the process and genes that interact directly with them, in this particular case, a sub module of the MAPK signaling pathway. The Petri net model provides a framework for the interaction with biologists who will be able to validate or enrich the recognized patterns; in the following section we describe in detail such interaction.

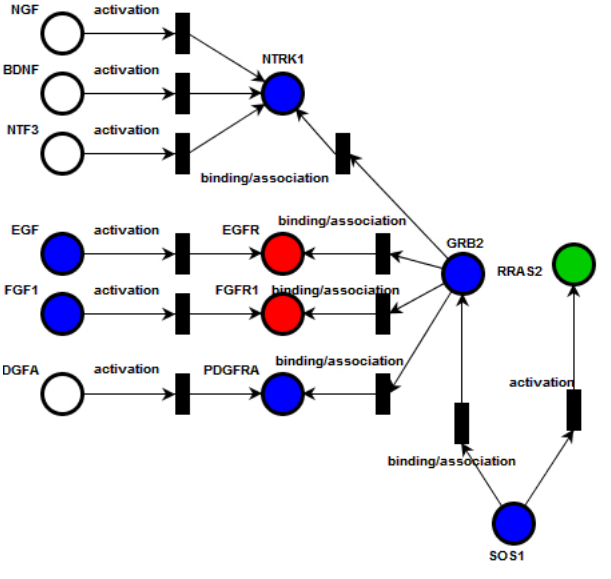

Fig. 4. This figure represents the integration of a segment of the MAPK signaling pathway and recognized genes graphs. Places in blue are genes or compounds that interact with genes with a significant expression level. Places in red denoted genes with a sub expression value, while places in green have an over expression level.

4.5 Interaction of Biologist with the Framework

The signaling pathway model represented by visual Petri nets provides biologists with an intuitive abstraction to interact. The Petri net model could be refined by incorporating personal knowledge, new data or by modifying the structure of the pathway. Operations of addition, deletion or modification of places and transitions are provided by the Petri net tool. Thus, recognized patterns require proper operations to be manipulated.

Sub-graphs or building blocks provide biologists with the capacity to manipulate patterns recognized within a signaling pathway for a better interpretation of expression microarray data. Interactions with the final output could be endless according with the interpretation or biological pursue. As demonstrated, an integrative perspective of data requires the coordination of different algorithms and computational models.

5 Conclusion

In this work, we have proposed a methodology to facilitate the interpretation of CC gene expression data and the inference of hypothesis by providing a signaling pathway context. Clustering methods, statistical discrimination, data preprocessing and systems modeling are integrated tasks to aid biologist to clarify the inner compensatory gene expression mechanisms of cervical cancer cells. The steps proposed by the methodology achieve the following: data reduction of expression profiles, selection of significantly altered genes and a visual representation of signaling pathways probably involved in the CC progression.

The facility to interrogate expression levels of thousands of genes in one experiment gives biologists a fresh look of cellular machinery compensatory events. The integration of cellular context and high-throughput expression microarray data increases the understanding of cellular systems, by providing a more interpretative model for cancer biology. At the same time the hybrid approach provides a framework to validate hypothesis.

Finally, a pattern within a signalling pathway might be represented by repetitive mutated substructure within the cascade, for instance a motif. A possible limitation of this methodology is the constant validation by biologists. In this methodology, we proposed a Petri net representation to visualize and, more significantly, to interact with identified patterns within a signalling pathway for interpretation of CC progression rather than automatization of pattern analysis.

References

1. Ferlay, J., Bray, F., Pisani, P., Parkin, D.M.: GLOBOCAN 2002; cancer incidence, mortality and prevalence worldwide. Iarc. Cancer Base No. series 5 Version 2.0. IARC Press, Lyon (2004)
2. zur Hausen, H.: Papilloma viruses in the causation of human cancers - a brief historical account. Virology 384, 260–265 (2009)

3. Jayshree, R.S., Sreenivas, A., Tessy, M., Krishna, S.: Cell intrinsic and extrinsic factors in cervical carcinogenesis. Indian J. Med. Res. 103, 286–295 (2009)
4. Ramaswamy, S., Golub, T.R.: DNA microarrays in clinical oncology. J Clin. Oncol. 20, 1932–1941 (2002)
5. Segal, E., Friedman, N., Kaminski, N., Regev, A., Koller, D.: From signatures to models: Understanding cancer using microarrays. Nat. Genet. 37, 38–45 (2005)
6. Irizarry, R.A., Hobbs, B., et al.: Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4(2), 249–264 (2003)
7. Spirin, V., Mirny, L.A.: Protein complexes and functional modules in molecular networks. Proc. Natl. Acad. Sci. USA 100, 12123–12128 (2003)
8. Jain, A.K., Dubes, R.C.: Algorithms for Clustering Data. Prentice-Hall, Englewood Cliffs (1988)
9. The Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. Nucleic Acids. Res. 32, 258–261 (2004)
10. Goh, K.-I., Cusick, M.E., Valle, D., Childs, B., Vidal, M., Barabasi, A.L.: The human disease network. Proc. Natl. Acad. Sci. 104, 8685–8690 (2007)
11. Barabsi, A.L., Oltvai, Z.: Network biology: understanding the cells functional organization. Nat. Rev. Genet. 5, 101–113 (2004)
12. Nam, D., Kim, S.Y.: Gene-set approach for expression pattern analysis. Brief. Bioinform. 79, 189–197 (2008)
13. Delongchamp, R., Lee, T., Velasco, C.A.: Method for computing the overall statistical significance of a treatment effect among a group of genes. BMC Bioinformatics 7, S11 (2006)
14. BioCarta pathways, `http://www.biocarta.com/`
15. Caspi, R., Foerster, H., Fulcher, C.A., et al.: The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids. Res. 36, D623–D631 (2008)
16. Kanehisa, M., et al.: From genomics to chemical genomics: new developments in KEGG. Nucleic Acids. Res. 34, 354–357 (2006)
17. Hoffmann, R., Valencia, A.: Implementing the iHOP concept for navigation of biomedical literature. Bioinformatics 21(ii), 252–258 (2005)
18. Heiner, M., Koch, I.: Petri net based model validation in systems biology. In: Cortadella, J., Reisig, W. (eds.) ICATPN 2004. LNCS, vol. 3099, pp. 216–237. Springer, Heidelberg (2004)
19. Zhang, J., Wiemann, S.: KEGGgraph: a graph approach to KEGG Pathway in R and Bioconductor (2008)
20. Adler, P., Reimand, J., Janes, J., Kolde, R., Peterson, H., Vilo, J.: KEGGanim: pathway animations for high-throughput data. Bioinformatics 24(4), 588–590 (2008)
21. Cell Illustrator, `http://www.cellillustrator.org/`
22. KEGG Converter, `http://www.grissom.gr/keggconverter/`
23. Shmulevich, I.: Probabilistic Boolean networks: A rule-based uncertainty model for gene regulatory networks. Bioinformatics 18, 261–274 (2002)
24. Kim, S.Y.: Inferring gene networks from time series microarray data using Bayesian networks, Brief. Bioinform. 34, 228–235 (2003)
25. Aittokallio, T., Schwikowski, B.: Graph-based methods for analysing networks in cell biology Brief. Bioinform. 7(3), 243–255 (2006)
26. Nagasaki, M., et al.: Petri Net Based Description and Modeling of Biological Pathways. Algebraic Biology, 19–31 (2005)
27. Reddy, V.N., Mavrovouniotis, M.L., Liebman, M.N.: Petri net representations in metabolic pathways. In: Proceedings of the ISMB, pp. 328–336 (1993)

28. Pinney, J.W., Westhead, D.R., McConkey, G.A.: Petri Net representations in systems biology. Biochem. Soc. Trans. 31(Pt 6), 1513–1515 (2003)
29. Zielinski, R., et al.: The crosstalk between EGF, IGF, and Insulin cell signaling pathways-computational and experimental analysis. BMC Systems Biology 3, 88 (2009)
30. Gilbert, D., Heiner, M., Lehrack, S.: A unifying framework for modelling and analysing biochemical pathways using Petri nets. In: Calder, M., Gilmore, S. (eds.) CMSB 2007. LNCS (LNBI), vol. 4695, pp. 200–216. Springer, Heidelberg (2007)
31. Matsuno, H., Tanaka, Y., Aoshima, H., Doi, A., Matsui, M., Miyano, S.: Bio pathways representation and simulation on hybrid functional Petri net. In: Silico Biology (2003)
32. Affymetrix, `http://www.affymetrix.com`
33. FlexArray: statistical data analysis software for gene expression microarrays, `http://genomequebec.mcgill.ca/FlexArray`
34. Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., Speed, T.P.: Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics, Oxford, England 4(2), 249–264 (2003)
35. The R package, `http://cran.r-project.org/`
36. Cheadle, C., Vawter, M.P., Freed, W.J., Becker, K.G.: Analysis of micro array data using Z score transformation. J Mol. Diagn. 5, 73–81 (2003)