Estimation of Drug Treatment Effects on the Signaling Pathways

Seçkin Erce, Zerrin Işik*
Computer Engineering Department, Dokuz Eylül University, Tinaztepe Kampusu Buca, 35160 Izmir, Turkey.

* Corresponding author. Tel.: +90 232 301 7413; email: zerrin@cs.deu.edu.tr
Manuscript submitted January 10, 2017; accepted March 8, 2017.
doi: 10.17706/ijbbb.2017.7.4.213-220

Abstract: Developing a computational method to observe cell behaviors can help to understand the expected outcome of a drug treatment even before doing the wet-lab experiments. The main goal of this research is computing the effects of a drug treatment in the signaling level for a better understanding of cellular responses after a drug treatment. The proposed algorithm can work on various biological networks to quantitatively evaluate the effects of a drug treatment in the cell by using gene expression data. The method was applied on the integrated KEGG signaling pathway and the most effected proteins were identified after the application of 14 different drugs on lymphoma cells. The results showed that the most affected proteins are not the direct target proteins of the given drugs. Indeed, the protein activities in the distant parts of signaling pathways are highly changed upon drug treatments. The literature validation showed that some of the proteins, which are commonly effected by the treatment of several drugs, have cancer related cellular activities as well.

Key words: Bioinformatics, drug effect on pathways, KEGG signaling pathway, microarray gene expression.

1. Introduction

The human genome contains approximately 21,000 genes. At any given moment, for different tissues, some combinations of these genes are active and others are inactive. Scientists can answer this question for any cell sample or tissue by analyzing gene expression profiles which a microarray analysis technique. Microarray analysis can help to determine genome wide behaviors of genes under different experimental conditions such as oxidative stress, drug treatment, disease treatment etc. These experiments generate data for thousands of genes across multiple experimental conditions, so the statistical analysis of these data is curial to understand cellular mechanisms under different conditions [1]. Pathway-based analysis is fairly new perspective to interpret such large amount of gene expression profiles in the cellular signaling and metabolic levels [2].

Pathways are collections of genes and proteins that collaboratively perform a well-defined biological task. For example, proteins that work to synthesize metabolites within a cell are grouped into various metabolic pathways. Such biological pathway data is freely accessible in the last years. Pathway data sets also cover various types of cell signaling networks in which a group of genes work collaboratively for controlling cellular behavior as a response to an external perturbation signal. Therefore, integration of the graphical topology of signaling networks and microarray experimental data can explain how a drug treatment (a type of perturbation for the cell) changes cell behavior in the molecular level [3]. Traditional genome-wide microarray experiments reveal lists of the effected genes that are assumed to be a response of cells in the

genomic level. However, the analysis of only a list of effected genes cannot provide a fully understanding of the molecular basis of cellular processes. Hence, new algorithms are necessary to overcome this problem. A recent study proposed a score flow algorithm to quantitatively visualize cell responses in the signaling pathways [4]. The algorithm traverses a signaling pathway and calculates an activity score of each biological process attached to the pathway. In this study, we adapted this score flow algorithm on a large signaling network to evaluate effects of different drug treatments on the lymphoma cancer cell lines. The algorithm found out specific genes that are mostly affected due to the given drug in the cancer cells. Therefore, the proposed method might be beneficial for a better understanding of cellular responses under various perturbations applied on the cells.

2. Material and Methods

2.1. Data

2.1.1. Microarray data

There are various technologies to measure cell response in the genomic level. Microarray experiments aim to measure mRNA levels of genes upon a specific condition. After statistical analysis of microarray experiments, the expression levels of genes are measured in terms of increase or decrease in the amount of mRNA compared the control samples. We used a public microarray experiment B-cell lymphoma cancer cells were treated with 14 different drugs (Table 1) to observe their effects on the cancer cells [5]. In this dataset, every gene has 3 expression measurements for both drug-treated and control samples. In order to compute one measurement for both drug-treated and control samples, we took the median of 3 samples for each condition. So, both drug-treated and control samples are represented by only one measurement that can be provided as the input score of the gene. Then we translated each gene symbol to the corresponding gene identifier. We combined the score of a gene, which has multiple samples in the experiment, by taking the median values of this gene. Finally, for each gene, a single drug-treated and control expression value is provided to the algorithm.

2.1.2. Pathway data

Pathway Commons is free pathway information about different taxonomy [6]. It has an online interface that enables researchers to examine lots of information about biological pathways from different sources. Pathway Commons has a downloadable format for specific pathway data and provides a web service so everyone can use queries and access all data. A pathway includes two elements: node and edge. Source node represents a protein or drug; the target node shows a protein. The edge represents the type of biological relation or reaction between source and target nodes [6]. The goal of our study is to detect the genes are effected after a drug treatment. Therefore, we just chose “interacts-with” and “consumption-controlled-by” edge types covered in KEGG pathways that were downloaded from Pathway Commons database.

2.1.3. Drug targets

A drug target is a protein or enzyme, which is affected by the designed drug, and its original function in the cell is changed or corrupted after binding of the given drug its binding pocket. For example, the known targets of “Aclacinomycin A” drug are TOP1, TOP2A, and TOP2B proteins. Drug targets were derived from screens using cell culture or whole organisms and phenotypic or molecular readouts. Discovery of the direct target(s)of a drug is often the most challenging and time-consuming step of the drug development process [7]. In this project, we utilized the STITCH database for finding the targets of 14 drugs. STITCH is a searchable database that coordinates data obtained from metabolic pathways, crystal structures and drug–target connections [8]. Text mining and chemical structure comparability is utilized to predict relations between chemicals. Each protein-interaction can be followed back to the first information sources. It contains interactions for between 300,000 little atoms and 2.6 million proteins from 1,133 organisms.

Total number of known targets of each drug in our dataset is listed in Table 1.

2.2. Pre-processing of Pathways

We applied a processing on the original KEGG pathways to transform its topology into a tree structure. For this purpose, we used a well-known graph traversal algorithm: Breadth First Search (BFS). For a specific drug *A*, the pathway is transformed into its BFS tree. In this transformation the first interactors of drug A - that are known targets of *A* - are placed into the first level of the BFS tree. Then the next interactors of these proteins are placed to the second level, and so on. This transformation continues till no protein left. An example transformation of input pathway for "Aclacinomycin A" is shown in Fig. 1. As summary, each drug has its own network that is transformed from the larger signaling pathway by only using the known targets of this drug and their children nodes in the pathway.

Total Number of Edges, nodes and drug targets are listed in the "# of Edges", "# of Nodes" and "# of Drug Targets" columns, respectively. After transformation of pathways by using BFS algorithm, the depth of each drug network is given in the "# of Levels" column.

Table 1. The Details of Each Drug Network

Drug Name	# of Edges	# of Nodes	# of Levels	# of Drug Targets
Aclacinomycin A	13979	4791	12	3
Mitomycin C	251979	12955	8	18
Rapamycin	81717	13003	7	128
Doxorubicin hydrochloride	118752	12875	7	108
H-7 Dihydrochloride	59644	4430	11	2
Geldanamycin	116033	12826	9	21
Methotrexate	274265	13011	8	33
Vincristine	263037	11558	11	7
Blebbistatin	55838	5690	11	6
Monastrol	94305	8039	11	2
Camptothecin	148146	12991	7	20
Trichostatin A	139873	11581	10	44
Etoposide	131129	13034	8	63
Cycloheximide	116875	12797	10	86

After transformation of the input pathway into a drug tree (i.e., network), the edges were removed which connect genes in the same level or to the upper levels. After this edge removal process, single nodes, which have no incoming / outgoing edges were also removed. By applying these operations, potential cycles in the tree were eliminated. In Fig. 1, red edges and nodes were deleted to prevent such cycles in final drug tree. After the transformation and cycle removal stages, the network content (total number of edges, nodes, levels, drug targets) for each drug is listed in Table 1.

2.3. Score Flow Algorithm

The algorithm works iteratively and computes the score of the proteins in a level-wise manner. The initial scores of proteins are provided by the microarray experiment of each drug. According to BFS algorithm, children nodes are connected with their parent nodes. For each iteration of the algorithm, the proteins in the network are processed in level order, i.e., the proteins in the level *i* are processed before the proteins in the level *i+1*. The output-score of a protein-node is the summation of its input score (microarray experiment) and the weights of all incoming edges to that node. Topologically in a tree, a parent node transfers its

output-score to its children by partitioning its output-score between each child equally. Then this partitioned score is assigned as the weight of each edge, which connects the parent and a child node. The pseudo-code of this algorithm is given in Table 2.

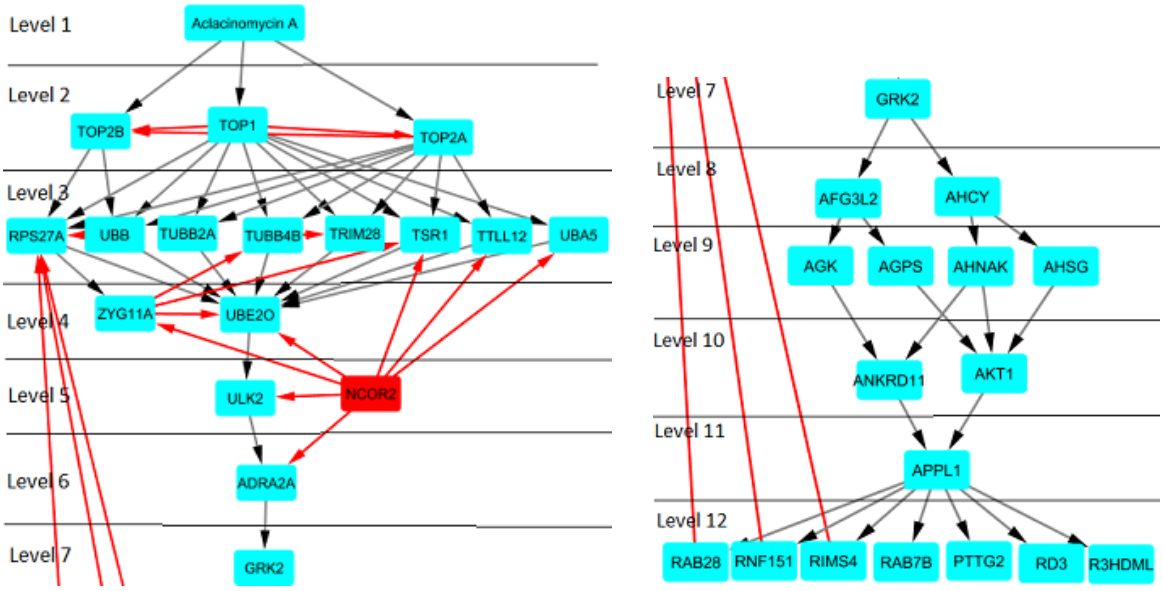

Fig. 1. Transformed network for AclacinomycinA drug. The red nodes and edges were deleted to prevent potential cycles in the original pathway.

Table 2. Pseudo Code for the Score Flow Algorithm

```
Score: indicates initial score of each node provided by microarray file
outScore: contains out-score of each node
marray: indicates self-score of each node provided by microarray file
outAdj(x): out-adjacency list of node x.
ElementCount(x): Number of edges from x node
All(x): List of all proteins in the network
Levelization information V0,...,V1-1 is obtained by running the BFS algorithm
Initialization:
     For each vertex x in All (x) do
      If marray(x) contains then
         Score(x) = marray (x)
        outscore(x) = marray (x)
      else
        Score(x)=0
        outscore(x) = 0
Score Computation:
    For each level = 0,...,l-1 do
       For each vertex x in Vi do
          For each vertex y in outAdj{x} do y
                outscore(y) = outscore(y)+ (outscore(x)/ElementCount(x))
```

3. Results

We applied the score flow algorithm for 14 drugs separately. One detailed example for Aclacinomycin A drug is given in the Fig. 2 that shows the score calculation between the first, second, third, and fourth levels of Aclacinomycin A drug network. Aclacinomycin A has three know protein targets (TOP1, TOP2A, TOP2B) that are in the second level of the drug network (tree). The drug itself does not transfer a score to its

children that's why the edge weights are set to zero for the edges connecting first and second levels. For example, TOP2B protein in the second level has an input score of 8.48, the weights of all incoming edges to this protein are zero. The output score of a protein is the summation of its input score (gene expression) and the weights of all incoming edges to this protein. So, the output-score of TOP2B will become 8.48 as well. This output-score is partitioned between its children (in total 32); each child of TOP2B will get a score of 0.33 that is set as the incoming edge weight of each child of TOP2B. This score flowing scheme is applied until reaching the deepest level of the drug network. After running of the score flow algorithm with this manner for each drug separately, the final output-scores of all proteins were recorded. The effect of a drug treatment on a protein in the network was identified based of difference between its initial microarray score and the final output-score computed by the score flow algorithm. The proteins with highest difference in their scores are selected as the most affected ones and listed in Table 3. The most affected proteins are not the direct target proteins of the given drugs, since they placed in the lower levels, most commonly in the 3rd level, of the tree. This result was also found in recent studies [9], [10]. These proteins have activities in the distant parts of signaling pathways and showed the most remarkable reactions in this pathway after the application of drug treatments. When we analyzed these proteins, some of them were found to be in common between different drugs. For example, *COX7A2* is identified as the highly effected protein in the drug network of Aclacinomycin A, H-7 Dihydrochloride, Methotrexate, and Mitomycin C drugs. Similarly, *NFE2L1* is identified in the drug networks of Methotrexate and Mitomycin C. *USP15* is also highly influenced protein in the drug networks of Camptothecin and Etoposide.

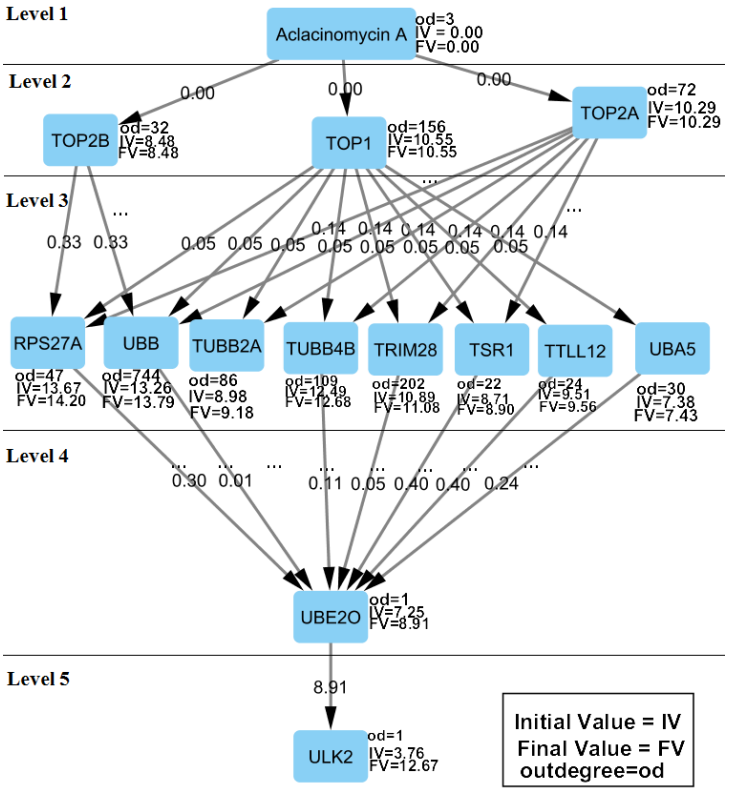

Fig. 2. Score calculations for the "Aclacinomycin A" drug between first five levels.

We performed a literature search about the highly affected proteins observed for several drug treatments. A previous study showed that the dys-regulation NFE2L1 protein might lead tumor [11]. Another study also suggested that NFE2L1 expression is related to the cell survival under stress condition [12]. A recent study found that USP15 protein regulates the TGF-β pathway and USP15 has an important

role in glioblastoma cancer [13]. These proteins, which are highly effected due to more than one drug treatment in our dataset, appear to have some roles in different cell activities related with the cancer development. As a summary, such previous studies in literature also support our results found by the score flow algorithm that might help a better interpretation of molecular responses of a cell after a drug treatment.

For each drug, three proteins with the highest score change (given in the "Difference" column) were chosen as the highly affected proteins.

Table 3. The Most Affected Three Proteins for All 14 Drugs

Drug Name	Protein Name	Initial Score	Final Score	Difference	Level
Aclacinomycin A	ZRANB2	9,45	71,99	62,55	3
Aclacinomycin A	COX7A2	11,33	73,02	61,69	6
Aclacinomycin A	ECH1	9,50	70,26	60,76	6
Blebbistatin	TAF1	5,44	133,08	127,64	3
Blebbistatin	USP7	8,30	96,49	88,19	3
Blebbistatin	YME1L1	10,07	88,60	78,53	3
Camptothecin	USP15	5,70	225,42	219,72	3
Camptothecin	TEX10	6,97	155,05	148,08	3
Camptothecin	TFPI2	3,65	124,86	121,21	4
Cycloheximide	YEATS4	9,28	227,75	218,47	3
Cycloheximide	ZBTB7C	6,10	207,20	201,10	3
Cycloheximide	WDR61	8,43	171,72	163,29	3
Doxorubicin hydrochloride	ZCCHC11	5,15	258,38	253,23	3
Doxorubicin hydrochloride	VSIG8	3,70	190,53	186,83	3
Doxorubicin hydrochloride	ZNFX1	9,96	173,33	163,37	3
Etoposide	USP15	6,04	345,46	339,42	3
Etoposide	WDR26	8,96	180,13	171,17	3
Etoposide	SMURF2	4,46	153,33	148,87	3
Geldanamycin	TK1	6,39	218,23	211,84	3
Geldanamycin	ZNF346	6,03	156,38	150,36	3
Geldanamycin	TALDO1	8,69	152,95	144,26	3
H-7 Dihydrochloride	COX7A2	11,39	120,69	109,30	6
H-7 Dihydrochloride	ZNF30	5,80	83,15	77,35	4
H-7 Dihydrochloride	ZNRF4	4,15	81,37	77,22	4
Methotrexate	YARS	9,80	142,33	132,53	3
Methotrexate	NFE2L1	7,16	139,13	131,97	3
Methotrexate	COX7A2	11,02	115,07	104,05	4
Mitomycin C	ZMPSTE24	9,97	109,39	99,42	3
Mitomycin C	COX7A2	11,07	105,53	94,46	4
Mitomycin C	NFE2L1	6,94	100,78	93,84	3
Monastrol	RXFP1	3,82	133,42	129,61	4
Monastrol	SIAH2	11,40	122,16	110,76	4
Monastrol	TMEM223	8,45	115,43	106,98	4
Rapamycin	RCC1L	6,01	172,69	166,68	3
Rapamycin	TRIP6	6,75	150,61	143,86	3
Rapamycin	ZFP36	10,53	147,42	136,89	3
Trichostatin A	YTHDF3	7,86	236,42	228,57	3
Trichostatin A	YARS2	8,58	182,77	174,19	3
Trichostatin A	UBE2NL	4,21	168,68	164,48	3
Vincristine	PSMB10	0,00	131,70	131,70	4
Vincristine	SBDS	10,92	129,93	119,01	4
Vincristine	MRPL52	9,60	125,57	115,98	4

4. Conclusion

Traditional analysis of microarray experiments should integrate more information from the molecular signaling levels to have a better understanding of cellular responses. Therefore, signaling pathway and interaction network analysis became more attractive to give a new perspective for the classic analysis of gene expression data [14]-[16]. New algorithms have started to highlight important regulator proteins and

cellular processes in pathways; such results had not been found by applying naive differentially expression analysis [9], [17]-[20].

In this study, we applied a score flow algorithm on an integrated KEGG signaling pathway and showed the molecular effects of different drug treatments applied on lymphoma cancer cells. The algorithm highlighted specific proteins that are mostly affected due to the given drug. Some of these proteins were also found in other studies and they work on different cancer related cell activities. Therefore, the proposed method might be useful to understand cellular responses under various experimental conditions applied on the cells. We will continue this study with the literature analysis of top-10 mostly affected proteins for 14 drugs. We believe that such analysis might bring more insights to understand the cellular effects of these drugs applied on lymphoma cancer cells. This method might help to predict the synergistic behaviors of drug combinations in the molecular signaling level. As a future work, a visual tool for this method will be developed to open its usage for biologists and pharmaceutical developers.

References

[1] Kerr, M. K., & Churchill, G. A. (2007). Statistical design and the analysis of gene expression microarray data. *Genet Res, 89(5-6),* 509-14.

[2] Pas, M. F. W., Hemert, S., Hulsegge, I., Rebel, J. M. J., & Smits, M. A. (2008). Pathway analysis tool for analyzing microarray data of species with low physiological information. *Adv Bioinformatics,* 719468.

[3] Zhang, D.Y ., Ye, F., Gao, L., Liu, X., Zhao, X., & Che, Y., *et al*. (2009). Proteomics, pathway array and signaling network-based medicine in cancer. *Cell Div., 4,* 20.

[4] Isik, Z., Ersahin, T., Atalay, V., Aykanat, C., & Cetin-Atalay, R. (2012). A signal transduction score flow algorithm for cyclic cellular pathway analysis, which combines transcriptome and ChIP-seq data. *Mol Biosyst, 8,* 3224-31.

[5] Bansal, M., Yang, J., Karan, C., Menden, M. P., Costello, J. C., & Tang, H., *et al.* (2014). A community computational challenge to predict the activity of pairs of compounds. *Nat Biotechnol., 32(12),* 1213-22.

[6] Cerami, E. G., Gross, B. E., Demir, E., Rodchenkov, I., Babur, O., & Anwar, N., *et al*. (2011) Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res. 39(Database issue),* D685–D690.

[7] Warren, J. (2011). Drug discovery: Lessons from evolution. *Br J Clin Pharmacol., 71(4),* 497–503.

[8] Szklarczyk, D., Santos, A., Mering, C., Jensen, L. J., Bork, P., & Kuhn, M. (2016). STITCH 5: Augmenting protein-chemical interaction networks with tissue and affinity data. *Nucleic Acids Res. 4, 44(D1),* D380-4.

[9] Isik, Z., Baldow, C., Cannistraci, C. V., & Schroeder, M. (2015). Drug target prioritization with perturbed genes and network information. *Sci Rep.. 5,* 17417.

[10] Iskar, M., Zeller, G., Blattmann, P., Campillos, M., Kuhn, M., & Kaminska, K. H., *et al.* (2013). Characterization of drug-induced transcriptional modules: towards drug repositioning and functional understanding. *Mol Syst Biol., 9,* 662.

[11] Oh, D. H., Rigas, D., Cho, A., & Chan, J. Y. (2012). Deficiency in the nuclear-related factor erythroid2 transcription factor (Nrf1) leads to genetic instability. *FEBS J., 279(22),* 4121-30.

[12] Biswas, M., Kwong, E. K., Park, E., Nagra, P., & Chan, J. Y. (2013). Glycogen synthase kinase 3 regulates expression of nuclear factor-erythroid-2 related transcription factor-1 (Nrf1) and inhibits pro-survival function of Nrf1. *Exp Cell Res., 319(13),* 1922-31.

[13] Eichhorn, P. J., Rodon, L., Gonzalez-Junca, A., Dirac, A., Gili, M., & Martinez-Saez, E., *et al*. (2012). USP15 stabilizes TGF-β receptor I and promotes oncogenesis through the activation of TGF-β signaling in glioblastoma. *Nat Med., 18(3),* 429-35.

[14] Cline, M. S., Smoot, M., Cerami, E., Kuchinsky, A., Landys, N., & Workman, C., *et al.* (2007). Integration of biological networks and gene expression data using Cytoscape. *Nat Protoc*., *2(10)*, 2366-82.

[15] Kotelnikova, E. A., Pyatnitskiy, M., Paleeva, A., Kremenetskaya, O., & Vinogradov, D. (2016). Practical aspects of NGS-based pathways analysis for personalized cancer science and medicine. *Oncotarget, 7(32)*, 52493-516.

[16] Garcia-Campos, M. A., Espinal-Enriquez, J., & Hernandez-Lemus, E. (2015). Pathway analysis: State of the art. *Front Physiol., 6,* 383.

[17] Zhou, J. X., Isik, Z., Xiao, C., Rubin, I., Kauffman, S. A., & Schroeder, M. *et al*. (2016). Systematic drug perturbations on cancer cells reveal diverse exit paths from proliferative state. *Oncotarget, 7(7),* 7415–25.

[18] Kaushik, A., Ali, S., & Gupta, D. (2017). Altered Pathway Analyzer: A gene expression dataset analysis tool for identification and prioritization of differentially regulated and network rewired pathways. *Sci Rep., 7,* 40450.

[19] Tarca, A. L., Draghici, S., Khatri, P., Hassan, S. S., Mittal, P., Kim, J. S., *et al*. (2009). A novel signaling pathway impact analysis. *Bioinformatics*, *25,* 75-82.

[20] Mueller, A. J., Tew, S. R., Vasieva, O., Clegg, P. D., & Canty-Laird, E. G. (2016). A systems biology approach to defining regulatory mechanisms for cartilage and tendon cell phenotypes. *Sci Rep., 6,* 33956.

Seçkin Erce got his bachelor degree from Computer Engineering Department of Ege University, Turkey in 2008. His thesis was related with web 3.0 based virtual assistant designing and its implementation. He has started a master degree in 2014 in Computer Engineering Department of Dokuz Eylül University, Turkey under the supervision of Dr. Isik. His master thesis is in the bioinformatics field, specifically it is the investigation of drug treatment effects in the pathway levels.

He had been working in Ibtech, Finansbank's IT company from 08/2008 to 09/2010. He had completed projects such as IBAN transitions, FinBor, DBKK, SBT and Warehouse Tracking System. He worked in Astron, Yasar Holding IT Company from 01/2011 to 06/2011; worked on SAP projects. He is currently working for the government at general Directorate of Highways as Information Technology supervisor.

Zerrin Isik got a Ph.D. degree from Computer Engineering Department of Middle East Technical University in 2011. Her Ph.D. dissertation established a novel pathway enrichment system based on integration of gene expression, ChIP-sequencing data and cyclic signaling pathways to assess biological activity of specific cell processes. She worked as a post-doctoral researcher in Biotechnology Center of TU Dresden, Germany from 2011 to 2014. She works as an Assistant Professor in Computer Engineering Department of Dokuz Eylül University since 2014.

Her main research fields are bioinformatics and machine learning methods. She had developed new algorithms and tools for drug-target identification, cancer biomarker prediction, transcriptional and interaction network analysis, assessment of signaling pathways.