Molecular BioSystems

Cite this: *Mol. BioSyst.*, 2012, **8**, 3224–3231

www.rsc.org/molecularbiosystems

PAPER

A signal transduction score flow algorithm for cyclic cellular pathway analysis, which combines transcriptome and ChIP-seq data†

Zerrin Isik,[a] Tulin Ersahin,[b] Volkan Atalay,[c] Cevdet Aykanat[d] and Rengul Cetin-Atalay*[b]

Received 31st May 2012, Accepted 7th September 2012
DOI: 10.1039/c2mb25215e

Determination of cell signalling behaviour is crucial for understanding the physiological response to a specific stimulus or drug treatment. Current approaches for large-scale data analysis do not effectively incorporate critical topological information provided by the signalling network. We herein describe a novel model- and data-driven hybrid approach, or signal transduction score flow algorithm, which allows quantitative visualization of cyclic cell signalling pathways that lead to ultimate cell responses such as survival, migration or death. This score flow algorithm translates signalling pathways as a directed graph and maps experimental data, including negative and positive feedbacks, onto gene nodes as scores, which then computationally traverse the signalling pathway until a pre-defined biological target response is attained. Initially, experimental data-driven enrichment scores of the genes were computed in a pathway, then a heuristic approach was applied using the gene score partition as a solution for protein node stoichiometry during dynamic scoring of the pathway of interest. Incorporation of a score partition during the signal flow and cyclic feedback loops in the signalling pathway significantly improves the usefulness of this model, as compared to other approaches. Evaluation of the score flow algorithm using both transcriptome and ChIP-seq data-generated signalling pathways showed good correlation with expected cellular behaviour on both KEGG and manually generated pathways. Implementation of the algorithm as a Cytoscape plug-in allows interactive visualization and analysis of KEGG pathways as well as user-generated and curated Cytoscape pathways. Moreover, the algorithm accurately predicts gene-level and global impacts of single or multiple *in silico* gene knockouts.

Introduction

Recent genomic data collections have become publicly available for whole genomes of several species during the last decade. In parallel, omics-wide experimental technologies have been developed. Combined with the advent of supporting bioinformatics tools, the high-throughput technology has been commonly exploited in a range of disease conditions such as cancer and neurodegenerative pathologies.[1] These large-scale biological datasets are often integrated and represented in various forms of cell signalling networks, which are composed of a group of biomolecules working together to control cellular behaviour in response to a signal. It is widely recognized that a coordinated response of a combination of genes is responsible for most cellular behaviour and related phenotypes.[2–4] Hence, studying the complex architecture of signalling networks with novel algorithmic approaches with the experimental data can demonstrate how complex biological traits arise and propagate.

Traditional transcriptomics data-analysis methods identify a list of significant genes that are expected to be related to a particular cellular phenotype. However, analysis of the large-scale experimental data based only on a list of significant genes falls short of revealing the molecular basis of cellular events. Therefore, specific methodologies to manipulate and analyse these data collections still remain to be developed.

Cell signalling networks are often represented in the form of node-edge structured graphs. The nodes (vertices) and edges of these graphs represent biomolecules (proteins or small molecules) and physical interactions between them, respectively. KEGG,[5] BioGRID[6] and Reactome[7] are some of the data sources often used for integration of omics data into cell signalling networks. Several bioinformatics tools have been developed to associate large-scale data, especially microarray gene expression, with pathway graphs.[8–15] These tools aim to interpret the expression

[a] *Department of Bioinformatics, Biotechnology Center, Technical University Dresden, 01307 Dresden, Germany. E-mail: zerrin.isik@biotec.tu-dresden.de*
[b] *Department of Molecular Biology and Genetics, Faculty of Science, Bilkent University, 06800 Ankara, Turkey. E-mail: ersahin@bilkent.edu.tr, rengul@bilkent.edu.tr; Fax: +90-312-266-5097; Tel: +90-312-290-2503*
[c] *Department of Computer Engineering, Middle East Technical University, 06531 Ankara, Turkey. E-mail: volkan@ceng.metu.edu.tr*
[d] *Computer Engineering Department, Bilkent University, 06800 Bilkent, Ankara, Turkey. E-mail: aykanat@cs.bilkent.edu.tr*

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c2mb25215e

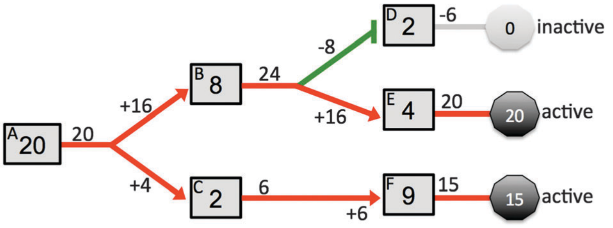


Fig. 1 Demonstration of the score flow operation on a three step hypothetical pathway with final cellular activity processes (apoptosis, cell cycle *etc*.). Raw data scores are shown in protein nodes in squares. Protein A is the initiator protein, which transforms the 8/10 fraction (10 is the sum of the scores of B and C) of its score to protein B and the 2/10 fraction to protein C. Then protein B gives out the sum of its raw score plus the incoming score from A (+16 + 8 = 24) to proteins D and E. Interaction with protein D is of inhibitory type, therefore protein D's incoming edge gets a negative value of −8. Hence, its target final cellular process is inactive with a negative score (−8 + 2 = −6). Other target processes are active according to the accumulated scores.

profiles by identifying experimental condition related genes or pathways based on traditional statistical tests. Hence they generate gene co-expression networks of only the selected pool of genes. Most of the methods perform pathway analysis based on either significant gene sets or gene functional class identifications. Although some tools provide quantitative enrichment scores for the genes or gene-groups, they do not use the topological structure of the pathway or the biological activity of a specific sub-cellular process responsible for the observed phenotype. Therefore, this study aims to design and implement a signal transduction score flow algorithm to quantitatively assess biological activities of cellular processes and to identify significant sub-paths (downstream process) within that pathway using not only a selected subset of genes but for all the gene nodes in a given pathway.

We previously described a feed-forward score flow algorithm for large-scale data annotation and its relation to cellular networks.[16] Our current study focuses on the design and implementation of a signal transduction score flow algorithm that quantitatively assesses biological activities of a cyclic cellular network and identifies significant sub-paths and target cellular processes in a given pathway. The cyclic network algorithm was also implemented as a Cytoscape plug-in, then applied on 30 different KEGG pathways by using two different data sets. Significance analysis of final activity scores of target processes was performed. *In silico* knock-out studies were analysed on a curated pathway. Our approach fuses and exploits both data and model, effectively benefiting from topological information brought in by cell signalling pathways. A pathway was converted into a graph and the individual gene scores were mapped onto the nodes of the graph. Gene scores were transferred *en route* to the biological pathway to form a final activity score, describing the behaviour of a specific process in the pathway while enriching the gene node scores.

Methods

Pathway node score calculation

Based on the omics data, an initial score was assigned to each protein node of a given pathway. The protein node scores were obtained by taking products of the rank scores extracted from experimental data (explained in *Raw Data Preparation Documentation*, ESI†). Initial raw data analysis for microarray and ChIP-seq was done by R and CisGenome frameworks respectively.[17] Then score computation on the pathway with the partitioned score transfer procedure was initiated. Usually, cell signaling flows from cell membrane towards the nucleus in order to activate certain cellular activities upon a signal from receptors. Therefore our algorithm simulates this signal flow after the initial score assignment. First the nodes (proteins), which are close to the membrane transmit their scores to their immediate edges, then to nodes in their immediate neighborhood. If a protein has two or more interacting partners in the immediate neighborhood, the initial score of the protein is partitioned to the interacting edges based on the weights of the interacting neighborhood nodes' raw data score (Fig. 1). Then the new edge score and the interacting node score are added up in order to calculate the output of the next step in the signaling path. The partitioning idea of this approach is based on the stoichiometric concentrations of protein–protein interactions. The number of interactions/reactions in a cellular system is based on the substance concentration; therefore we adopted raw data scores as the stoichiometric concentrations of the node and partitioned raw data scores, based on their interacting partners' stoichiometric concentrations. A fraction of the raw data score is transferred to one neighbour based on its raw score while the rest of the raw data score is transferred to the other interacting partner. During the transfer if the edge is of inhibitory type, then the edge has a negative value on the interacting partner (Fig. 1). However when there is negative or positive feedback, score calculation cannot be solved with the above-explained procedure. Therefore, we applied a modified breadth-first search (BFS) algorithm to overcome this problem (see Algorithm 1). The algorithm had to iterate 10–15 times over the entire cyclic graph until the convergence of gene node scores was attained. Fig. 2 illustrates the general process diagram of our pathway node score flow algorithm.

The score flow calculation algorithm was implemented as a Cytoscape plug-in (ESI†) and it is publicly available. This plug-in can be used by following the directions given in the ESI.† Score calculation can be performed on custom (manually) generated pathways as well as KEGG pathways.

In addition, in order to evaluate the significance of scores obtained with the algorithm, we randomized input data several times and reran the algorithm to calculate new activity scores. Then *p*-values of pathway enrichment scores were calculated so that the consistencies of the final activity scores could be assessed and demonstrated.

Datasets

We applied the score flow algorithm to two datasets: ChIP-seq and expression array data sets from Estradiol-treated MCF7 breast cancer cells (GSE11352 and GSE19013) obtained simultaneously,[18,19] and the gene expression profile of the Colo741 cell line transfected by KRas-G12D or KRas-G12V mutant proteins (GSE12398).[20] The datasets were pre-processed as explained in Raw Data Preparation Documentation (ESI†). Analysis of Estradiol-treated MCF7 ChIP-seq was performed

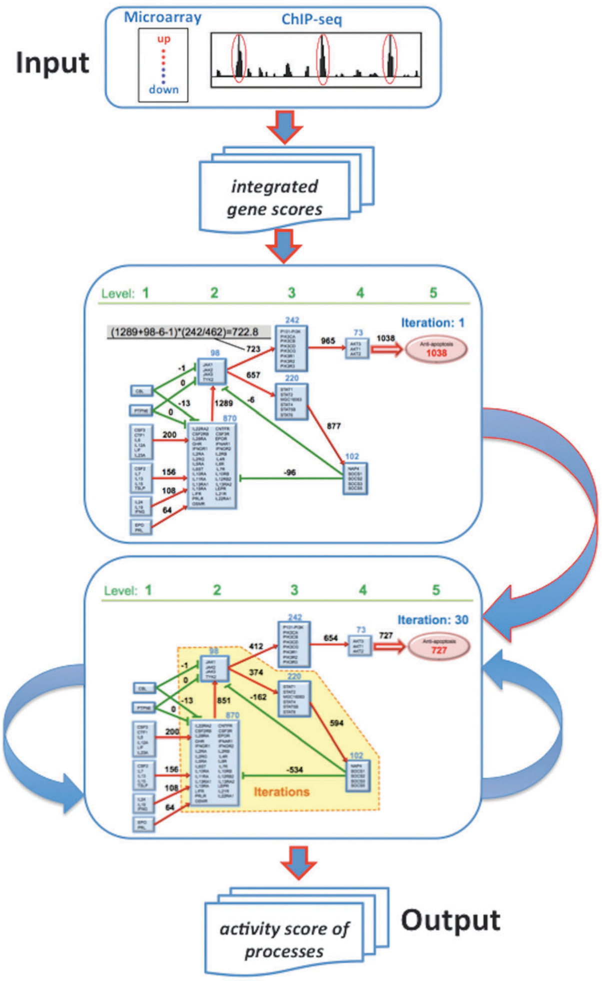


Fig. 2 Process diagram of the signal transduction score flow algorithm.

to determine the regulation role of the estrogen receptor transcription factor in the MCF7 breast cancer cell line.

We performed a row-wise normalization on raw array data. The gene expression experiments were provided as individual rank $I(x)$ scores for each node in the pathway.

The raw self-scores of nodes in the pathway were calculated by the product of individual gene scores extracted from ChIP-seq and expression array data by using the rank product technique.[21] The rank product method combines individual ranks of different biological measurements.

$$S(x) = \prod_{s=1}^{N} I_s(x), \qquad (1)$$

where $I_s(x)$ is the individual rank value of gene x coming from the data source s, and N is the total number of heterogeneous data sources. In order to integrate rank scores of genes extracted from individual gene expression and the ChIP-seq dataset, we applied eqn (1) and obtained the product of individual ranks, where $I_1(x)$ and $I_2(x)$ represent the individual ranking values of the microarray and ChIP-seq experiments for the gene x, respectively. $S(x)$ defines the raw *self-score* of gene x. If both of the ranks of x were missing, $S(x)$ value was set to 0. If gene x in a pathway has several *Entrez* gene identifiers, the mean of self-scores of these identifiers was calculated and the mean value was assigned as the self-score of x.

Pathway scoring algorithm

A pathway is converted into a directed graph $G = (V, E)$. A node in the graph represents a gene product or a target process linking the current signal to a final cellular activity. The edges represent the relations (*i.e.*, activation, inhibition) between the nodes. In G, let outAdj(x) denote the out-adjacency list of node x, that is, outAdj(x) = $\{y: (x,y) \varepsilon E\}$ and let inAdj(x) denote the in-adjacency list of node x, that is, inAdj(x) = $\{y: (y,x) \varepsilon E\}$.

If an edge (x,y) from node x to y is labelled activation, the total score of node x is then directly transferred. If edge (x,y) is inhibition, the total score of node x is transferred with a negative value as the score of node y. In order to consider the processing order of the genes in the actual pathway map, we performed score computations following the pathway nodes. For this purpose, the directed graph is converted into a cascade form by applying the multiple source breadth-first search (BFS) algorithm, which effectively propagates BFS levels starting from nodes of zero in-degree. Algorithm 1 displays the BFS-based algorithm used for this conversion. This cascade form enables us to solve the score convergence problems of some cyclic pathways.

Let $V_0, V_1, V_2, \ldots, V_{L-1}$ denote the levels of this cascade form of G, where V_0 denotes the set of nodes with zero in-degree. Note that V_l contains the nodes whose shortest path distance to the nodes in V_0 is equal to l, for l = 1, 2,. . ., $L - 1$. The proposed approach adopts an iterative process that updates the score of the nodes in a level-wise fashion. At each iteration of the algorithm, the nodes of the graph are processed in level order, *i.e.*, the nodes in level l are processed before the nodes in level $l + 1$. The processing of a node refers to transferring its score to the nodes in its out-adjacency list. At iteration k, a node x transfers its S^k_{out} to each node y in its out-adjacency list according to the following equation:

$$f^k(x,y) = \text{sign}(x,y) \times S^k_{\text{out}}(x) \times \frac{S(y)}{\sum\limits_{z \in \text{outAdj}(x)} S(z)} \qquad (2)$$

the out-score of node x is divided among the nodes in outAdj(x) according to the raw self-scores of these nodes. That is, nodes with small raw self-scores will get a small share of $S^k_{\text{out}}(x)$, compared to nodes having large self-scores. Note that the type of the edge from x to y is defined by sign(x,y), where sign(x,y) = 1 denotes activation and -1 denotes inhibition. Hence, the out-score of a node x is updated at each iteration k by summing up the out-score transfers from the nodes in its in-adjacency list as:

$$S^k_{\text{out}}(x) = S(x) + \sum_{z \in \text{inAdj}(x)} f^k(z,x).$$

Algorithm 2 describes the general steps of pathway scoring. In Algorithm 2, the for-loop inside the initialization for-loop computes the sum of the raw self-scores of the nodes in the out-adjacency of each node, which is equal to the denominator term of eqn (2). The scheme adopted in the while-loop of the score computation phase enables in-place accumulation of the contributions of the out-score of a given vertex x to the out-scores of the nodes in its adjacency list. Thus, the scheme avoids the need for maintaining a flow value (see eqn (2)) for each edge of graph G. The reason for the iterative approach is the cyclic signalling pathways; the out-scores of the nodes in a cycle need to be computed many times for the convergence of node scores in the cycle. For this purpose, we execute the while-loop until obtaining converged out-scores for all nodes in the graph. The convergence on the out-score of a node x is defined as:

$$S^{k}_{\text{out}}(x) - S^{k-1}_{\text{out}}(x) \leq \varepsilon$$

where ε is the error threshold for the convergence criteria and is set to 10^{-6}. Note that the proposed algorithm does not necessitate the expensive cycle-finding process in graph G. Instead, we performed passes over the entire graph level by level (as indicated in pseudo code) to compute the converged out-scores for all the nodes.

The graph G represents an overall pathway containing one or more final biological processes. In G, different biological processes are represented by a different subset of target nodes, where the distinguishing property of a target node is having zero out-degree. Let P denote the set of biological processes in a pathway represented by G and let $T(p)$ denote the subset of target nodes representing biological processes $p \varepsilon P$. The final activity score for a biological process p is computed by taking the sum of all possible biological processes leading to p in pathway G:

$$S_{\text{tot}}(p) = \sum_{t \in T(p)} S_{\text{out}}(t)$$

The BFS-based levelization/cascading algorithm runs in linear time in the size of the pathway graph G. That is, it is an $O(V + E)$-time algorithm. The while-loop of Algorithm 2 processes each vertex once, thus processing each edge only once. The initialization for-loop of Algorithm 2 also makes a single scan over all vertices and edges of G. So, Algorithm 2 can be considered as a linear-time algorithm if a constant number of iterations suffice for convergence.

Significance analysis of activity scores

In order to determine significance of final activity scores, we calculated the p-value of each activity score by applying input data randomization. For this purpose, the score ratio of a biological process z was defined as

$$\text{SR}(z) = \frac{S_{\text{tot}}(z)_{\text{control}}}{S_{\text{tot}}(z)_{\text{exp}}},$$

where $S_{\text{tot}}(z)_{\text{control}}$ and $S_{\text{tot}}(z)_{\text{exp}}$ are final activity scores of the process z obtained with original control and experiment data, respectively. The $\text{SR}(z)$ value is crucial to identify which experimental condition has more effect on the activity of a specific process.

Randomization of the input data was performed as follows:

1. For each node j in a pathway, randomly select a gene identifier k from the entire chip, then assign control and experiment self-scores of gene k to node j.
2. Run the score flow algorithm with these random data.
3. Compute the new ratio score of each process obtained with random data.
4. Repeat steps 1, 2 and 3 for M times.

The p-value $P(z)$ of process z was calculated by taking proportion of new ratio scores obtained with random data that yield bigger or smaller scores than original $\text{SR}(z)$

$$P(z) = \frac{1}{M}\sum_{n=1}^{M} C(\text{NR}(z)_n, \text{SR}(z)),$$

where $\text{NR}(z)_n$ represents the new ratio score obtained with randomized data at iteration n, and M is the total number of iterations performed for the randomization procedure and set to 10 000. The function C compares the values of $\text{SR}(z)$ and $\text{NR}(z)_n$ based on the magnitude of $\text{SR}(z)$ and returns either 0 or 1.

$$C(\text{NR}_n, \text{SR}) = \begin{cases} 1, & \text{if SR} < 1 \text{ and } \text{NR}_n \leq \text{SR} \\ 1, & \text{if SR} > 1 \text{ and } \text{NR}_n \geq \text{SR} \\ 0, & \text{otherwise} \end{cases}.$$

We set significance threshold of $P(z)$ to 0.1, hence the activity score of the process is assumed to be significant for an experiment if its p-value is less than this threshold.

Cytoscape plug-in

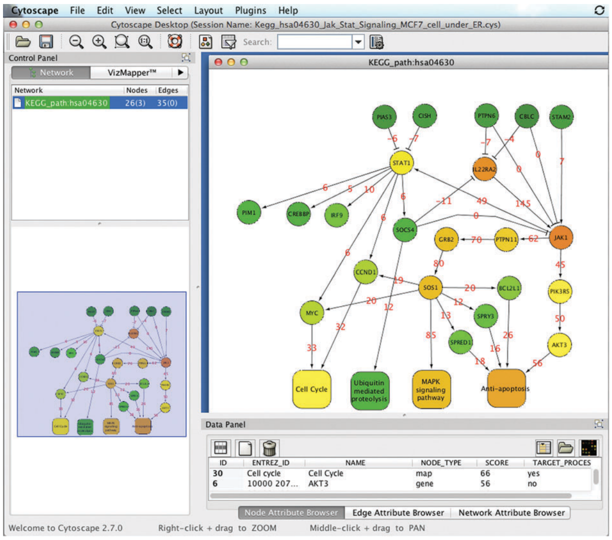
The score flow algorithm was implemented as Cytoscape plug-in to make the algorithm publicly available for molecular biologists. Cytoscape enables to visualize and to compute activity score of each target process.[13] In this environment, the user can load manually curated and custom generated pathways or upload KEGG pathways online. Each node in the graph should contain a unique *ID* (assigned by Cytoscape), *NAME* (process or gene name), *ENTREZ ID* (Entrez gene Id), *NODE TYPE* (defined by the use *i.e.* "gene" or "activity process"), *TARGET PROCESS* flag (for gene set to "no", for process "yes") and *SCORE* (initially set to zero, then calculated by the Score Flow algorithm). Circle and rectangle shapes represent the genes and target processes, respectively.

The plugin requires two input files, first the above-mentioned Cytoscape pathway file and the raw gene score file as tab delimited text file. Each line of the score file contains three attributes: Entrez id of gene, name and raw self-score. Upon uploading both files, the signal transduction score flow algorithm can be run over the pathway using Cytoscape plug-in's menu (Fig. 3). The calculated activity scores of genes and processes can be exported in a tab delimited text file. The installation and the step-by-step usage of the Cytoscape plug-in are given as ESI.†

Results

Application of the score flow algorithm to paired transcriptome and ChIP-seq data

Estrogen receptor (ER) is a hormonal transcription factor that plays important roles in breast cancer development. Upon binding to its ligand estrodiol, ER functions primarily through binding to the transcription regulatory regions of target genes containing the estrogen response element (ERE) consensus motifs.



Cytoscape Desktop (Session Name: Kegg_hsa04630_Jak_Stat_Signaling_MCF7_cell_under_ER.cys)

Network	Nodes	Edges
KEGG_path:hsa04630	26(3)	35(0)

ID	ENTREZ_ID	NAME	NODE_TYPE	SCORE	TARGET_PROCES
30	Cell cycle	Cell Cycle	map	66	yes
6	10000 207…	AKT3	gene	56	no

Welcome to Cytoscape 2.7.0   Right-click + drag to ZOOM   Middle-click + drag to PAN

Fig. 3 Cytoscape representation of the Jak-STAT pathway scored with Estradiol-treated MCF7 cells data. The circles and rectangles represent the genes and processes, respectively. The color intensity of nodes from green to red represents the final gene enrichment or process activity scores after the score flow algorithm is applied.

The set of experimental data that we analysed (from the NCBI-GEO database; see Materials) was paired Estradiol E2-treated MCF7 breast adenocarcinoma cells for ChIP-seq and expression array data. After initial data integration and the rank analysis of the raw ChIP-seq data, 1900 putative peak regions neighbouring 485 genes from the array data were identified. The raw gene scores were applied on KEGG pathways by using Cytoscape plug-in. The algorithm had to run 10–15 times over the entire cyclic graph until it verified the convergence threshold. The activity scores of significant signalling pathways in MCF7 cells treated with Estradiol are presented in Table 1. ER receptor activation in the estrogen-receptor positive breast cancer cell line MCF7 was shown to be clearly differentially activating the cellular processes involved in *Apoptosis* in many cellular pathways. We also observed an increased activity in *Proliferation*, *Survival* and *Cell cycle* end-cellular processes (Table 1). This is in correlation with the proliferative effect of E2 on MCF7 cells as also demonstrated by previous studies.[22–24] E2 carcinogenesis involves two distinct pathways: oxidative metabolism of estrogen through the Catechol Pathway and small GDP binding proteins with MAPK pathway activation.[25] Catechol Pathway leads to apoptosis and MAPK signalling leads to survival and cell proliferation. Our data analysis clearly demonstrates the action of these two mechanisms in E2 treated MCF7 cells when compared to untreated control cells.

In silico gene knockout operation on the PI3K/AKT pathway

Proteins residing in central positions in the network topology and having many interactions with other proteins can be considered hub-proteins. There are some proteins which act as hubs, collecting high scores in our method as well. The scores of target processes in a signalling cascade would be affected by the deletion of such hub-nodes. With the aim of determining the weights of such hub-nodes, we simply deleted the hub-gene node and the in- and out-edges of that node from the pathway and ran the algorithm again. The scored pathway as a result of this gene knock-out operation was compared to the original pathway's scores. The significance of the final activity scores was evaluated by randomization of input data. After randomization, the *p*-value of the final activity scores in knockout pathways was still consistent. Therefore, we were able to assess the critical role of hub-proteins in high score collecting nodes in a pathway leading to a cellular end process.

The *in silico* gene knock-out was applied on a PI3K/AKT pathway, which was manually constructed using literature information with Cytoscape (ESI† *.cys* file). The pathway contains 83 genes, six target process nodes (*DNA repair*, *Translation*, *Migration*, *Angiogenesis*, *Apoptosis*, and *Cell Cycle*), and 160 edges (105 activation and 55 inhibition) (ESI†, Fig. S1 and Cytoscape Files).

In the PI3K/AKT pathway, there are two significant hub nodes: serine/threonine kinase *Akt* and tumour suppressor gene p53. *Akt* promotes cell survival and had been shown to be constitutively expressed in a variety of human tumours.[26–28] p53 is an important hub-protein in cell signalling such as apoptosis, cell cycle and DNA repair. Therefore, we decided to knock out the p53 protein node from the native pathway. After the *in silico* p53 knockout operation, the new pathway was used during the score computation. The wild-type and *in silico* p53-knocked out PI3K/Akt pathways were analysed with an expression array dataset from adenocarcinoma cell line Colo741 carrying oncogenic mutant form of KRas (G12D) and the wild-type experiment control.[20,29] Our score flow algorithm provided comparative activity scores of original and knockout pathways (Fig. 4).

As expected, the final activity score of the *Apoptosis* process was significantly reduced in the p53 knockout pathway (Fig. 4C, second row), confirming p53 as the key regulator of the *Apoptosis* process (Fig. 4). In parallel the *Cell Cycle* process was scored with increased activity. With the microarray data we used, in which Ras mutations were studied in a BRAF mutated context, the G12D mutation was also shown to be associated with processes like cell cycle and apoptosis.[26] In addition, no change is observed in the *DNA repair* process because these cells were not challenged to induce their DNA repair mechanism.

Discussion

The present study describes the novel signal transduction score flow algorithm that not only computes the experimental data-driven enrichment of the gene nodes and connecting edges of a given cellular pathway but also provides the activity scores for all target biological processes.

Due to the detailed node level biochemical data availability, metabolic pathways were often dynamically modelled with ordinary differential equations.[30] Additionally flux balance analysis using boolean expressions incorporated with ordinary differential equations was used to simulate metabolic regulatory pathways in an iterative approach similar to our algorithm.[31] However lack of protein node stoichiometry knowledge in cell signalling pathways is a major drawback in dynamic modelling of the cell signalling networks using large scale omics data.

Table 1 Significant activity scores of signaling pathways for control and estradiol (E2)-treated samples of MCF7 cells

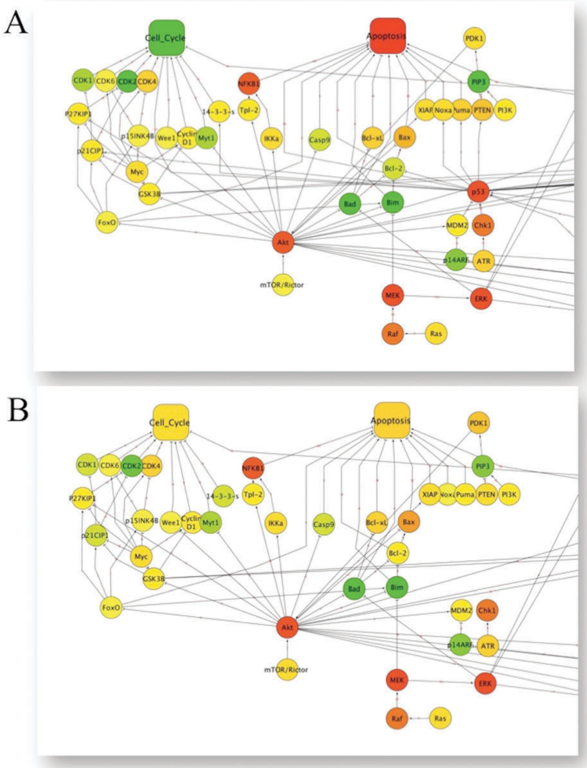
		Activity score of process		
KEGG pathway	Final process	Control	ER	*p*-Value
Acute myeloid leukaemia (hsa05221)	Proliferation	49	**929**	0.034
Alzheimer's disease (hsa05010)	Apoptosis	54	**552**	0.037
Apoptosis (hsa04210)	Apoptosis	168	**1354**	0.050
	Degradation	56	**653**	0.023
Chronic myeloid leukaemia (hsa05220)	Proliferation	**137**	68	0.004
Endometrial cancer (hsa05213)	Cell growth	**131**	63	0.006
	Proliferation	**163**	94	0.004
ErbB signalling (hsa04012)	Degradation	**6**	5	0.022
Focal adhesion (hsa04510)	Apoptosis	37	**161**	0.053
	Cell motility/FA formation	38	**172**	0.052
	FA-turnover	27	**708**	0.016
	Proliferation	66	**257**	0.045
	Survival	23	**165**	0.025
Glioma (hsa05214)	Cell growth	118	**276**	0.035
Jak-Stat signalling (hsa04630)	Anti-apoptosis	41	**118**	0.048
	Cell cycle	23	**66**	0.048
	MAPK	24	**86**	0.029
	Ubiquitin mediated proteolysis	6	**12**	0.033
MAPK signalling (hsa04010)	Apoptosis	37	**95**	0.051
	Cell cycle	62	**1122**	0.022
	P53 signalling	22	**46**	0.062
	Proliferation	156	**1766**	0.033
	Wnt signaling	**13**	7	0.006
Melanoma (hsa05218)	Survival	31	**131**	0.034
Neurotropic signalling (hsa04722)	Plasticity	23	**358**	0.018
	Regulation of actin cytoskeleton	25	**359**	0.037
Non-small cell lung cancer (hsa05223)	Proliferation	**160**	91	0.004
Pathways in cancer (hsa05200)	Block of differentiation	41	**1052**	0.009
	Proliferation	239	**1216**	0.099
Regulation of actin cytoskeleton (hsa04810)	Adherent junction	**26**	13	0.007
	MAPK	109	**2296**	0.017
Renal cell carcinoma (hsa05211)	Cell-junction, Migration, Invasion	147	**627**	0.055
	Proliferation	**149**	80	0.004
Thyroid cancer (hsa05216)	Proliferation	151	**203**	0.054
	Survival	112	**165**	0.043

In general, there are two approaches used to interpret the large-scale experimental data after pre-processing. Usually the scored or ranked gene categories based on the experimental input are mapped on static cellular signalling pathways. According to the pathway topology, the experiment-related genes or gene sets are selected as differentially expressed genes correlated with the observed phenotype.[32] There are also studies that use the pre-processed or raw high-throughput data to infer molecular pathways related to the experiment. Therefore the novel score flow algorithm presents a heuristic approach that uses the gene score partition as a solution for protein node stoichiometry during dynamic scoring of the pathway of interest. Compared to other tools, our algorithm performs a simulation of cell signalling flow on the cyclic pathway topology rather than assigning static gene scores to pathway nodes.

The data driven enrichment of pathway nodes and edges can be computed on network topology using differentially expressed gene data by various tools. In general, these tools calculate a single pathway impact score or a list of enriched genes from that pathway. SPIA and GSEA tools are the most similar publicly available tools comparable to our algorithm. Signalling Pathway Impact Analysis (SPIA) tool estimates the impact of experimental perturbations on pathways and it is implemented in R.[33] SPIA aims to identify the enriched pathways using differentially expressed genes and pathway topological information. When compared to our method, SPIA provides only a general behaviour of the pathway *i.e.*, activation or inhibition without activity scores for pathway nodes, edges and activity processes (ESI†, Table S1). Besides, SPIA does not provide a visual graph representation of the pathways whereas our algorithm can be applied on any hand-curated pathway with its Cytoscape plug-in.

Gene Set Enrichment Analysis (GSEA) was also compared with our algorithm on the Colo741-KRas dataset. Based on the GSEA results, only Reactome *Apoptosis* gene set was significantly enriched (ESI†, Tables S2 and S3).

There are also recent studies, which exploit high-throughput data considering the flow of the cell signalling within a pathway. In one of these studies, a signalling network is represented by an electrical circuit, where interactions are resistors, proteins are interconnecting junctions and the information flow analysis identifies hub-proteins in the interactome networks.[34] Although the information score flow approach seems to be similar to our signal transduction flow algorithm, we model gene signals as the integrated scores and score flow is transferred into child nodes based on their edge types and self-score states. Cyclic feedback loops are also not considered in previous studies.[32,34] When the score flow algorithm is compared to similar tools, the most significant difference is the stoichiometric concentration based score partition during the flow of the signal and the implementation of the cyclic feedback loops in the pathways.



C

Biological Process	wt	G12D
Apoptosis	40	38
Apoptosis p53-KO	9	8
Angiogenesis	21	21
Angiogenesis p53-KO	23	23
DNA repair	106	110
DNA repair p53-KO	104	108
Cell Cycle	0	0
Cell Cycle p53-KO	7	7
Migration	50	59
Migration p53-KO	51	51

Fig. 4 Enrichment scores of apoptosis and cell cycle processes in the manually curated PI3K/Akt pathway with KRas (G12D) mutation data (A) and *in silico* p53-knockout (p53-KO) enrichment (B) with Colo741 data. Activity scores of wildtype (wt) *versus* G12D mutation are indicated in the table (C). Down-regulated and up-regulated genes or processes are represented in color tones of green and red, respectively.

In this study we also present the application of the tool on datasets with complementary transcriptome and ChIP-seq data. The results that we observed with the signal transduction score flow algorithm were in correlation with the literature data. Moreover, gene-level and global impacts of single or multiple gene knockouts were examined by *in silico* knockout analysis. The algorithm allows visualization of the impact of deleting or inhibiting a protein node, not only on the first level downstream protein but also related signalling pathways and the various target cellular processes. Thus, it is possible to visualize the side effects of inhibiting one protein, since its influence on target processes other than the expected ones will be demonstrated as well. It would be of great value to be able to predict the drug combination that could not only increase apoptosis in cancer cells but also decrease survival and cell cycle. This *in silico* tool may suggest hypotheses about how a drug of interest acts on the molecular cellular pathways, and it may predict the synergistic effects of different inhibitors.

Algorithm 1

BFS-based algorithm for levelizing graph *G*.

```
Input:
Directed graph G stored in-adjacency and out-adjacency
list format
outAdj(x): out-adjacency list of node x

Initialization:
for each vertex x ∈ V do
        if in-degree(x) = 0 then
                color(x) = BLACK
                d(x) = 0
                ENQUEUE(Q,x)
        else
                color(x) = WHITE

Levelization:
while Q ≠ ∅ do
        x = DEQUEUE(Q)
        for each vertex y ∈ outAdj(x) do
                if color(y) = WHITE then
                        color(y) = BLACK
                        d(y) = d(x) + 1
                        $V_{d(y)}$ = $V_{d(y)}$ U {y}
                        ENQUEUE(Q,y)
return ($V_0$,$V_1$,$V_2$,...,$V_{L-1}$)
```

Algorithm 2

Pathway scoring.

```
Input:
Directed graph G stored in in-adjacency and out-
adjacency list format
Score: indicates self-score of each node given by our
method
outScore: contains out-score of each node
outAdj(x): out-adjacency list of node x
sign : keeps edge types: activation (1) or inhibition (-
1)
P = {p}: set of biological processes
T (p): set of target nodes representing process P in G
Levelization information $V_0$,$V_1$,$V_2$,...,$V_{L-1}$ obtained by
running Algorithm 1.

Initialization:
for each vertex x ∈ V do
        outScore(x) = Score(x)
        totOutSelfScore(x) = 0
        for each vertex y ∈ outAdj(x) do
                totOutSelfScore(x) = totOutSelfScore(x)
        + Score(y)

Score Computation:
while not converged do
 for each level i = 0,1,2,...,L-1 do
  for each vertex x ∈ $V_i$ do
   for each vertex y ∈ outAdj(x) do
   outScore(y) = outScore(y) + sign(x,y) * outScore(x) *
[Score(y)/totOutSelfScore(x)]

Output:
for each biological process p ∈ P do
        TotalScore(p) = 0
        for each target node t ∈ T (p) do
                TotalScore(p) = TotalScore(p) +
        outScore(t)

return {TotalScore(p)}$_{p \in P}$
```

Acknowledgements

ZI was supported by The Scientific and Technological Research Council of Turkey (TUBITAK). We thank Ms R. Nelson for editing the English of the final version of our manuscript.

References

1 J. J. Hornberg, F. J. Bruggeman, H. V. Westerhoff and J. Lankelma, *Biosystems*, 2006, **83**, 81–90.
2 P. Brazhnik, A. de la Fuente and P. Mendes, *Trends Biotechnol.*, 2002, **20**, 467–472.
3 E. Demir, O. Babur, U. Dogrusoz, A. Gursoy, G. Nisanci, R. Cetin-Atalay and M. Ozturk, *Bioinformatics*, 2002, **18**, 996–1003.
4 E. Demir, O. Babur, U. Dogrusoz, A. Gursoy, A. Ayaz, G. Gulesir, G. Nisanci and R. Cetin-Atalay, *Bioinformatics*, 2004, **20**, 349–356.
5 M. Kanehisa, S. Goto, M. Furumichi, M. Tanabe and M. Hirakawa, *Nucleic Acids Res.*, 2010, **38**, D355–D360.
6 C. Stark, B. J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz and M. Tyers, *Nucleic Acids Res.*, 2006, **34**, D535–D539.
7 L. Matthews, G. Gopinath, M. Gillespie, M. Caudy, D. Croft, B. de Bono, P. Garapati, J. Hemish, H. Hermjakob, B. Jassal, A. Kanapin, S. Lewis, S. Mahajan, B. May, E. Schmidt, I. Vastrik, G. Wu, E. Birney, L. Stein and P. D'Eustachio, *Nucleic Acids Res.*, 2009, **37**, D619–D622.
8 F. Al-Shahrour, R. Diaz-Uriarte and J. Dopazo, *Bioinformatics*, 2004, **20**, 578–580.
9 B. Mlecnik, M. Scheideler, H. Hackl, J. Hartler, F. Sanchez-Cabo and Z. Trajanoski, *Nucleic Acids Res.*, 2005, **33**, W633–W637.
10 K. D. Dahlquist, N. Salomonis, K. Vranizan, S. C. Lawlor and B. R. Conklin, *Nat. Genet.*, 2002, **31**, 19–20.
11 N. Goffard and G. Weiller, *Nucleic Acids Res.*, 2007, **35**, W176–W181.
12 A. Nikitin, S. Egorov, N. Daraselia and I. Mazo, *Bioinformatics*, 2003, **19**, 2155–2157.
13 P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski and T. Ideker, *Genome Res.*, 2003, **13**, 2498–2504.
14 Z. Hu, J. Mellor, J. Wu, T. Yamada, D. Holloway and C. Delisi, *Nucleic Acids Res.*, 2005, **33**, W352–W357.
15 H. J. Chung, C. H. Park, M. R. Han, S. Lee, J. H. Ohn, J. Kim and J. H. Kim, *Nucleic Acids Res.*, 2005, **33**, W621–W626.
16 Z. Isik, V. Atalay and R. Çetin-Atalay, *J. Mach. Learn. Res. – Proc. Track*, 2010, **8**, 44–54.
17 H. Ji, H. Jiang, W. Ma, D. S. Johnson, R. M. Myers and W. H. Wong, *Nat. Biotechnol.*, 2008, **26**, 1293–1300.
18 M. Hu, J. Yu, J. M. Taylor, A. M. Chinnaiyan and Z. S. Qin, *Nucleic Acids Res.*, 2010, **38**, 2154–2167.
19 C. Y. Lin, V. B. Vega, J. S. Thomsen, T. Zhang, S. L. Kong, M. Xie, K. P. Chiu, L. Lipovich, D. H. Barnett, F. Stossi, A. Yeo, J. George, V. A. Kuznetsov, Y. K. Lee, T. H. Charn, N. Palanisamy, L. D. Miller, E. Cheung, B. S. Katzenellenbogen, Y. Ruan, G. Bourque, C. L. Wei and E. T. Liu, *PLoS Genet.*, 2007, **3**, e87.
20 M. Monticone, E. Biollo, M. Maffei, A. Donadini, F. Romeo, C. T. Storlazzi, W. Giaretti and P. Castagnola, *Mol. Cancer*, 2008, **7**, 92.
21 R. Breitling, P. Armengaud, A. Amtmann and P. Herzyk, *FEBS Lett.*, 2004, **573**, 83–92.
22 R. X. Song, Z. Zhang, Y. Chen, Y. Bao and R. J. Santen, *Endocrinology*, 2007, **148**, 4091–4101.
23 H. Seeger, D. Wallwiener, E. Kraemer and A. O. Mueck, *Maturitas*, 2006, **54**, 72–77.
24 C. Martinez-Campa, P. Casado, R. Rodriguez, P. Zuazua, J. M. Garcia-Pedrero, P. S. Lazo and S. Ramos, *Breast Cancer Res. Treat.*, 2006, **98**, 81–89.
25 J. D. Yager and N. E. Davidson, *N. Engl. J. Med.*, 2006, **354**, 270–282.
26 J. A. Engelman, *Nat. Rev. Cancer*, 2009, **9**, 550–562.
27 P. Liu, H. Cheng, T. M. Roberts and J. J. Zhao, *Nat. Rev. Drug Discovery*, 2009, **8**, 627–644.
28 E. Tokunaga, E. Oki, A. Egashira, N. Sadanaga, M. Morita, Y. Kakeji and Y. Maehara, *Curr. Cancer Drug Targets*, 2008, **8**, 27–36.
29 M. A. White, C. Nicolette, A. Minden, A. Polverino, L. Van Aelst, M. Karin and M. H. Wigler, *Cell*, 1995, **80**, 533–541.
30 J. C. Sible and J. J. Tyson, *Methods (San Diego, Calif.)*, 2007, **41**, 238–247.
31 M. Terzer, N. D. Maynard, M. W. Covert and J. Stelling, *Wiley Interdiscip. Rev.: Syst. Biol. Med.*, 2009, **1**, 285–297.
32 S. Efroni, C. F. Schaefer and K. H. Buetow, *PLoS One*, 2007, **2**, e425.
33 A. L. Tarca, S. Draghici, P. Khatri, S. S. Hassan, P. Mittal, J. S. Kim, C. J. Kim, J. P. Kusanovic and R. Romero, *Bioinformatics*, 2009, **25**, 75–82.
34 P. V. Missiuro, K. Liu, L. Zou, B. C. Ross, G. Zhao, J. S. Liu and H. Ge, *PLoS Comput. Biol.*, 2009, **5**, e1000350.