JMLR: Workshop and Conference Proceedings 8: 44-54 Machine Learning in Systems Biology

Evaluation of Signaling Cascades Based on the Weights from Microarray and ChIP-seq Data

Zerrin Isik ZERRIN.SOKMEN@CENG.METU.EDU.TR
Volkan Atalay VOLKAN@CENG.METU.EDU.TR
Department of Computer Engineering
Middle East Technical University
Ankara, TURKEY

Rengul Cetin-Atalay RENGUL@BILKENT.EDU.TR
Department of Molecular Biology and Genetics
Bilkent University
Ankara, TURKEY

Editor: Sašo Džeroski, Pierre Geurts, and Juho Rousu

Abstract

In this study, we combined the ChIP-seq and the transcriptome data and integrated these data into signaling cascades. Integration was realized through a framework based on data- and model-driven hybrid approach. An enrichment model was constructed to evaluate signaling cascades which resulted in specific cellular processes. We used ChIP-seq and microarray data from public databases which were obtained from HeLa cells under oxidative stress having similar experimental setups. Both ChIP-seq and array data were analyzed by percentile ranking for the sake of simultaneous data integration on specific genes. Signaling cascades from KEGG pathway database were subsequently scored by taking sum of the individual scores of the genes involved within the cascade. This scoring information is then transferred to en route of the signaling cascade to form the final score. Signaling cascade model based framework that we describe in this study is a novel approach which calculates scores for the target process of the analyzed signaling cascade, rather than assigning scores to gene product nodes.

Keywords: evaluation of signaling cascades, chip-seq, gene expression

1. Introduction

Large scale experiments enable researchers to access the transcriptome information related to the state of several thousands of genes under a particular experimental condition. Traditional analysis methods for microarray data output a list of significant genes relevant to the performed experiments. In order to associate the list of genes to a specific cellular process secondary tools and databases are used. Therefore, research focuses on the analysis of the biological pathways rather than individual genes (Cordero et al., 2008). Several gene prioritization methods determine the similarity between candidate genes and genes known to play a role in defined biological processes or diseases (Aerts et al., 2006; De-Bie et al., 2007; Lopez-Bigas and Ouzounis, 2004; Kent et al., 2005). Therefore, it is clear that available data from multiple sources (e.g. Gene Ontology annotations, protein domain databases, biological networks, published literature, gene expression data etc.) would enrich the anal-

ysis. A variety of methods have been developed to analyze and visualize microarray data in the context of biological networks (Al-Shahrour et al., 2004; Dahlquist et al., 2002; Chung et al., 2005; Mlecnik et al., 2005; Goffard and Weiller, 2007). These methods identify significant functional terms or biological pathways by applying several statistical significance tests. They also overlay gene expression data into molecular pathways to reveal experiment specific gene regulations (Ingenuity; Nikitin et al., 2003). Additionally, these tools apply graph theory and calculate significance scores on the pathways. Nevertheless, they rely on the initially identified differentially expressed significant gene list.

Chromatin ImmunoPrecipitation (ChIP) combined with genome re-sequencing (ChIP-seq) technology provides protein DNA interaction data. Transcription factors (TFs) bind to specific DNA sequences and turn transcription of target genes on or off. ChIP-seq technology is expected to be popularly used for the analysis of gene expression signatures, similarly to microarray technology. ChIP-seq experiments and computational analysis methods in literature have been at their initial stages (Rosenbluth et al., 2008). Although there is a few number of raw data analysis tools for ChIP-seq data, further gene list-molecular process enrichment methods should be considered. In this study we integrated ChIP-seq and gene expression experiments using publicly available ChIP-seq and microarray data from HeLa cells under oxidative stress having similar experimental conditions. We constructed an enriched model to evaluate the signaling cascades under the control of specific biological process (Figure 1).

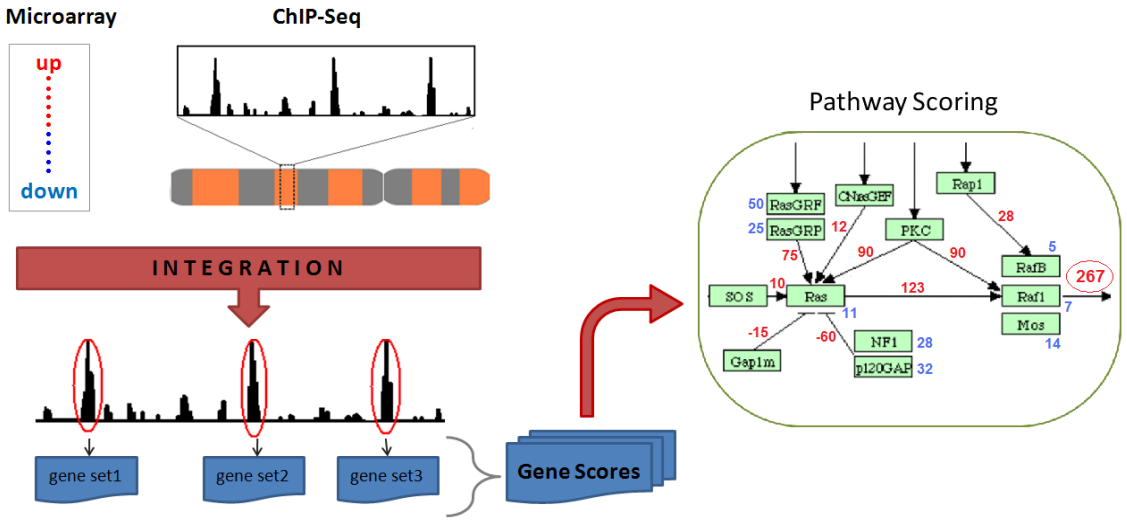

Figure 1: Process diagram of the proposed method. The integration stage combined ChIP-seq and transcriptome data to obtain scores of genes related with a specific transcription factor. In the next stage, signaling cascades activated by the TF were identified by exploring scores of each signaling cascade.

2. System and Methods

2.1 Data Processing

Experimental data of this study was obtained from NCBI GEO database. We selected ChIP-seq and microarray data from GEO datasets (NCBI GEO id: GSE14283 and GSE4301). The ChIP-seq data by Kang et al. aimed to identify transcription regulation role of OCT1 TF on HeLa cells, cervical cancer derived immortal cell line, under oxidative stress (Kang et al., 2009). Raw ChIP-seq data of OCT1 TF contained approximately 3.8 million reads. Initially, significant peak regions in raw data were explored by applying peak detection method, CisGenome tool (Ji et al., 2008). Peak detection method scans the genome with a sliding window (width=100, slide=25) and identifies regions with read counts greater than 10 reads for significant binding regions. We obtained 5080 putative peak regions over entire human genome. In order to compute significance of each peak region, we set a percentile rank value for each peak region by considering total number of reads involved in that region.

$$ReadRank(r) = \frac{cf_l + 0.5(f_r)}{T} \tag{1}$$

where cf_l was the cumulative frequency for all scores lower than the score of the peak region r, f_r was the frequency of the score of peak region r, and T was the total number of peak regions. $ReadRank(r)$ score ranged from 0 to 1. After identification of OCT1 high quality peak regions, we mapped to the TSS of the genes within a region ±10000 bp. Total number of neighboring genes associated with high quality peak regions was 260.

The microarray data used in the study was also obtained from HeLa cells under oxidative stress condition (Murray expression data) (Murray et al., 2004). We calculated fold-change (gene expression log ratio) of two channels for control and oxidative stress experiments.

$$FoldChange(x) = log_2(\frac{\overline{ch2_x}}{\overline{ch1_x}}) \tag{2}$$

where $\overline{ch1_x}$ and $\overline{ch2_x}$ represented the mean value of channel 1 and channel 2 of gene x, respectively. We observed that half of the genes had very low fold changes (less than 0.2 fold). In order to assign a rank value of the gene expression, we applied Equation 3 which involved similar computation to that of $ReadRank$.

$$ExpRank(x) = \frac{cf_l + 0.5(f_x)}{T} \tag{3}$$

where cf_l was the cumulative frequency for all fold-change values lower than the fold-change value of the gene x, f_x was the frequency of the fold-change value of gene x, and T was the total number of genes in chip. If the magnitude of fold change was very close to 0, the rank value was close to 0. Otherwise, rank value of a gene varied between 0 and 1 according to magnitude of its fold change.

2.2 Integration of ChIP-seq and Microarray Data

The gene set extracted from OCT1 ChIP-seq data and Murray expression data for a gene were associated by taking their weighted linear combinations.

$$Score(x) = c_{chip}ReadRank(x) + c_{exp}ExpRank(x) \tag{4}$$

where $ReadRank(x)$ was the ChIP-seq read rank value of gene x given by Equation 1, $ExpRank(x)$ was the expression rank value of gene x indicated by Equation 3, and c_{chip} and c_{exp} were the coefficients of two data sources. In order to consider their effects equally, 0.5 was assigned to both c_{chip} and c_{exp}.

2.3 Scoring of Signaling Cascades

In order to assign scores to signaling cascades which control biological process; we used KEGG pathway as the model driven cell signaling scoring approach. KEGG pathways were converted into the graph structures by using KGML files. A node in the graph represented gene product, chemical compound or target process linking current signal to other KEGG pathways. The edges represented the relations (i.e. activation, inhibition) between the nodes. Each cell signaling pathway cascade was enumerated from a specific KGML file that leads to biological process of the selected pathway. If the edge between two nodes is labeled as activation, the total score of that node was transferred directly. If the edge is inhibition, the total node score was transferred with a negative value (Figure 2). If a gene, involved in a pathway, had no score, the value of $Score(x)$ was set to zero. In order to consider processing order of the genes in actual pathway map, we performed score computations following the signaling cascade nodes. Total score of a signaling cascade was computed by applying score flow mechanism up to the target node: biological process. Algorithm 1 describes general steps of the biological score computation.

Two different score transfer approaches were applied between the neighboring nodes. The first one was called "direct score transfer" that adopted the direct score transferring from a parent node to its children by edge type: 1 for activation and -1 for inhibition. The second approach, "partitioned score transfer" divides the score of effector (parent) node on the children according to the score of the child node. In other words, each child node received a partitioned score from the parents based on its self $Score(x)$. Therefore the nodes with small self scores did not share the same parent score with the nodes of high scores. Hence the outcoming score of a parent node was distributed to all of its children according to the magnitude of their self $Score(x)$ (see Figure 2-C and 2-D).

Total score of a signaling cascade P was computed by taking the sum of all possible biological processes leading to P which is the target biological process linking current pathway to the other pathways in KEGG database.

$$TotalScore(P) = \sum_{s=1}^{N} outputScore[s, s] \tag{5}$$

where $outputScore[s, s]$ was total path score of the biological process s, N was the total number of the same biological processes leading to P. The average score of the signaling cascade P was computed to discover oxidative stress effected signaling cascades and assign a significance score to them.

$$AverageEnrichmentScore(P) = \frac{TotalScore(P)}{N} \tag{6}$$

where $TotalScore(P)$ was the total score of the signaling cascade P, N was the total number of genes involved in that signaling cascade. The current source code of the framework is available upon request.

Algorithm 1 : Computing Score of Signaling Cascades

Input:

Graph P, has *nodes* array and A matrix
$nodes[m]$: keeps the id of each node in P
$A[m, m]$: adjacency matrix of P (neighboring relation could be an activation (1), inhibition (-1) or no relation (0), diagonal of A was set to 0)
$Score[m]$: indicates self score of each node given by our method
$outputScore[m, m]$: contains output score of each node
$level[m, m]$: hash table to store the level of each node according to the visiting order in BFS
highestLevel: the highest level of P according to visiting order in BFS
partition: a boolean parameter defines the output score transfer method. If it is set to 1, partitioned score transfer method is used. Otherwise direct score transfer method is used.

Initialization:

startNodes = nodes having only outcoming edges (no incoming edge)
{Run Breadth-First Search (BFS) algorithm to identify neighboring order of the nodes}
(level, highestLevel) = BFS (A, nodes, startNodes)
Set entire matrix of $outputScore[m, m] = 0$

Score Computation:

for $i = 1$ to *highestLevel* **do**
  $levelMembers = level[i]$ {keeps the members in the level i}
  for $t = 1$ to $length(levelMembers)$ **do**

    $j = levelMember[t]$ {set the node id which is the member of level i and at the position t}
    $outputScore[j, j] = Score[j]$

    for $s = 1$ to $length(nodes)$ **do**
      $outputScore[j, j] = outputScore[j, j] + outputScore[s, j]$

    if $outputScore[j, j] < 0$ **then**
      $outputScore[j, j] = 0$ {negative score is originated by only inhibition edges}

    if $partition == 0$ **then**
      for $k = 1$ to $length(nodes)$ **do**
        if $A[j, k]! = 0$ **then**
          $outputScore[j, k] = A[j, k] * outputScore[j, j]$

    if $partition == 1$ **then**
      $totalScore = 0$
      for $k = 1$ to $length(nodes)$ **do**
        if $A[j, k]! = 0$ **then**
          $totalScore = totalScore + Score[k]$ {compute total score of the neighbors for node j}
      for $k = 1$ to $length(nodes)$ **do**
        if $A[j, k]! = 0$ **then**
          $outputScore[j, k] = A[j, k] * (outputScore[j, j] * Score[k]/totalScore)$

Output: Diagonal entries in *outputScore* matrix provides the outcoming scores of target biological processes in graph P.

3. Results and Discussion

KEGG pathways were used to evaluate our model driven pathway scoring approach by integrating rank scores from ChIP-seq and array data. Mapping gene scores onto pathways provided the determination of specific regulation motifs driving different responses in several signaling cascades. An example about this pathway enrichment is illustrated in Figure 2. In this figure Jak-STAT signaling cascade started with the initial activation nodes on the left and ended with the target node: *Apoptosis* biological process. The gene scores were assigned to the nodes and the scores were reflected to en route of the signaling cascade by applying direct (Figure 2-A, 2-B) and partitioned (Figure 2-A, 2-B) score transfer approaches under control and oxidative stress conditions. Finally, the score of the target biological process *Apoptosis* under the control of Jak-STAT signaling cascade was calculated. The total score for *Apoptosis* biological process computed by oxidative stress expression data (Figure 2-A, 2-C) was higher than that of the control (Figure 2-B, 2-D).

In order to highlight the novelties of our model driven framework for transcriptome data analysis, we also applied *kegArray* tool (Kanehisa et al., 2006) to Murray oxidative stress data over Jak-STAT signaling cascade (Figure 3). Several tools, similar to kegArray, map expression data over pathways; however, they could not assign a score to the target biological process. On the other hand, our method provided a quantitative measure to evaluate biological activity of a pathway in a specific process such as *Apoptosis* or *Cell cycle*.

We applied our framework to 4 KEGG pathways: Jak-STAT (hsa04630), Apoptosis (hsa04210), TGF-β (hsa04350), and MAPK (hsa04010) signaling pathways. These pathways had several target cellular processes. When the average scores of target biological processes were compared by considering the results of the two score transfer approaches, Apoptosis biological process in Jak-STAT signaling cascade produced a score of 224.73 obtained with direct score transfer and a score of 45.69 with partitioned score transfer approach under the oxidative stress condition. For all KEGG pathways direct score transfer approach result in higher scores for the oxidative stress compared to the control (Table 1). On the other hand, when the partitioned score transfer approach was applied on the same KEGG pathways, discriminative scores were obtained with both the control and the oxidative stress data (Table 1) which is more in correlation with the cellular machinery. Cells respond to the same signal by activating different cellular processes with different extends. In Table 1, when the score transfer was done by partitioned scoring method, *Apoptosis* got higher scores and *MAPK* pathway got lower scores when compared to normal. This was an expected result. Oxidative stress damages the biomolecules of the cell, therefore program cell death (*Apoptosis*) pathway was scored to be more active then *MAPK* cell proliferation (survival) pathway. The response of a cell to a condition either normal or stressed is expected to be differential; therefore as a result of our analysis some of the target processes were activated whereas others were down-regulated. For example, *Cell cycle* biological process always provided higher scores with the control microarray data for all pathways whereas *Apoptosis*, *Proliferation and differentiation*, and *MAPK signaling* gave differential scores, which is not surprising.

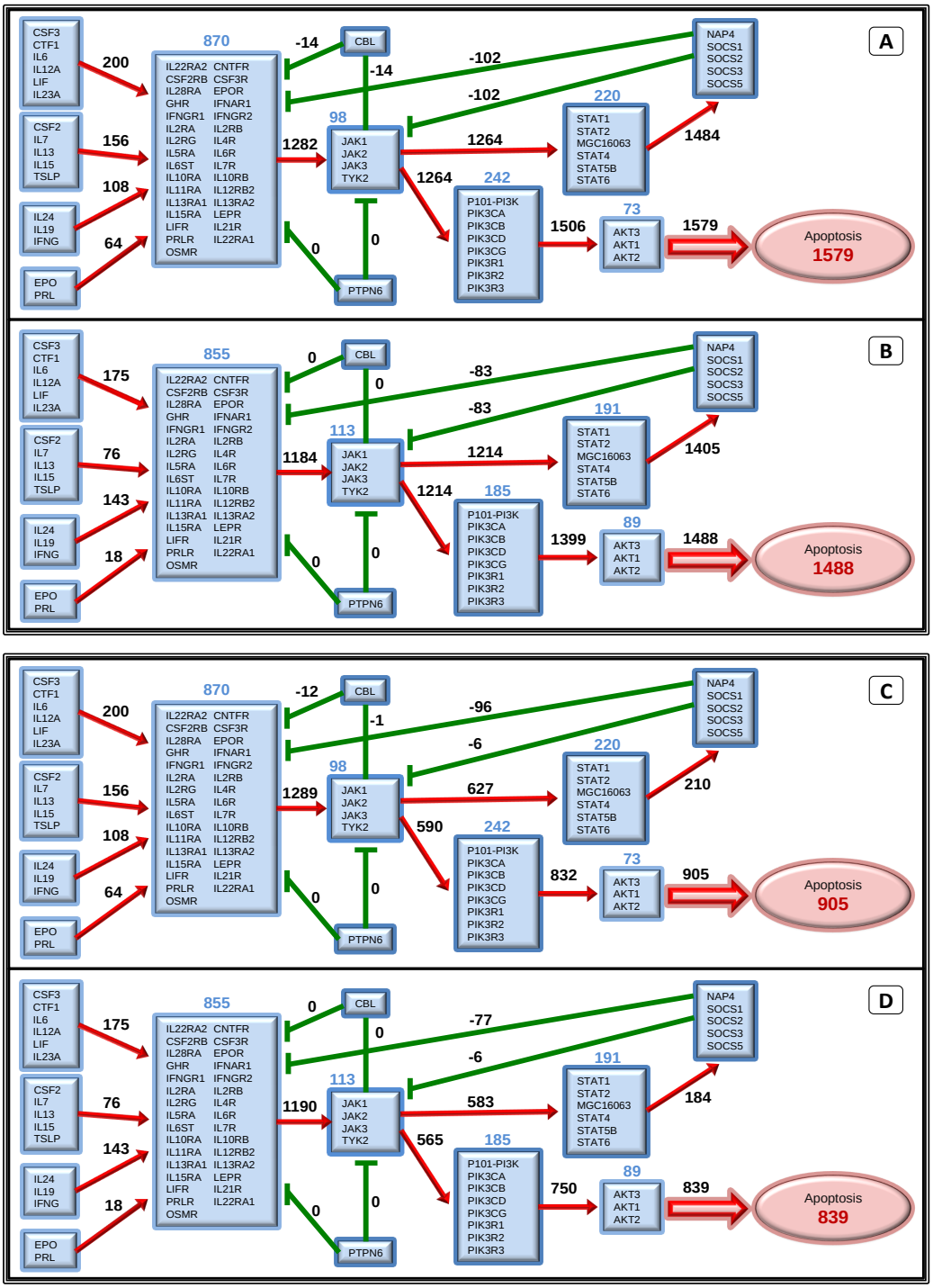

Figure 2: Calculation of combined ChIP-seq and array scores to a target process (*Apoptosis*) under oxidative stress (A, C) and control experiment (B, D). The scores were calculated by direct (A, B) and partitioned (C, D) score transfer approaches. The number on each node (gene) represented self-score of the gene. Red and green edges represented activation and inhibition properties, respectively.

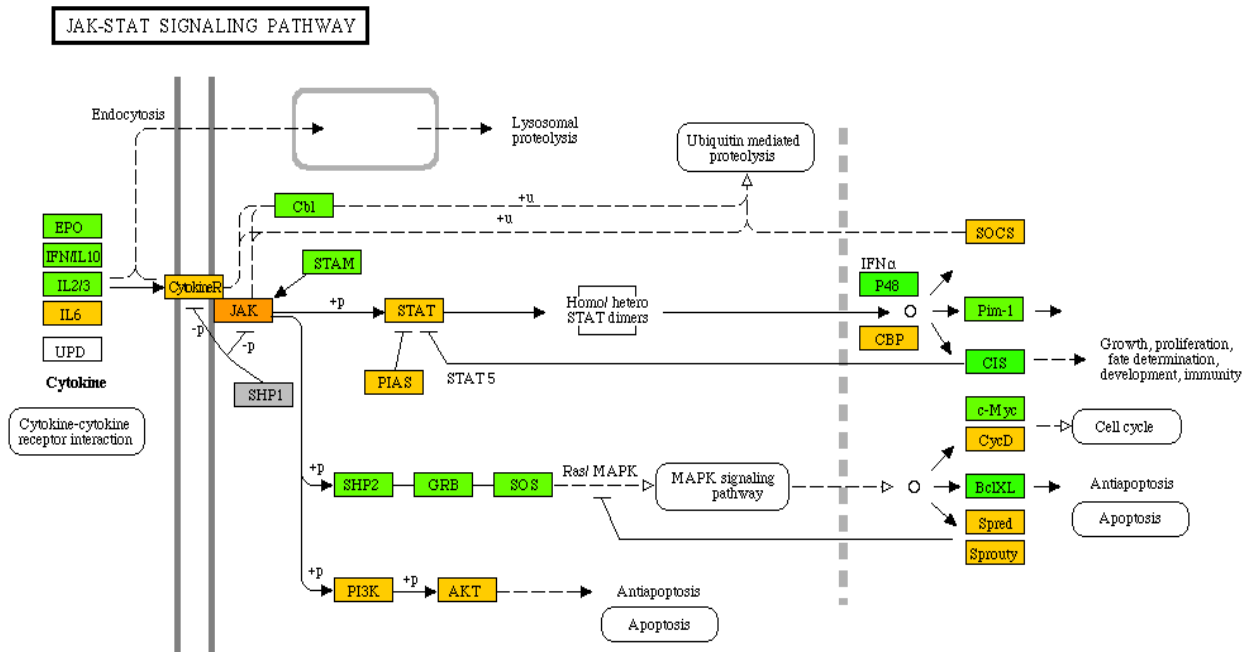

Figure 3: Murray oxidative stress data was mapped onto Jak-STAT signaling cascade by using kegArray tool. Green and orange colors indicate down-regulation and up-regulation values, respectively.

KEGG ID	Biological Process	*Average Cascade Scores of Target Processes*			
		Direct Score Transfer		**Partitioned Score Transfer**	
		Control	*Oxidative Stress*	*Control*	*Oxidative Stress*
hsa04630	Apoptosis	217.80	224.73	41.23	45.69
	Cell cycle	209.50	213.76	21.92	17.73
	Ubiquitin mediated proteolysis	99.5	105.92	10.26	12.00
	MAPK signaling	51.38	52.23	10.42	6.03
hsa04350	Cell cycle	2.92	3.07	2.29	2.11
	MAPK signaling	0.81	1.40	0.94	0.77
	Apoptosis	0.96	1.22	1.44	0.92
hsa04210	Survival	37.66	46.81	12.33	14.93
	Apoptosis	45.22	45.91	14.79	17.61
	Degradation	33.62	37.08	13.76	11.77
hsa04010	Proliferation & differentiation	172.73	199.24	14.67	11.90
	Cell cycle	22.61	33.66	2.50	2.00
	Apoptosis	14.75	26.33	1.08	1.45
	p53 signaling	7.42	10.13	0.85	0.63
	Wnt signaling	1.65	2.57	0	0

Table 1: Assigned scores of hsa04630 (Jak-STAT), hsa04350 (TGF-β), hsa04210 (Apoptosis), hsa04010 (MAPK) signaling pathways for control and oxidative stress experiments by applying direct (two columns on the left side) and partitioned (two columns on the right side) score transfer approaches. The average cascade score was calculated by using Equation 6.

4. Conclusion

In general, current approaches which integrate transcriptome data to molecular pathways are either data driven or model driven (Hahne et al., 2008; Viswanathan et al., 2008). In this study, we applied a hybrid approach which integrates large scale (i.e. transcriptome, ChIP-seq) data to quantitatively evaluate target process through a signaling cascade under the control of a biological process. In our framework signaling cascades acted as models. We used the integrated data as the attribute of a node and we transferred this information to en route of the pathway as scores. The scores reflected the current activity of the analyzed pathway. Our hybrid approach utilized equally the signaling cascade intrinsic properties (i.e. edge and node specifications) and scored genes based on large scale data. We used ChIP-seq data in order to further enrich the scores of genes in addition to transcriptome data. ChIP-seq and other large scale data can be further integrated into this framework. Our framework in its current state was applied to directed acyclic graphs. Current pathway analysis tools do not assign roles to the genes which are not differentially expressed for the enrichment. On the other hand, our hybrid approach considers signal relaying molecules even though they are not differentially expressed. We believe that our hybrid approach represents better cellular machinery rather than ignoring a gene product which is not differentially expressed but present in the biological process.

Acknowledgments

ZI was supported by The Scientific and Technological Research Council of Turkey (TUBITAK).

References

S. Aerts, D. Lambrechts, S. Maity, P.V. Loo, B. Coessens, F. De-Smet, L.C. Tranchevent, B. De-Moor, P. Marynen, B. Hassan, P. Carmeliet, and Y. Moreau. Gene prioritization through genomic data fusion. *Nature Biotechnology*, 24:537–544, 2006.

F. Al-Shahrour, R. Diaz-Uriarte, and J. Dopazo. Fatigo: a web tool for finding significant associations of gene ontology terms with groups of genes. *Bioinformatics*, 20(4):578–580, 2004.

H.J. Chung, C.H. Park, M.R. Han, S. Lee, J.H. Ohn, J. Kim, J. Kim, and J.H. Kim. Arrayxpath ii: mapping and visualizing microarray gene-expression data with biomedical ontologies and integrated biological pathway resources using scalable vector graphics. *Nucleic Acids Res.*, 33:W621–W626, 2005.

F. Cordero, M. Botta, and R.A. Calogero. Microarray data analysis and mining approaches. *Brief. in Funct. Genomics and Proteomics*, pages 1–17, 2008.

K.D. Dahlquist, N. Salomonis, K. Vranizan, S.C. Lawlor, and B.R. Conklin. Genmapp: a new tool for viewing and analyzing microarray data on biological pathways. *Nature Genet.*, 31:19–20, 2002.

T. De-Bie, L.C. Tranchevent, L.M. Oeffelen, and Y. Moreau. Kernel based data fusion for gene prioritization. *Bioinformatics*, 23:i125–i132, 2007.

N. Goffard and G. Weiller. Pathexpress: a web-based tool to identify relevant pathways in gene expression data. *Nucleic Acids Res.*, 35:W176–W181, 2007.

F. Hahne, A. Mehrle, D. Arlt, A. Poustka, S. Wiemann, and T. Beissbarth. Extending pathways based on gene lists using interpro domain signatures. *BMC Bioinformatics*, 9 (3):D354–D357, 2008.

Ingenuity. http://www.ingenuity.com.

H. Ji, H. Jiang, W. Ma, D.S. Johnson, R.M. Myers, and W.H. Wong. An integrated software system for analyzing chip-chip and chip-seq data. *Nature Biotechnology*, 26(11): 1293–1300, 2008.

M. Kanehisa, S. Goto, M. Hattori, K.F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, and M. Hirakawa. From genomics to chemical genomics: new developments in kegg. *Nucleic Acids Res.*, 34:D354–D357, 2006.

J. Kang, M. Gemberling, M. Nakamura, F.G. Whitby, H. Handa, W.G. Fairbrother, and D. Tantin. A general mechanism for transcription regulation by oct1 and oct4 in response to genotoxic and oxidative stress. *Genes Dev.*, 23(2):208–222, 2009.

W.J. Kent, F. Hsu, D. Karolchik, R.M. Kuhn, H. Clawson, H. Trumbower, and D. Haussler. Exploring relationships and mining data with the ucsc gene sorter. *Genome Res.*, 15: 737741, 2005.

N. Lopez-Bigas and C.A. Ouzounis. Genome-wide identification of genes likely to be involved in human genetic disease. *Nucleic Acids Res.*, 32:31083114, 2004.

B. Mlecnik, M. Scheideler, H. Hackl, J. Hartler, F. Sanchez-Cabo, and Z. Trajanoski. Pathwayexplorer: web service for visualizing high-throughput expression data on biological pathways. *Nucleic Acids Res.*, 33:W633–W637, 2005.

J.I. Murray, M.L. Whitfield, N.D. Trinklein, R.M. Myers, P.O. Brown, and D. Botstein. Diverse and specific gene expression responses to stresses in cultured human cells. *Molecular and Cellular Biology*, 15(5):2361–2374, 2004.

A. Nikitin, S. Egorov, N. Daraselia, and I. Mazo. Pathway studio-the analysis and navigation of molecular networks. *Bioinformatics*, 19:2155–2157, 2003.

J.M. Rosenbluth, D.J. Mays, M.F. Pino, L.J. Tang, and J.A. Pietenpol. A gene signature-based approach identifies mtor as a regulator of p73. *Molecular and Cellular Biology*, 28 (19):5951–5964, 2008.

G.A. Viswanathan, J. Seto, S. Patil, G. Nudelman, and S.C. Sealfon. Getting started in biological pathway construction and analysis. *PLoS Comput Biol*, 4(2):e16, 2008.