Revealing Disease Mechanisms via Coupling Molecular Pathways Scaffolds and Microarrays: A Study on the Wilm's Tumor Disease

Alexandros Kanterakis[1], Vassilis Moustakis[1,3], Dimitris Kafetzopoulos[2], and George Potamias[1]

[1]Institute of Computer Science, FORTH, Greece
[2]Institute of Molecular Biology and Biotechnology, FORTH, Greece
[3]Department of Production Engineering and Management, TUC, Chania, Greece

Abstract. Moving towards the realization of genomic data in clinical practice, and following an individualized healthcare approach, the function and regulation of genes has to be deciphered and manifested. This is even more possible after the later advances in the area of molecular biology and biotechnology that have brought vast amount of invaluable data to the disposal of researchers. Two of the most significant forms of data come form microarray gene expression sources, and gene interactions sources – as encoded in Gene Regulatory Pathways (GRPs). The usual computational task involving microarray experiments is the gene selection procedure while, GRPs are used mainly for data annotation. In this study we present a novel perception of these resources. Initially we locate all functional paths encoded in GRPs and we try to assess which of them are compatible with the gene-expression values of samples that belong to different clinical categories (diseases and phenotypes). Then we apply usual feature selection techniques to identify the paths that discriminate between the different clinical phenotypes providing a paradigm shift over the usual gene selection approaches. The differential ability of the selected paths is evaluated and their biological relevance is assessed. The whole approach was applied on the Wilm's tumor domain with very good and indicative results.

1. Introduction

The interdisciplinary research field of molecular biology and bioinformatics is continuously enriched by the advances in many areas such as sequence analysis, genome annotation and analysis of gene and protein regulation. These advances have brought to the present the post-genomic era where, as the basic knowledge is tamed, we are mainly seek for methods that integrate various and heterogeneous

types of established biological knowledge. The major question to deal with relates to the regulation of the function of genes, targeting the ways that this function affects the overall phenotype of a living organism. A first step towards this is the combined processing of clinical and genomic data. Clinical data as the explicit expression of the phenotypic features of an organism should be integrated with the genomic data that represent the genotypic signature of the organism. This effort can help researchers to gain insights about the role of gene function in pathology, locate the risks and susceptibilities of each unique person and thus, provide individualized healthcare [1].

From a biology point of view, the goal is to provide a systematic, genome-scale view of genes interactions and functionality [2]. The advantage of this approach is that it can identify emergent properties of the underlying molecular system as a 'whole' – an endeavor of limited success if targeted genes, reactions or even molecular pathways are studied in isolation [3]. Individuals show different phenotypes for the same disease – they respond differently to drugs and sometimes the effects are unpredictable. Many of the genes examined in early clinico-genomic studies were linked to single-gene traits, but further advances engage the elucidation of multi-gene determinants of drug response. Differences in the individuals' background DNA code but mainly, differences in the underlying gene *regulation* mechanisms alter the expression or function of proteins being targeted by drugs, contribute significantly to variation in the responses of individuals. The challenge is to accelerate our understanding of the molecular mechanisms of these variations and to produce targeted individualized therapies.

In this paper we present an integrated methodology that couples and 'amalgamates' knowledge and data from both Gene Regulatory Pathway (GRP) and Microarray (MA) gene-expression sources. The methodology comprises two main parts. In the first part we decompose a number of targeted pathways – pathways involved in particular disease phenotype, into all possible functioning paths (i.e., part of a molecular pathway). Then, by introducing gene expression knowledge from a MA experiment we rank all paths according to the 'compatibility difference' they exhibit among the samples of different clinical phenotypes. In the second part of the methodology we substitute genes with paths and gene expression with compatibility value ranks. At the end we apply feature selection techniques to identify those functional paths that differentiate between the targeted phenotypic classes, and we assess their prediction power (classification) performance. As a proof of concept we apply the technique on a microarray experiment that targets the *Wilms' tumor* (WT; nephroblastoma) disease. We were able to identify significant paths in various molecular pathways that reveal distinct mechanisms between different WT phenotypes. The targeted WT phenotypes concern the tumor grade histological feature. Results are discussed about their biological relevance.

A preliminary implementation of the methodology is made in a system called MinePath. MinePath aims to uncover potential gene-regulatory 'fingerprints' and mechanisms that govern the molecular and regulatory profiles of diseases.

2. MAs and GRNs as sources of biomedical knowledge

2.1. Microarrays

Microarrays [4], [5] are devices able to measure simultaneously the expression of thousands of genes, revolutionizing the areas of molecular diagnostics and prognostics. A number of pioneering studies have been conducted that profile the expression-level of genes for various types of cancers such as leukemia, breast cancer and other tumors [6], [7]. The aim is to add molecular characteristics to the classification of diseases so that diagnostic procedures are enhanced and prognostic predictions are improved. These studies demonstrate the great potential and power of gene-expression profiling in the identification and prediction of various disease phenotypes and prognostic disease factors.

Gene-expression data analysis depends on Gene Expression Data Mining (GEDM) technology, and the involved data analysis is based on two basic approaches: (a) hypothesis testing - to investigate the induction or perturbation of a biological process that leads to predicted results, and (b) knowledge discovery - to detect underlying hidden-regularities in biological data. For the latter, one of the major challenges is gene-selection [8], [9]. Possible prognostic genes for disease outcome, including response to treatment and disease recurrence, are then selected to compose the molecular signature (gene-markers) of the targeted disease.

2.2. Gene Regulatory Pathways

GRPs are network structures that depict the interaction of DNA segments during the transcription of genes into mRNA. The prominent and vital role of GRPs in the study of various biology processes is a major sector in contemporary biology research, where numerous thorough studies have been conducted and reported [10], [11]. From a computational point of view, GRPs can be conceived as analogue of biochemical computers that regulate the level of expression of target genes [12]. Each network has inputs, usually proteins or transcription factors that initiate the network function. The outputs are usually certain proteins (encoded by specific genes). The network by itself acts as a mechanism that determines cellular behavior where the nodes are genes and edges are functions that represent the molecular reactions between the nodes. These functions can be perceived as Boolean functions, where nodes have only two possible states ("on" and "off"), and the whole network being represented as a simple directed graph [13]. The notion of

GRPs is by itself an abstraction of the underlying chemical dynamics of the cell, thus the expectation of high reliability in terms of modeling is limited. It is indicative that most of the relations in known and established GRPs have been derived from laborious and extensive laboratory experiments and careful study of the existing biochemical literature. Thus GRPs are far from being complete, at least with respect to their ability to capture and model all the internal cell dynamics of complex living organisms.

Current efforts focus on the reconstruction of GRPs by exploring gene-expression data. For example in [14] it is reported that network topologies, as extracted from gene co-expression events, could discover motifs and regulatory hubs that can characterize the entire cellular states and guide further pharmaceutical research. Very few methods of gene regulatory inference are considered superior, mainly because of the intrinsically noisy property of the data, 'the curse of dimensionality', and the lack of knowledge about the 'true' underlying structure of the networks.

The study of the function, structure and evolution of GRPs in combination with microarray gene-expression profiles and data is essential for contemporary biology research. First of all, researchers have uncovered a multitude of biological facts, such as protein properties and genome sequences. But this alone is not sufficient to interpret biological systems and understand their robustness, which is one of the fundamental properties of living systems at different levels [15]. This is mainly because cell, tissues, organs, organisms or any other biological systems defined by evolution are essentially complex physicochemical systems. They consist of numerous dynamic networks of biochemical reactions and signaling interactions between active cellular components. This cellular complexity has made it difficult to build a complete understanding of cellular machinery to achieve a specific purpose [16]. To circumvent this complexity microarrays and molecular networks can be combined in order to document and support the detected and predicted interactions [17]. The advances and tools that each discipline carries can be integrated in a holistic and generic perspective so that the chaotic complexity of biology networks can be 'screened' and traced down.

2.3. Coupling MAs and GRPs

Microarray experiments involve more variables (genes) than samples (patients). This fact, leads to results with poor biological significance. There is an open debate whether we should concentrate on gathering more data or on building new algorithms in order to improve biological significance. Simon et al. in [18] published a very strict criticism on common pitfalls on microarray data mining while in [19] comments about the bias in the gene selection procedure are presented. Moreover, due to limitations in DNA microarray technology higher differential expressions of a gene do not necessarily reflect a greater likelihood of the gene being related to a disease and therefore, focusing only on the candidate genes with

the highest differential expressions might not be the optimal procedure [20]. Another significant aspect is the noisy content microarray experiments. Appropriate statistical analysis of noisy data is very important in order to obtain meaningful biological information [21], [22]. Evidence on this is given by the fact that different methods produce gene-marker lists that are strikingly different [23]. As a result, and because the immature state of microarray technology, reproducibility of microarray experiments and the accompanied statistical prediction models are pretty low, except when protocols are uniformly and strictly followed [24], [25].

In the light of the aforementioned observations and in order to overcome the posted limitations we have to consider MA-based gene-expression profiles just as an instance of biological information, strongly connected - rather than isolated, from other sources of related biological knowledge. In other words, gene-expression profiles should be examined, explored and interpreted not as 'static' but as instances of the underlying regulatory framework, as encoded by established and known GRPs.

3. Methodology

Existing GRPs databases provide us with widely utilized networks of proved molecular validity. The most known are network that describe important cellular processes such as cell-cycle, apoptosis, signaling, and regulation of important growth factors. Online public repositories contain a variety of information that includes not only the network per se but links and rich annotations for the respective nodes (genes) and edge (regulation). In the current study we utilize the KEGG pathways repository. KEGG provides a format representation standardized by its own markup description language (KGML).

The gene regulatory relations we consider are restricted to what might be observed in a microarray experiment: a change in the expression of a regulator gene modulates the expression of a target gene mainly via protein-DNA interactions. In other words, there are genes that causally regulate other genes. A change in the expression of these genes might change dramatically the behavior of the whole network. The identification and prediction of such changes is a challenging task in bioinformatics. Moreover, we have to identify real, true networks and use them as scaffolds [26] to methods that infer gene regulatory networks out from gene expression data. This approach can aim several areas of biology research such as genomic medicine [27], microarray data mining [28] and phylogenetic analysis [29]. We have implemented our approach on coupling GRPs and MA data in a system called MinePath.

3.1. Pathway decomposition

MinePath relies on a novel approach for GRP processing that takes into account all possible functional interactions of the network, the network's *scaffolds*. The

different GRP scaffolds correspond to the different functional paths that can be followed during the regulation of a target gene.

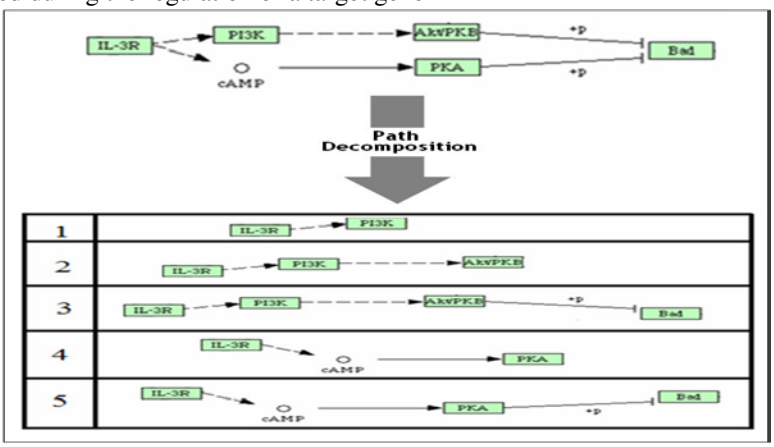

Fig. 3.1. Funtion-path decomposition – the GRP scaffolds: Top: A target part of the KEGG cell-cycle GRP; Bottom: The five decomposed fictional paths (scaffolds) for the targeted path part – all possible functional routes taking place during network regulation machinery.

Different GRPs are downloaded from the KEGG repository. With an XML parser (based on the specifications of the KGML representation of GRPs) we obtain all the internal network semantics (see next sub-section). In a subsequent step, all possible and functional network paths are extracted as exemplified in Fig 3.1. Each functional path is annotated with the possible valid values according to Kauffman's principles that follow a binary setting: each gene in a functional path can be either 'ON' or 'OFF'. According to Kauffman [13], the following functional gene regulatory semantics apply: (a) the network is a directed graph with genes (inputs and outputs) being the graph nodes and the edges between them representing the causal (regulatory) links between them; (b) each node can be in one of the two states: 'ON' or 'OFF': 'ON' and 'OFF' states correspond to the gene being expressed (i.e., the respective substance being present) or not expressed, respectively; and (c) time is viewed as proceeding in discrete steps - at each step the new state of a node is a Boolean function of the prior states of the nodes with arrows pointing towards it.

Since the regulation-edge connecting two genes defines explicitly the possible values of each gene, we can set all possible state-values that a gene may take in a path. Thus, each extracted path contains not only the relevant sub-graph but the state-values of the involved genes as well. The only requirement concerns the following assumption: for a path being functional it should be 'active' during the GRP regulation process; in other words we assume that all genes in a path are functionally active. For example, assume the functional path A → B ('→' is an activation/expression regulatory relation). If gene A is on an 'OFF' state then, gene B is not allowed to be on an 'ON' state - B could become 'ON' only and only if it

is activated/expressed by another gene in a different functional path (e.g., C → B). The assumption follows a 'closed world assumption', that is: if what we know is just the 'A → B' gene-gene interaction then, B could be activated only from A; if A is inactive there is no causal evidence for B being active. If we had allowed non-functional genes to have arbitrary values then the significant paths would be more likely to be 'noisy' rather than exhibiting some form of biological importance.

After parsing the targeted GRPs, the involved genes are stored in a database that acts as a repository for future reference. Through this repository we can query paths being parts of target GRPs, GRPs that contain specific genes or target a specific regulatory relation. Moreover, the stored paths can be combined and analyzed in the view of specific microarray experiments and respective gene-expression sample profiles. As the database repository contain and retrieves functional paths from a variety of different GRPs (e.g., cell-cycle, apoptosis etc), we may combine different molecular pathways and networks – a major need for molecular biology and a big challenge for systems biology and contemporary bioinformatics research.

3.2. Combining gene-expression profiles and functional paths

The next step is to locate microarray experiments and respective gene-expression data for which we expect (suspect) the targeted GRPs play an important role. For example the cell-cycle and apoptosis GRPs play an important role in tumorgenesis and cancer progression.

With a gene-expression/functional-path *matching* operation, the valid and most prominent GRP functional paths are identified. These paths uncover and present potential underlying gene regulatory mechanisms that govern the gene-expression profile of the samples under investigation. Such a discovery may guide to the finer classification of samples as well as to the re-classification of diseases, providing the most prominent molecular evident for that.

3.3. Matching GRP paths with MA data

The samples of a binary transformed (discretized) gene-expression matrix are matched against targeted molecular pathways and respective GRP functional paths (retrieved form the described repository). We follow a gene-expression value discretization process presented elsewhere (please refer to [9]). As already exemplified, GRP and MA gene-expression data matching aims to differentiate GRP paths and identify the most prominent functional paths for the given samples. In other words, the quest is for the paths that exhibit high matching scores for one of phenotypic class and low matching scores for another. This is a paradigm shift from mining for genes with differential expression to mining for subparts of GRP with

differential function. The algorithm for differential path identification is inherently simple (see Fig. 3.3).

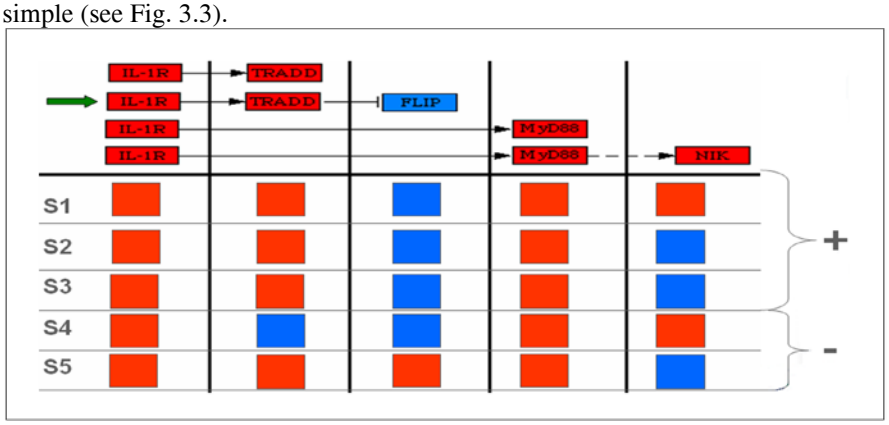

Fig. 3.3. Matching Functional-paths (scaffolds) and gene-expression profiles: Samples S1, S2, S3 belong to the '+' class and samples S4, S5 belong to the '-' class. The first path (IL-1R → TRADD) satisfies samples 1,2,3,5. Second path (IL-1R → TRADD ⊣ FLIP) satisfies samples S1, S2, S3. Third path satisfies all samples and the forth path doesn't satisfy any sample. The green arrow indicates that the second path yields the maximum differential power and it contains a potential function differentiation since it is consisted only with samples that belong to the '+' class. ('→': activation; ' ⊣': inhibition).

– For each path we compute the number of samples that is consistent for each disease phenotypic class. Suppose that there are S_1 and S_2 samples belong to the two classes, respectively. Assume that path P_i is consistent with $S_{i;1}$ and $S_{i;2}$ samples form the first and second class, respectively. Formula 1,

$$\left|\left(S_{i;1} / S_1\right) - \left(S_{i;2} / S_2\right)\right| \quad (1)$$

computes the *differential power* of the specific path with respect to the two classes. Ranking of paths according to formula 1 provides the most differentiating and prominent GRP functional paths for the respective disease phenotypes. These paths present evidential molecular mechanisms that govern the disease itself, its type, its state or other targeted disease phenotypes (e.g., positive or negative response to specific drug treatment). The formula can be enriched so that longer consistent paths acquire stronger power. It can also be relaxed so that 'consistent' is a continuous indicator rather than a Boolean value. Finally we may introduce 'unknown' values for missing and erroneous gene expression values.

4. Revealing Regulating Wilm's tumor Molecular Mechanisms

The presented MA-GRP coupling methodology was applied on a study for expression profiling of the Wilm's tumor (WT, nephroblatoma) disease [30]. In the orig-

inal publication the researchers report new candidate genes for various WT clinical phenotypes.

WT samples were divided according to the *histological risk grade* ('low/intermediate' and 'high'), relapse of tumor ('no', 'yes'), survival ('relapse-free' and 'death'), metastasis ('no', 'yes') and response to chemotherapy ('good', 'poor'). The results presented in this paper focus on the histological risk grade as a target WT phenotype. In the original published study a set of 20 differentially expressed genes are reported for this WT phenotype [30].

From the ArrayExpress online microarray experiments' repository (http://www.ebi.ac.uk/microarray-as/ae/) we downloaded the expression values and the clinical annotation of 138 samples from the WT study - 108 of them being classified as histology risk 'low/intermediate', and 30 as 'high'. For this study we targeted 17 GRPs – the selection was made on the basis of their susceptibility and incrimination to the WT disease and on established biological and clinical knowledge of their involvement in cell regulatory tumor growth mechanisms. The path decomposition process resulted into a total of 8937 functional paths. Most of these paths didn't show any special differential ability over the samples. In order to identify the significant paths the matching gene-expression formula (formula 1 presented in section 3.3) was applied. A threshold value of 0.5 was set to filter-out not differential paths (the threshold was fixed after experimentation with various cut-off values). Filtering resulted into a set of 87 functional-paths for further exploration.

The next step was to find the most relevant and discriminant functional-paths, and build a classifier able to distinguish between the two phenotypic classes - 'high' and 'low' (including 'intermediate' samples) histological risk grade, respectively. The whole dataset is presented as a binary-{0,1} array-matrix of 87 lines for functional paths, and 138 columns for samples. The value "1' in the position *i,j* of this array means that the *i* path is 'active' for sample *j*. Active means that all genes that comprise this path are either 'ON' or 'OFF' according the interaction relationships of the genes of the path. Respectively, a '0' value means that the genes involved in the path do not exhibit the same value as the expression value of the respective sample. The array-matrix can be seen as an indicator of which paths are functional on which samples. Furthermore, it comprises a resemblance of normal gene-expression matrices - instead of genes being either active or inactive, according to their expression over different kind of samples, we have paths being functional or non-functional over the same set of samples. This gives us the ability to apply whatever feature selection processes to select the most relevant and discriminant functional-paths. For this, we rely on a feature/gene-selection algorithmic process presented in [9].

Initially a Wilcoxon rank-sum test ($p<0.005$) was applied that reduced the functional-paths from 87 to 54. Then, ranking and selection of the most discriminant functional-paths was performed – ranking is based on an information-theoretic entropic formula, and selection encompasses a naïve Bayes classification process

[9]. The whole process resulted into a complex of four discriminant and indicative functional-paths (see Fig. 4.)

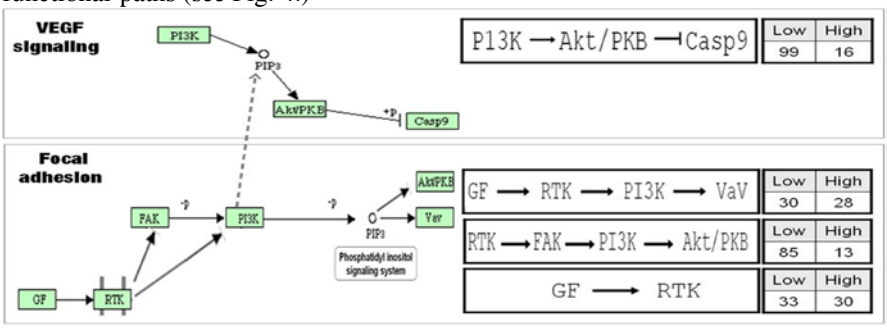

Fig 4. Indicative 'low'/'high' histological risk functional-paths for the WT disease: The GRP name, its KEGG graph representation, the identified functional-paths, and the coverage/discrimination statistics, e.g., (99,16) in the upper part of the figure, indicates that the specific path [P13K → Akt/PKB -| Casp9] covers 99 'low' and 16 'high' risk samples, respectively. The presence of 'P13K' in both GRPs is quite interesting for the biology of the WT disease, and respective therapeutic targets.

The four selected functional-paths are involved in two different GRPs: one from 'VEGF signaling', and three from 'Focal adhesion'. The three functional-paths from 'Focal adhesion' are subparts of the whole 'Focal adhesion' GRP.

We performed a Leave One Cross Validation (LOOCV) procedure in order to assess the discrimination/classification performance that these paths exhibit (note that each functional-path is now considered as a feature). A 95% LOOCV (131/138 samples) accuracy figure was achieved when, the fitness (i.e., train vs. train) figure was inferior, 91% (126/138, 8 misses for 'high' classified as 'low', and 4 misses for the inverse case). This finding is quite interesting: beside the high accuracy performance data 'overfitting' is reduced. This is a strong indication for the high relevance of the four identified functional-paths (at least for the available dataset).

In addition, we applied the same feature/gene selection algorithmic process on the original gene-sample matrix (i.e., the normal gene-selection setting for microarray gene-expression profiles). This resulted into the same LOOCV accuracy and in 89% (123/128) fitness (17 genes selected) accuracy figures. A potential speculation on this finding is the following: with the presence of thousands of genes and of a limited number of samples, gene-selection processes are lean to overfitting events. We believe that this explains the diversity of results produced by available gene-selection techniques, their instability on different population (for the same disease) cohorts, and the inability to relate statistical significance with biological relevance. In contrast, the introduced methodology is able to identify not just discriminant gene-markers but, discriminant, indicative and 'stable' gene-regulatory mechanisms that govern disease phenotypes and clinical manifestations.

In a preliminary attempt to find biological evidence for our findings we focus on the involvement of 'P13K' and 'AkT'gene-products in both identified GRPs (see Fig. 4). The related literature report the 'P13K/AkT' complex to be implicated in WT disease, as well as it is the main component of WT therapeutic targets

[31]. Certainly, further biological validation of the approach is needed, a task for future research.

Conclussions

We have presented an integrated methodology for the coupling of both GRPs and MA gene expression profiles. In the heart of the methodology is the decomposition of GRPs into functional-paths (or, scaffolds), the matching of these paths with samples' gene expression profiles, and the application of feature selection techniques for the identification of the most relevant and discriminant ones.

Application of the methodology on gene-expression data for the Wilms' tumor disease showed that: we can identify a limited number of functional-paths that exhibit significantly differential behavior between different WT phenotypes ('low/intermediate' vs. 'high' histological grade risk). The findings provide valuable insights for further research over the function and role of the involved genes and their underlying regulatory machinery.

Among others, our on-going and future R&D work include: (a) further experimentation with various real-world microarray studies and different GRP targets (accompanied with the evaluation of results form molecular biology and clinical research experts); (b) extension of path decomposition to multiple GRPs; (c) elaboration on more sophisticated path/gene-expression profile matching formulas and operations; (d) incorporation of different gene nomenclatures in order to cope with microarray experiments from different platforms and nomenclatures; and (e) porting of the whole methodology in a Web-Services and scientific workflow environment.

Acknowledgements: This work was supported by the European Commission's Sixth Framework Programme in the context of the ACGT (FP6-2005-IP-026996) Integrated project.

References

[1] J. Bell, 'Predicting disease using genomics', Nature 429, 453-456 (2004).
[2] T. Ideker, T. Galitski and L. Hood, 'A new approach to decoding life: systems biology', Annu Rev Genomics Hum Genet, 2, 343-372 (2001).
[3] F.S. Collins, E.D. Green, A. E. Guttmacher and M. S. Guyer, 'A Vision for the Future of Genomics Research', Nature, 422(6934), 835-847 (2003).
[4] H.F. Friend, 'How DNA microarrays and expression profiling will affect clinical practice', Br Med J., 319, 1-2 (1999).
[5] D.E. Bassett, M.B. Eisen, and M.S. Boguski, 'Gene expression informatics: it's all in your mine', Nature Genetics, 21(Supplement 1), 51-55 (1999).
[6] T.R. Golub et al., 'Molecular classification of cancer: class discovery and class prediction by gene expression monitoring', Science, 286, 531-537 (1999).
[7] L.J. van 't Veer et al., 'Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer', Nature, 415, 530-536 (2002).
[8] M.E. Troyanskaya, M.E. Garber, P.O. Brown, D. Botstein, and R.B. Altman, 'Nonparametric methods for identifying differentially expressed genes in microarray data', Bioinformatics, 18 (11), 1454-1461 (2002).

[9] G. Potamias, L. Koumakis and V. Moustakis, 'Gene Selection via Discretized Gene-Expression Profiles and Greedy Feature-Elimination', LNAI, 3025, 256-266 (2004).
[10] J. M. Bower and H.Bolouri, Computational Modeling of Genetic and Biochemical Networks, Computational Molecular Biology Series, MIT Press, 2001.
[11] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, Molecular Biology of the Cell, Garland Science, New York, 2002.
[12] Arkin and J. Ross, 'Computational functions in biochemical reaction networks', Biophys J., 67(2), 560-578 (1994).
[13] S. A. Kauffman, The Origins of Order: Self-Organization and Selection in Evolution, Oxford Univ. Press, New York, 1993
[14] N.M. Babu, N.M. Luscombe, L. Aravind, M. Gerstein and S.A. Teichmann, 'Structure and evolution of transcriptional regulatory networks', Curr. Opin. Struct. Biol., 14, 283-291 (2004).
[15] H. Kitan, 'Robustness from top to bottom', Nat. Genet.,38, 133 (2006).
[16] H. Kitano, 'Systems biology: a brief overview', Science, 295(5560), 1662-1664 (2002).
[17] K. Kwoh and P. Y. Ng, 'Network analysis approach for biology', Cell. Mol. Life Sci., 64, 1739-1751 (2007).
[18] R. Simon, M. D. Radmacher, K. Dobbin and L. M. McShane, 'Pitfalls in the Use of DNA Microarray Data for Diagnostic Classification', Journal of the National Cancer Institute, 95(1), 14-18, (2003).
[19] Ambroise and G. J. McLachlan, 'Selection bias in gene extraction on the basis of microarray gene-expression data', PNAS, 99(10), 6562-6566, (2002).
[20] S. Draghici, S. Sellamuthu and P. Khatri, 'Babel's tower revisited: a universal resource for cross-referencing across annotation databases', Bioinformatics, 22(23), 2934-2939 (2006).
[21] D.K. Slonim, 'From pattern to pathways: gene expression data analysis comes of age', Nature Genetics, 32, 502-508 (2002).
[22] J. Quackenbush, 'Computational Analysis of Microarray Data', Nature Reviews Genetics, 2, 418-427 (2001).
[23] W. Pan, 'A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments', Bioinformatics, 18(4), 546-554 (2002).
[24] MTRC: Members of the Toxicogenomics Research Consortium, 'Standardizing global gene expression analysis between laboratories and across platforms', Nature Methods, 2, 351-356 (2005).
[25] Robert et al. 'Robust interlaboratory reproducibility of a gene expression signature measurement consistent with the needs of a new generation of diagnostic tools', BMC Genomics, 8:148 (2007).
[26] T. Ideker and D. Lauffenburger, ,Building with a scaffold: emerging strategies for high- to low-level cellular modeling', Trends in Biotechnology, 21(6), 255-262 (2003).
[27] M.A. Hoffman, 'The genome-enabled electronic medical record', Journal of Biomedical Informatics, 40(1), 44-46 (2007).
[28] P. Jares, 'DNA Microarray Applications in Functional Genomics', Ultrastructural Pathology, 30, 209-219, (2006).
[29] R. Jothi, T. M Przytycka and L. Aravind, 'Discovering functional linkages and uncharacterized cellular pathways using phylogenetic profile comparisons: a comprehensive assessment', BMC Bioinformatics, 8:173 (2007).
[30] Zirn B, Hartmann O, Samans B, Krause M, Wittman S, Mertens F, Graf N, Eilers M, Gessler M. Expression profiling of Wilms tumor reveals new candidate genes for different clinical conditions. Int. J. Cancer: 118, 1954-1962 (2006).
[31] International Society of Paediatric Oncology - SIOP Education Book. http://www.siop.nl/content/files/SIOP_Educational_Book_2008.pdf Acessed 8 march, 2009.