Methods in Molecular Biology (2016) 1375: 137–153
DOI 10.1007/7651_2015_252

Published online: 02 July 2015

Integrating Microarray Data and GRNs

L. Koumakis, G. Potamias, M. Tsiknakis, M. Zervakis, and V. Moustakis

Abstract

With the completion of the Human Genome Project and the emergence of high-throughput technologies, a vast amount of molecular and biological data are being produced. Two of the most important and significant data sources come from microarray gene-expression experiments and respective databanks (e,g., Gene Expression Omnibus—GEO (http://www.ncbi.nlm.nih.gov/geo)), and from molecular pathways and Gene Regulatory Networks (GRNs) stored and curated in public (e.g., Kyoto Encyclopedia of Genes and Genomes—KEGG (http://www.genome.jp/kegg/pathway.html), Reactome (http://www.reactome.org/ReactomeGWT/entrypoint.html)) as well as in commercial repositories (e.g., Ingenuity IPA (http://www.ingenuity.com/products/ipa)). The association of these two sources aims to give new insight in disease understanding and reveal new molecular targets in the treatment of specific phenotypes.

Three major research lines and respective efforts that try to utilize and combine data from both of these sources could be identified, namely: (1) de novo reconstruction of GRNs, (2) identification of Gene-signatures, and (3) identification of differentially expressed GRN functional paths (i.e., sub-GRN paths that distinguish between different phenotypes). In this chapter, we give an overview of the existing methods that support the different types of gene-expression and GRN integration with a focus on methodologies that aim to identify phenotype-discriminant GRNs or subnetworks, and we also present our methodology.

Keywords: Microarray, Gene expression, Gene regulatory networks, Pathways, Functional pathways, Bioinformatics, Systems biology

1 Introduction

In recent years, high-throughput data capture technology, as with microarray platforms, have vastly improved life scientists' ability to detect and quantify gene, protein, and metabolite expression. The most common type, two-color microarrays, can measure the expression of tens of thousands of genes with a single chip (1). Applications include measuring gene expression in different developmental stages, identifying biomarkers for particular phenotypes or diseases, and monitoring treatment response.

In the systems biology framework, scientists follow a "holistic" approach in order to explore and study the behaviour of biological components. System biology provides a global view of the dynamic interactions in a biological system. On the molecular level, the purpose of the underlying systems biology computational approaches is to ascertain the interactions and dynamic behavior of molecules within a cell (2). The molecular mechanisms determine how cells interact and how they develop and maintain higher levels of organization and function. Systems biology tries to formulate these mechanisms in mathematical models.

Currently bioinformatics community focuses on more enhanced methods for gene selection on microarrays mainly by adding and *amalgamating* knowledge from other sources, such as GRNs. Integrating GRN information into the class comparison, discovery, and prediction process is an important issue in bioinformatics, mainly because the provided information possesses a true biological content. By changing the focus from individual genes to a set of genes or pathways, the gene set analysis (GSA) approach enables the understanding of cellular processes as an intricate network of functionally related components. A performance evaluation of GSA methodologies (3) concluded that the inclusion of additional biological features such as topology or covariates would be more useful than simple gene selection approaches. In addition, utilizing more domain knowledge is likely to reveal more insights in the analysis.

Similar to bioinformatics, systems biology community took advantage of the human genome and the microarray technology to reconstruct and validate gene regulatory networks in an automatic way. GRN reconstruction or reverse engineering aims toward the inference GRN models from data (in most of the cases from gene expression data). In the literature, a large number of computational methods are reported with the target of inferring gene regulatory networks from expression data (4).

A relatively new line of research in the field is the identification of the most discriminant GRNs, or parts of GRNs that differentiate between specific phenotypes by coupling GRNs and microarray data. Assessment of the discriminant power of (sub)networks is based on the identification of those genes whose expression values are consistent, i.e., could be justified, by their corresponding interaction pattern in the target GRN.

The study of the function, structure, and evolution of GRNs in combination with microarray gene-expression profiles is essential for contemporary biology research. Due to limitations in DNA microarray technology—due to the different platforms utilised, to the different experimental protocols, and mainly to small sample sizes, 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 (5, 6).

Based on our knowledge, we propose a taxonomy of the methodologies that combine gene-expression data and GRNs in order to identify and assess discriminant pathway and subpathways (Fig. 1) and a taxonomy of methodologies which identify and assess discriminant pathway and sub-pathways (Fig. 2).

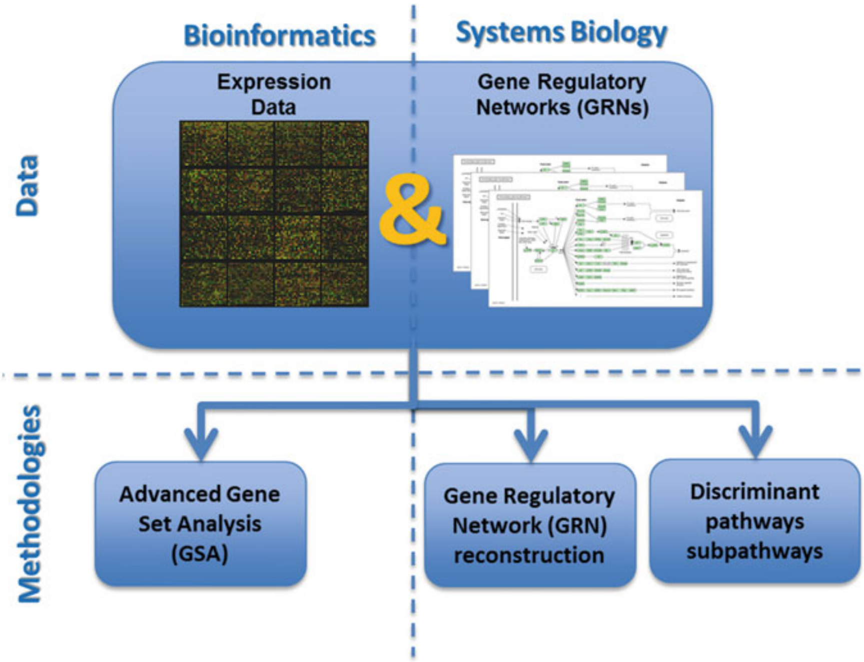

Bioinformatics
Systems Biology
Data
Expression Data
&
Gene Regulatory Networks (GRNs)
Methodologies
Advanced Gene Set Analysis (GSA)
Gene Regulatory Network (GRN) reconstruction
Discriminant pathways subpathways

Fig. 1 Integration of microarray data with gene regulatory networks

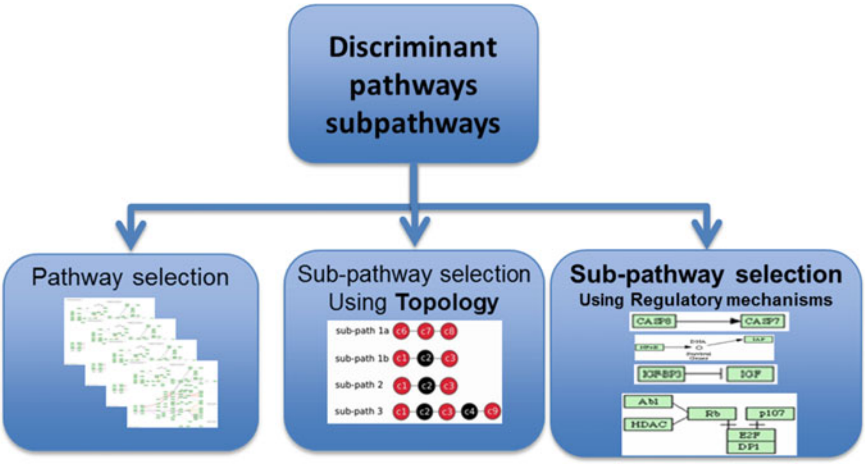

Discriminant pathways subpathways
Pathway selection
Sub-pathway selection Using Topology
sub-path 1a
sub-path 1b
sub-path 2
sub-path 3
Sub-pathway selection Using Regulatory mechanisms
CASP8
CASP7
IGFBP3
IGF
Abl
HDAC
Rb
p107
E2F
DP1

Fig. 2 Taxonomy of discriminant pathways and sub-pathways

A general observation concerns the different levels of knowledge extraction from the GRNs employed by the different methods. The first category naming *pathway selection* focuses on the identification of differentially expressed pathways using microarray data. Within this approach information about the topology, the existing subpaths, as well as the reactions/relationships between genes in a pathway are ignored. The second category *subpathway* *selection using topology* goes one step further and tries to identify the discriminant pathways or subpathways. Within this approach identification and selection of the most discriminant paths ignore the present gene relations/regulations. The last and most informative category is the *subpathway selection using regulatory mechanisms.* This approach takes advantage of the GRN topology as well as the type of GRN gene relations (e.g., activation or inhibition).

Initial efforts used GRN information as groups (plain list) of associated genes in order to identify the most discriminant and phenotype-differentiating genes. Molecular pathways effectively reduced the resulting sets of genes, extracted from a gene set analysis approach, and in some cases improved prediction performance. But GRNs encompass much more knowledge form just a plain list of genes. Recently, more and more methods take advantage of the GRNs topology and the underlying gene interaction patterns.

Pathway selection methodologies show similarities with gene signatures in terms of the level of information used over the years. Although GRNs hold important information about the structure and correlation among genes that should not be neglected, most of the currently available methods in pathway selection do not fully exploit it. In the literature, one can find three categories of methodologies that focus on the identification and selection of discriminant pathways and subpathways, based on the different levels of knowledge extraction from target GRNs. Initially the focus was on the identification of differentially expressed pathways (as a whole) using microarray data. Then the efforts concentrated on the knowledge of the GRN topology using decomposition mechanisms to reveal discriminant subpathways based on the graph theory concepts and network visualization toolkits. Recently more advanced methodologies are developed, which takes in consideration not only the topology of the GRNs but also the regulation type (activation/inhibition) of the interaction link that connects two or more genes.

One can easily identify three main categories of methodologies according to the level of the utilised GRN information. The categories are pathway selection using GRNs as list of genes, subpathway selection using the topology of GRNs, and subpathway selection methodologies using the underlying GRN gene regulatory interactions. The last category—being in its infancy—exhibits the fewer methodologies so far, but it takes the most out of GRNs and gene-expression data compared to the other two, and is a promising alternative for the identification of the regulatory mechanisms that underlie and putatively govern various phenotypes.

The subpathway selection using the underlying GRN gene regulatory interactions approach solves the major problem of the set enrichment strategies that refers to the conflicting constrains between GRNs and gene-expression data. A typical example of the conflicting constrains is reflected in the situation when two significantly up-regulated genes increase the enrichment of the set in microarray expression data, even if the first gene inhibits the other in a GRN.

2 Method

We introduce a new methodology for the identification of differentially expressed functional paths or subpaths within a gene regulatory network (GRN) using microarray data analysis. The analysis takes advantage of interactions among genes (e.g., activation, inhibition) as nodes of a graph network, which are derived from expression data.

We propose a novel perception of GRNs and gene expression data (Fig. 3). Initially we locate all functional paths encoded in GRNs 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). The differential power of the selected paths is computed and their biological relevance is assessed. The whole approach is applied on a set of microarray studies with the target of revealing putative regulatory mechanisms that govern the treatment responses of specific phenotypes.

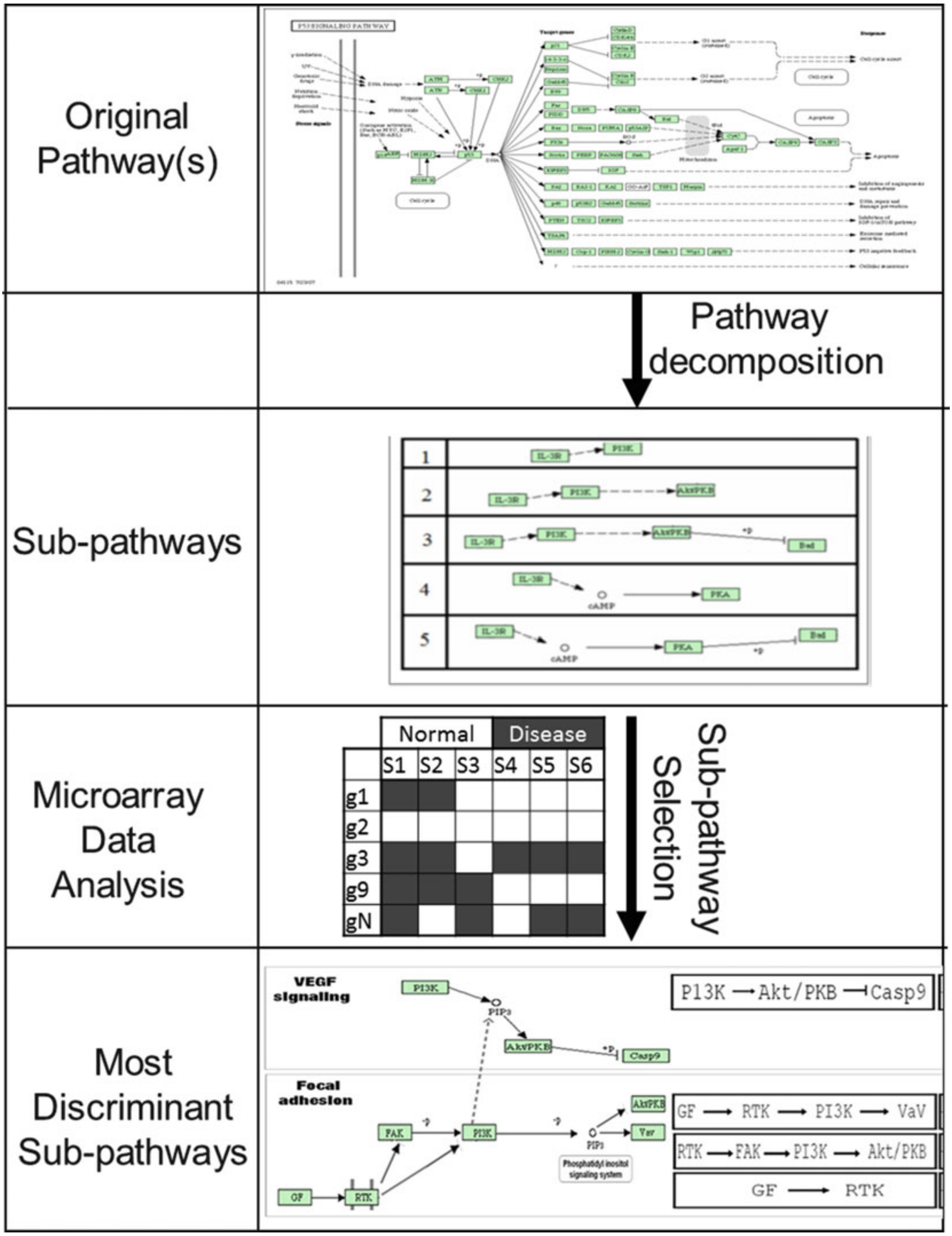
GRN and gene-expression data matching aims to differentiate GRN paths and identify the most prominent functional sub-paths for the given samples. In other words, the quest is for the subpaths that exhibit high-matching scores for one of phenotypic class and low-matching scores for the other. This is a paradigm shift from the mining of differential genes to the mining of GRN functional subpaths. The algorithm for differential subpath identification is inherently simple.

Figure 4 provides an indicative example of the gene expression limitation, where samples S1, S2, S3 belong to the "+" class and samples S4, S5 belong to the "−" class. At the first sight, we can see that no gene or no group of genes can discriminate 100 % our two classes ("+" and "−").

Figure 5 highlights the paradigm shift from the mining of differential genes to the mining of GRN functional subpaths. Given the same example as previously, we check our samples against known sub-paths of GRNs.

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 fourth 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 contains only with samples that belong to the "+" class ("→": activation; "—|": inhibition).

We rely on a novel approach for GRN processing that takes into account all possible functional interactions in the network. Gene-



	Normal			Disease		
	S1	S2	S3	S4	S5	S6
g1	■	■	□	□	□	□
g2	□	□	□	□	□	□
g3	■	■	□	■	■	■
g9	■	■	■	□	□	□
gN	■	□	■	□	■	■

Fig. 3 Flow of operations

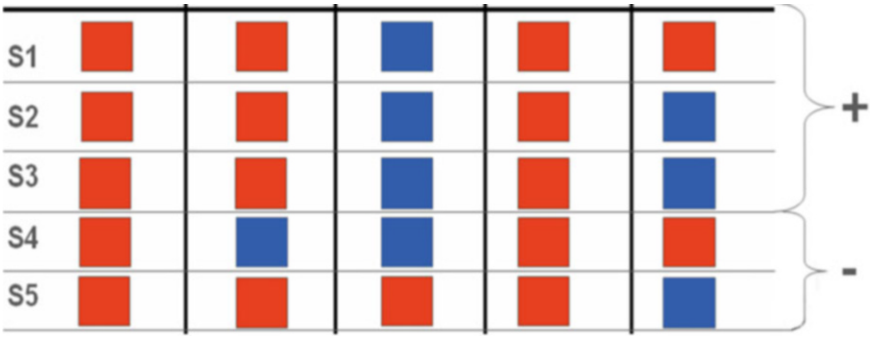


Fig. 4 Gene expression data example

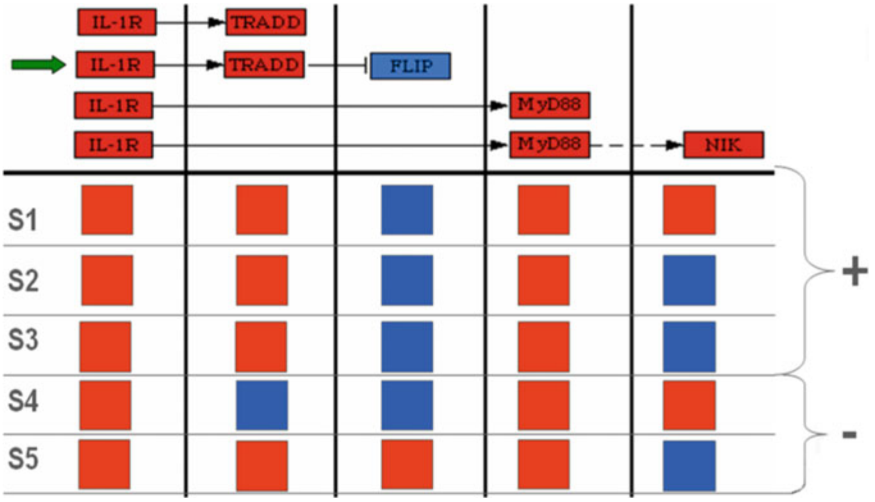


Fig. 5 Matching functional sub-paths and gene-expression profiles

expression samples profiles and their phenotype assignments are extracted form microarray data, and all targeted GRNs are evaluated for the identification of the most informative ones.

The method unfolds into three modular steps.

1. *Data preprocessing*: On the one hand, gene expression values are discretized into two states with values 1 and 0 for up-regulated and down-regulated genes. On the other hand, each target GRN is decomposed into its constituent subpaths.
2. *Data annotation*: Each subpath is interpreted on the basis of its functional active-state, represented by a binary ordered-vector with active states, resulting into its active-state ordered vector <1,1,0> for the corresponding genes.

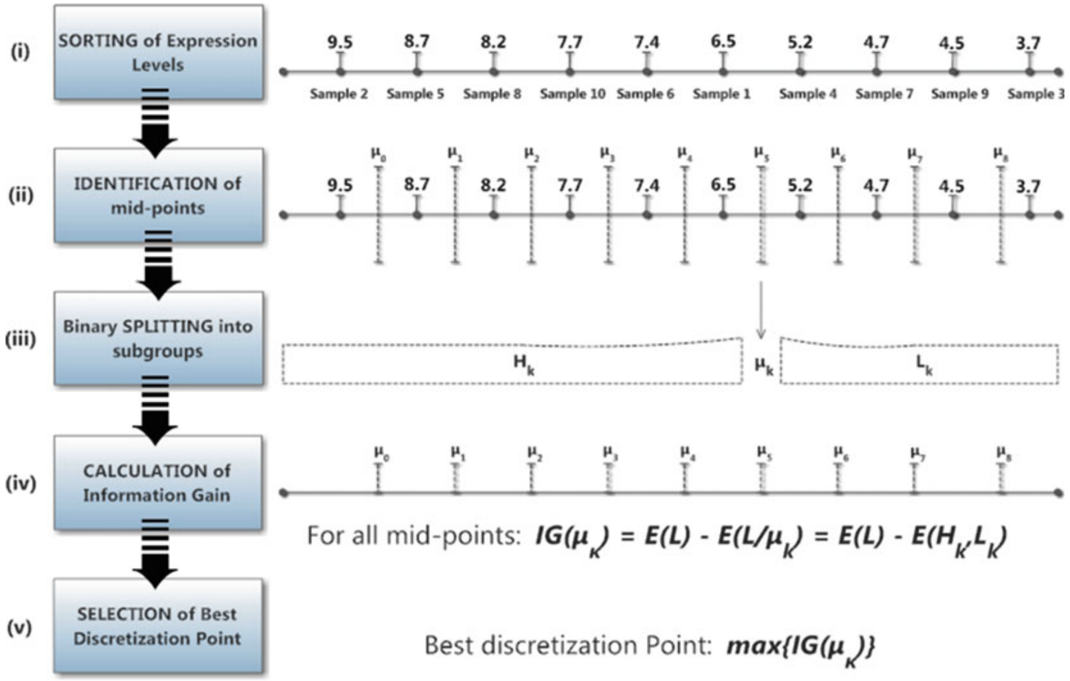
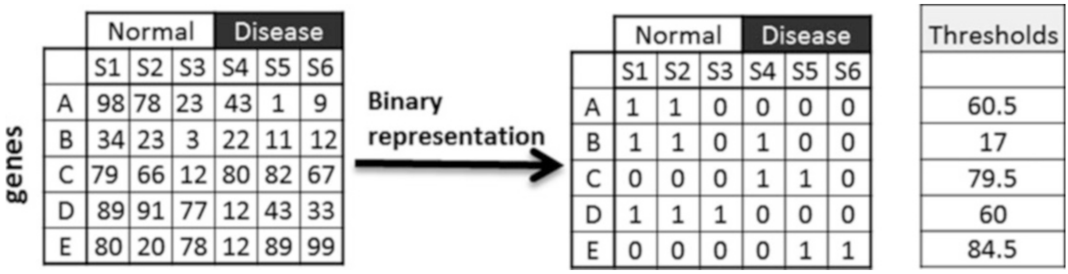

(i)
SORTING of Expression Levels
9.5
8.7
8.2
7.7
7.4
6.5
5.2
4.7
4.5
3.7
Sample 2
Sample 5
Sample 8
Sample 10
Sample 6
Sample 1
Sample 4
Sample 7
Sample 9
Sample 3
(ii)
IDENTIFICATION of mid-points
μ_0
μ_1
μ_2
μ_3
μ_4
μ_5
μ_6
μ_7
μ_8
(iii)
Binary SPLITTING into subgroups
H_k
μ_k
L_k
(iv)
CALCULATION of Information Gain
For all mid-points: $IG(\mu_K) = E(L) - E(L/\mu_K) = E(L) - E(H_K, L_K)$
(v)
SELECTION of Best Discretization Point
Best discretization Point: $max\{IG(\mu_K)\}$

Fig. 6 The gene discretization process

3. *Analysis* (*data mining*): The binary ordered-vector of each subpath is aligned and matched against all (discretized) binary gene-expression sample profiles. The subpaths are taking the place of sample descriptor features and utilized for the construction of subpath based phenotype prediction models.

2.1 Data Preprocessing

We utilize discretization of the gene-expression continuous values into the core of the gene-selection process. Discretization of a given gene's expression values means that each value is assigned to an interval of numbers that represents the expression-level of the gene in the given samples. A variable set of such intervals may be utilized and assigned to naturally interpretable values e.g., *low*, *high*. Given the situation that, in most of the cases, we are confronted with the problem of selecting genes that discriminates between two classes (i.e., disease-states) and we believe that it is convenient to follow a two-interval discretization of gene-expression patterns. Below we give a general statement of the discretization problem when two classes are present, followed by an algorithmic process that heuristically solves it. Therefore, expression value represented with 0 indicates a nonexpressed or underexpressed gene, whereas value of 1 indicates overexpressed gene. These values are being derived using the following process (as also shown in Fig. 6):



genes
Binary representation

	Normal			Disease		
	S1	S2	S3	S4	S5	S6
A	98	78	23	43	1	9
B	34	23	3	22	11	12
C	79	66	12	80	82	67
D	89	91	77	12	43	33
E	80	20	78	12	89	99

	Normal			Disease			Thresholds
	S1	S2	S3	S4	S5	S6	
A	1	1	0	0	0	0	60.5
B	1	1	0	1	0	0	17
C	0	0	0	1	1	0	79.5
D	1	1	1	0	0	0	60
E	0	0	0	0	1	1	84.5

Fig. 7 Microarray discretization, an indicative example

1. The expression levels of gene $\boldsymbol{A}$ over the total number of samples are sorted in descending order.
2. The midpoints between each two consecutive values are calculated.
3. For each midpoint, the samples are clustered into two subgroups, $\boldsymbol{H}$ and $\boldsymbol{L}$.
4. For each midpoint, an information gain formula is applied, which computes the entropy (7) of the system in respect to its division into subgroups. $\boldsymbol{IG(\mu_\kappa)}$ is the Information Gain of the system for midpoint $\boldsymbol{\mu_\kappa}$. $\boldsymbol{E(L)}$ is the total entropy of the system taking into account their prior assignment into classes (e.g., case–control), whereas $\boldsymbol{E(L/\mu_\kappa)} = \boldsymbol{E(H_\kappa,L_\kappa)}$ is the entropy of the system taking into account its division into subgroups around midpoint $\boldsymbol{\mu_\kappa}$.
5. Finally, the midpoint that results in the highest information gain is selected as the one which best discriminates against the two subgroups, and all the samples in the $\boldsymbol{H}$ group are considered to be overexpressed getting a value of **1**, whereas the ones in the $\boldsymbol{L}$ group are the nonexpressed/underexpressed, getting a value of **0**.

This discretization process is applied to each gene separately, and the final dataset is a matrix of discretized values. A similar approach has been used before in other expression profiling studies (8, 9). Figure 7 shows an indicative example of a "dummy" microarray with five genes (rows) and six samples (columns) categorized into two classes, normal and diseased. To the left of the figure we can see the absolute or normalized values of our "dummy" microarray and to the right we have the discretized matrix when we applied the proposed methodology.

On the other hand, the origin of concurrent knowledge about GRNs does not come from any concrete theoretic framework. However, although incomplete, this knowledge covers almost every biology function such as metabolism, genetic/environmental information processing, cellular processes, human diseases, and drug development, while it is constantly under refinement and enrichment. We chose to incorporate KEGG data for our analysis. Since its first introduction in 1995, KEGG DB for pathways has been widely used as a reference knowledge base for understanding biological pathways and functions of cellular processes. The knowledge from KEGG has proven of great value by numerous works in a wide range of fields (10).

Although it has been shown that KEGG has some errors (11), they are not so prominent and can be counterbalanced by the simplicity, the variety and the standard ontology that KEGG provides. Through KEGG public database, pathways can be downloaded in KGML[1] format. KGML (stands for KEGG Markup Language) is an exchange format of KEGG graph objects including GRNs. The GRN is described through standard graph annotation. Nodes can be either genes, groups of genes, compounds, or other networks. Edges can be one of the gene relations known from the biology theory (activation, inhibition, expression, indirect, phosphorylation, diphosphorylation, ubiquination, association, and dissociation). Each gene relation has a different semantic that depicts the precise biology phenomenon that happens during the regulation of the specific network.

Our approach relies on a novel processing for GRN that takes into account all possible functional interactions of the network. The different interactions correspond to the different functional subpaths that can be followed during the regulation of a target gene.

GRNs are downloaded from the KEGG repository. With an XML parser (based on the specifications of KEGG's KGML representation of GRNs), we obtain all the internal network semantics. In a subsequent step, all possible functional network subpaths are extracted as exemplified in Fig. 8.

2.2 Data Annotation

We exploit microarray experiments and respective gene-expression data for which we expect (suspect) the targeted GRNs play an important role. 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 the fine classification of samples as well as the reclassification of diseases, based on the most prominent molecular evidence. The samples of a binary transformed (discretized) gene-expression matrix are matched against targeted molecular pathways and respective GRN functional paths (retrieved form the pathway decomposition).

A translation between the genes identifiers used in the gene expression data to the corresponding KEGG identifiers is needed.

[1] http://www.kegg.jp/kegg/xml/

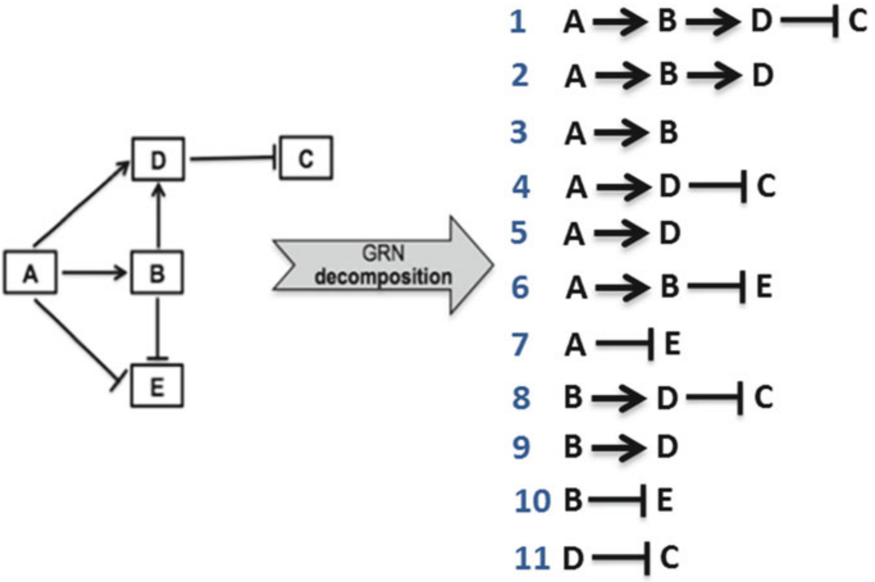


Fig. 8 Functional-path decomposition: Left: A target part of an artificial GRN; Right: The ten decomposed functional sub-paths

Both the GRNs and the gene expression data have to use the same ids. GRNs use gene ids while gene expression platforms use probes. A probe is a specific segment of single-strand DNA that is complementary to a desired gene. For example, if the gene of interest contains the sequence AATGGCACA, then the probe will contain the complementary sequence TTACCGTGT. When added to the appropriate solution, the probe will match and then bind to the gene of interest.

Due to the large number of databases and associated IDs, the conversion of gene identifiers is one of the initial and central steps in many workflows related to genomic data analysis. In the literature and the web, we can find several freely available ID conversion tools. Although each tool has distinct features and strengths, as reviewed by Khatri et al. (12), they all adopt a common core strategy to systematically map a large number of interesting genes in a list to the associated biological annotation.

The mapping from a thesaurus to another rises the many to one issue which in our case many probes from the gene expression dataset are assigned to the same KEGG gene ID. We check the multiple probes for the gene and place a logic OR for the assessment of the gene's value. This is actually the selection of the value of the probe with the highest intensity out of all the probes that map to the same gene.

Then we need to identify the subpaths that exhibit high-matching scores for one of phenotypic class and low-matching scores for the others. Each GRN subpath is interpreted according to Kauffman's principles and semantics (13):

1. The network is a directed graph with genes (inputs and outputs) being the graph nodes and the edges between them representing the *causal* links between them, i.e., the *regulatory* reactions.
2. Each node can be in one of the two states, "ON," the gene is expressed or up-regulated (i.e., the respective substance being present) or, "OFF," the gene is not-Expressed or down-regulated.
3. 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.

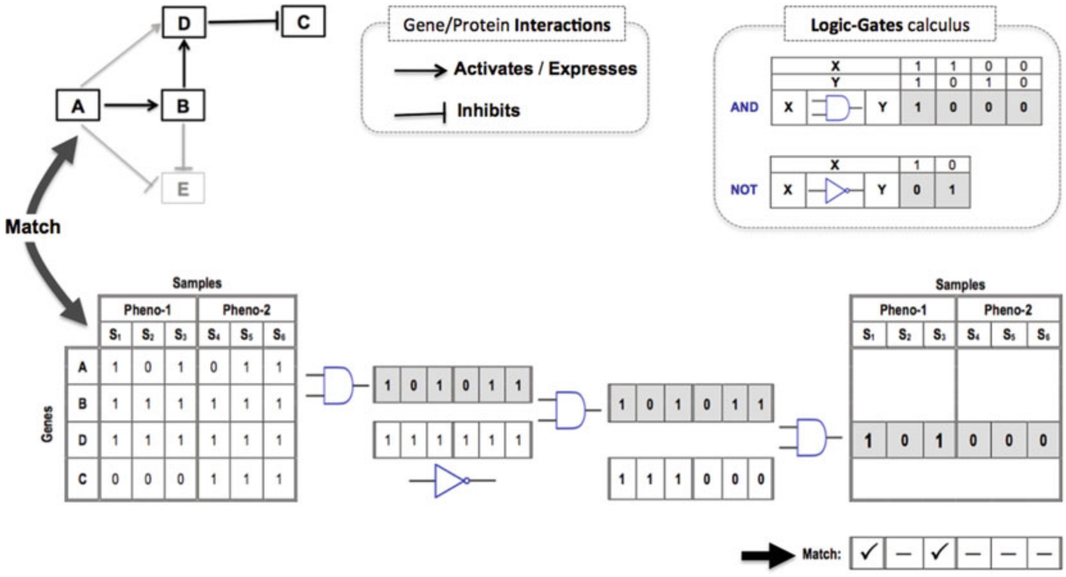
In order to cope with and reveal functional regulatory mechanisms we impose the following requirement over the formed subpaths: for a subpath to be considered as functional it should be "active" during the GRN regulation process—in other words we assume that all genes in a subpath are functional. For example, consider the reaction A → B, if A is "ON" then the activation/expression ("→") regulatory reaction is active, resulting into the activation/expression of gene B ("ON")—the same holds for an inhibition (—|) reaction. In the case that gene A is "OFF" then the reaction is considered as inactive with the state of the regulated gene B to remain undetermined. Under this assumption, a *path-module* is just a subpath (atomic or more complex) for which all its reactions are considered as active. So, the state of all genes engaged in a path-module that forms an *ordered regulation pattern*, e.g., the pattern of the complex regulatory mechanism A → D —| C is <"ON," "ON," "OFF">.

The samples of a binary transformed (discretized) gene-expression matrix are matched against functional path-modules of target GRNs. We follow an information-theoretic gene-expression discretization process.

2.3 Data Analysis

As an example, assume the gene-expression binary profiles of six artificial samples for genes A, B, D and C—with "1" to denote "ON" and "0" to denote "OFF"—three of them are assigned to phenotype-1 (S_1, S_2, and S_3) and the other three to phenotype-2 (S_4, S_5, and S_6)—refer to Fig. 9.

Furthermore, assume the artificial GRN shown in the left part of Fig. 9, and its subpath A → B → D —| C (in bold). We follow a *logic-gates* process that aims to match the path-module instance of the subpath with the respective samples' binary instances. The process results into the formation of an ordered pattern that indicate the samples for which the target sub-path is consistent with ("1"s) or not ("0"s), i.e., the respective path-module A="ON" → B="ON" → D="ON" —| C="OFF" is active.



	Pheno-1			Pheno-2		
	S1	S2	S3	S4	S5	S6
A	1	0	1	0	1	1
B	1	1	1	1	1	1
D	1	1	1	1	1	1
C	0	0	0	1	1	1

X	1	1	0	0
Y	1	0	1	0
AND	1	0	0	0

X	1	0
NOT	0	1

	Pheno-1			Pheno-2		
	S1	S2	S3	S4	S5	S6
	1	0	1	0	0	0
Match:	✓	–	✓	–	–	–

Fig. 9 Matching gene-expression sample profiles with GRN functional path-modules: a logic-gates approach

Note that for the finally inferred pattern of Fig. 9, <1,0,1,0,0,0>, value "1" occurs in positions one and three, which means that the examined path-module is active for samples one and three; in all other samples it is inactive ("0"). As samples one and three belong to phenotype-1, the target path-module matches 2 out of 3 phenotype-1 samples, and zero phenotype-2 samples. In general, assume that there are S_1 and S_2 samples that belong to phenotype-1 and phenotype-2, respectively, and that path-module P_i matches $S_{i;1}$ and $S_{i;2}$ samples form phenotype-1 and phenotype-2, respectively. Formula 1, computes the *differential power* of a path-module with respect to the two phenotypic classes;

Formula 1

$$S_{i;1}/S_1 - S_{i;2}/S_2$$

The formula posses a *polarity* characteristic according the class phenotype: positive for class S_1 and negative for class S_2; e.g., for the above example, the differential power of path-module A="ON" → B="ON" → D="ON" —| C="OFF" is (2/3) – 0 = 0.67, and as it positive it is interpreted and considered as a regulation mechanism that governs phenotype-1.

After the decomposition of each of these pathways into its functional components, each subpath has been matched against the respective samples' gene-expression profiles of the respective microarray studies. The result is an array of sub-paths with binary values for every sample in the form of a discretized microarray. Then using the machine learning library WEKA (14) we can extract the most discriminant subpaths using ranking algorithms. A feasibility study of the methodology approach is presented in the following section.

2.4 Experiments

Most of breast cancer (BRCA) cases are estrogen responsive, implying the activation of a series of growth-promoting pathways, for example, the estrogen receptor (ER) related ErbB signaling GRN. In an effort to reveal the underlying regulatory mechanisms that govern BRCA patients' treatment responses we applied our methodology on a public gene-expression study from the GEO, the GSE7390[2] dataset targeting the ER phenotypic status of the respective patients, i.e., ER+ (ER positive) vs. ER– (ER negative).

We targeted 14 pathways all of which are engaged within the "Pathways in Cancer" integrated pathway of KEGG (hsa05200) namely: ECM-receptor interaction (hsa04512), Cytocin-cytocin receptor interaction (hsa04060), Adherens junction (hsa04520), Wnt signaling (has04310), Focal adhesion (hsa04510), Jak-STAT signaling (hsa04630), ErbB signaling (hsa04012), MAPK signaling (hsa04010), mTOR signaling (hsa04150), VEGF signaling (hsa04370), Apoptosis (hsa04210), p53 signaling (hsa04115), Cell cycle (hsa04110), and TGF-β signaling (hsa04350).

The visualization of the results for the ErbB signaling (hsa04012) can be found in Fig. 10 where with the help of the Cytoscape[3] graph library. The graph preserves the KEGG layout topology. It is enriched with the expressed regulatory mechanisms (relations) between genes that differentiate between the two phenotypes and the color coding is as follows:

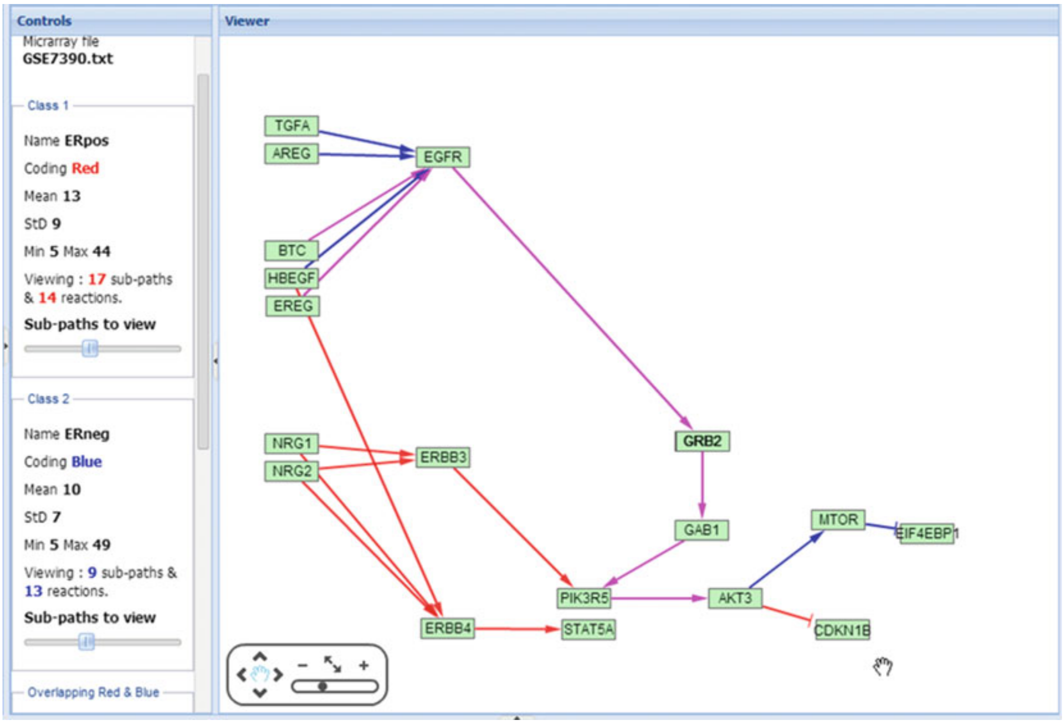
- Red indicates relations active at class 1 which in our example is the ERpos.
- Blue indicates relations active at class 2 (ERneg).
- Magenta indicates overlapping relations in the two classes.
- Orange for subpaths that are always active.

The figure highlights only the "interesting" subpaths which in our case are the most discriminant subpaths for the specific two phenotypes.

Inspecting the reduced network, it is clear that there is a pathway starting from NRG (1 and 2) and ends at inhibiting the CDKN1B for ERpos phenotype; and a pathway starting from TGFA or AREG or HBEFG that ends-up at inhibiting EIF4EBP1 for ERneg phenotype.

[2] http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=gse7390

[3] http://www.cytoscape.org/



Controls

Micrarray file
GSE7390.txt

Class 1

Name ERpos
Coding Red
Mean 13
StD 9
Min 5 Max 44
Viewing : 17 sub-paths & 14 reactions.
Sub-paths to view

Class 2

Name ERneg
Coding Blue
Mean 10
StD 7
Min 5 Max 49
Viewing : 9 sub-paths & 13 reactions.
Sub-paths to view

Overlapping Red & Blue

Fig. 10 Results of GSE7390 over 14 cancer related pathways

According to recent literature, the aforementioned results are quite relevant to the estrogen-receptor status. Based on a search of the related biomedical literature we focus our exploration on the mechanisms underlying the resistance to pure estrogen antagonists (e.g., fulvestrant). Recent studies show the significant role of both ErbB3 and ErbB4 as alternative targets for the treatment of BRCA patients. As Sutherland notes in ref. (15): "the initial growth inhibitory effects of fulvestrant appear compromised by cellular plasticity that allows rapid compensatory growth stimulation via ErbB-3/4. Further evaluation of pan-ErbB receptor inhibitors in endocrine-resistant disease appears warranted." In addition, Hutcheson et al. in ref. (16) investigated whether induction of ErbB3 and/or ErbB4 may provide an alternative resistance mechanism to antihormonal action. Their conclusion is that fulvestrant treatment is sensitive to the actions of the ErbB3/4 ligand HRGb1 (NRG1) with enhanced ErbB3/4-driven signaling activity, and significant increases in cell proliferation.

3 Discussion and Conclusions

Current trend in GRNs and gene expression data is the subpathway selection using regulatory mechanisms, which seems that it is at its first steps and could possibly gain a momentum. Our assumption for that momentum amplifies with the similarities we can find between the discriminant gene regulatory (sub)networks and microarray gene selection methodologies.

Apart the proposed procedure, only four (4) other tools take advantage of the underlying GRN gene regulation mechanisms, naming GGEA (17), SPIA (18), TEAK (19), and PATHOME (20). The main difference of the proposed methodology from these four systems is the handling of the gene regulatory mechanisms. To our knowledge all the other methodologies count with a +1 the activations and −1 the inhibitions. Each subpath gets a final score which is also used as a ranking mechanism. Contrary, our approach strictly checks and takes into account only subpaths that are functional (according to the gene relations and the expression values). Our approach is binary and leads to distinction between functional and nonfunctional subpaths per sample instead of a representation of the sub-path per class (the sum).

Our methodology relies on a novel approach for GRN processing that takes into account all possible functional interactions of the network. The phenotype information is extracted from microarrays and all the selected GRNs are evaluated for the identification of the most informative GRNs at the specific phenotype. The efficient ranking of subpaths provides the most differentiating and prominent GRN functional subpaths for the respective target phenotypes. The formula posses a polarity characteristic according the class phenotype, i.e., positive for class S1 and negative for class S2. These subpaths 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 methodology was applied on a gene-expression study with the target of identifying putative mechanisms that underlie and govern the treatment response of breast cancer patients according to their ER-status profiles. Results were quite indicative and strongly supported by the relevant biomedical literature.

It is known that integrating heterogeneous data sources is more effective than working within the boundaries of a single scientific technology/field. Bioinformatics and systems biology has proven that taking advantage of the knowledge from each other can aid the relevant scientific communities in their research endeavours or even reveal and create new research domains. In most of the cases there are levels of integration as well as levels of knowledge to be utilised. Extracting out the most of the knowledge will always give us more natural and meaningful, as well as more accurate results.

Acknowledgment

This work was supported by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement N° 270089 and by the European Union (European Social Fund—ESF) and by the European Union (European Social Fund—ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF)—Research Funding Program: Heracleitus II Investing in knowledge society through the European Social Fund.

References

1. Brown PO, Botstein D (1999) Exploring the new world of the genome with DNA microarrays. Nat Genet 21:33–37
2. Huang Y, Zhao Z, Xu H, Shyr Y, Zhang B (2012) Advances in systems biology: computational algorithms and applications. BMC Syst Biol 6(3)
3. Hung J-H, Yang T-H, Zhenjun H, Weng Z, DeLisi C (2012) Gene set enrichment analysis: performance evaluation and usage guidelines. Brief Bioinform 13(3):281–291
4. Heckera M, Lambecka S, Toepferb S, van Somerenc E, Guthke R (2009) Gene regulatory network inference: data integration in dynamic models—a review. Biosystems 96(1):86–103
5. Ein-Dor L, Kela I, Getz G, Givol D, Domany E (2005) Outcome signature genes in breast cancer: is there a unique set? Bioinformatics 21 (2):171–178
6. Iwamoto T, Pusztai L (2010) Predicting prognosis of breast cancer with gene signatures: are we lost in a sea of data? Genome Med 2(11):81
7. Shannon CEA (1948) Mathematical theory of communication. Bell Sys Tech J 27 (3):379–423
8. Potamias G, Koumakis L, Moustakis V (2004) Gene selection via discretized gene-expression profiles and greedy feature-elimination. Meth Appl Artif Intelligence 3025:256–266
9. Li L, Weinberg CR, Darden TA, Pedersen LG (2001) Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. Bioinformatics 17(12):1131–1142
10. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Yamanishi Y (2008) KEGG for linking genomes to life and the environment. Nucleic Acids Res 36:480–484
11. Ott MA, Gert V (2006) Correcting ligands, metabolites, and pathways. BMC Bioinformatics 7(1):517
12. Khatri P, Draghici S (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. Bioinformatics 21:3587–3595
13. Kauffman SA (1993) The origins of order: self-organization and selection in evolution. Oxford University Press, New York
14. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Ian H (2009) The WEKA data mining software: an update. SIGKDD Explorations 11(1)
15. Sutherland RL (2011) Endocrine resistance in breast cancer: new roles for ErbB3 and ErbB4. Breast Cancer Res 13(3):106
16. Hutcheson IR et al (2007) Heregulin beta1 drives gefitinib-resistant growth and invasion in tamoxifen-resistant MCF-7 breast cancer cells. Breast Cancer Res 9(4):50
17. Geistlinger L, Csaba G, Küffner R, Mulde N, Zimmer R (2011) From sets to graphs towards a realistic enrichment analysis of transcriptomic systems. Bioinformatics 27(13):366–373
18. Tarca AL, Draghici S, Khatri P, Hassan SS, Mittal P, Kim JS, Kim CJ, Kusanovic JP, Romero R (2009) A novel signaling pathway impact analysis. Bioinformatics 25(1):75–82
19. Judeh T, Johnson C, Kumar A, Zhu D (2013) TEAK: Topology Enrichment Analysis frameworK for detecting activated biological subpathways. Nucleic Acids Res 41(1):1425–1437
20. Nam S, Chang HR, Kim KT et al (2014) PATHOME: an algorithm for accurately detecting differentially expressed subpathways. Oncogene 33(41):4941–4951