BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 3 2014, pages 406–413
doi:10.1093/bioinformatics/btt685

Systems biology

Advance Access publication December 12, 2013

Genes related to differentiation are correlated with the gene regulatory network structure

Matan Bodaker[1], Eran Meshorer[2], Eduardo Mitrani[1] and Yoram Louzoun[3,*]

[1]Department of Cell and Developmental Biology and [2]Department of Genetics, The Alexander Silberman Institute of Life Sciences, Hebrew University of Jerusalem, Jerusalem 91904, Israel and [3]Department of Mathematics and Gonda Brain Research Center, Bar-Ilan University, Ramat Gan 52900, Israel

Associate Editor: Igor Jurisica

ABSTRACT

Motivation: Many secondary messengers, receptors and transcription factors are related to cell differentiation. Their role in cell differentiation can be affected by their position in the gene regulatory network. Here, we test whether the properties of the gene regulatory network can highlight which genes and proteins are associated with cell differentiation. We use a previously developed purely theoretical algorithm built to detect nodes that can induce a state change in Boolean gene regulatory networks, and show that most genes predicted to participate in differentiation in the theoretical framework are also experimentally known to be associated with such differentiation. These results show that genes related to differentiation are associated with specific features of the genetic regulatory network. The proposed algorithm produces a better classification than simple network measures such as the nodes degree or centrality. Boolean networks were used in many previous theoretical models. Here, we show a direct application of such networks to the detection of genes and subnetworks related to differentiation. The subnetwork emerging from the genes and edges that are predicted to be associated with differentiation are the most active molecular pathways experimentally described to be involved in cell differentiation.

Availability and implementation: http://peptibase.cs.biu.ac.il/homepage/Boolean_network_conversion_code.zip.

Contact: louzouy@math.biu.ac.il

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on June 3, 2013; revised on October 25, 2013; accepted on November 19, 2013

1 INTRODUCTION

During differentiation, stem cells are converted to become new types of cells. The converted cells can become similar to the cells that triggered the differentiation, which is often the case with differentiation of adult stem cells, or they may become a different cell type. This can happen in transdifferentiation or during differentiation of tissue stem cells from a nearby non-neighboring cell population. For example, the cells in the mesenchyme may send differentiation signals to the stem cells of the epithelium (Alonso and Fuchs, 2003; Krause *et al.*, 2001).

Some genes and their products are classically related to cell differentiation in the literature, as they play a key role in at least one case where one cell type is converted to another. These genes include, e.g. the members of signaling pathways related to cell differentiation such as the epidermal growth factor (EGF) signaling pathway (Herbst, 2004), the transforming growth factor beta signaling pathway (Moustakas *et al.*, 2002) and the nerve growth factor signaling pathway (Klesse *et al.*, 1999). These genes and their products share different biophysical and functional properties. Some are growth factors acting as extracellular signals. Others are the corresponding ligand receptors. Other signaling proteins transmit the differentiation signal to the nucleus. There, transcription factors and other proteins regulate chromatin structure and gene expression.

To model differentiation we use Boolean networks, which have been shown to produce multiple features of regulatory gene networks (Kauffman, 1993). Furthermore, Boolean networks can approximate well the continuous expression profiles of genes (Martin *et al.*, 2007; Serra *et al.*, 2007) (Serra *et al.*, 2004). We developed a theoretical algorithm to detect, using a random Boolean gene regulatory network, which genes and their products can participate in computer-simulated differentiation-like events (Bodaker *et al.*, 2013). Here, we show that genes related to this *in silico* differentiation are associated with differentiation as documented in the literature.

Multiple supervised machine learning algorithms have been proposed to classify genes into categories, using gene expression patterns (Eisen *et al.*, 1998; Gibbons and Roth, 2002; Zhou *et al.*, 2002) or properties of the protein interaction network (Deng *et al.*, 2003a; Hishigaki *et al.*, 2001; Karaoz *et al.*, 2004; Schwikowski *et al.*, 2000; Sharan *et al.*, 2007; Vazquez *et al.*, 2003). Here, we propose a novel approach to define genes related to differentiation based on first principles and not on blind machine learning algorithms.

The function of genes and proteins is typically studied using their sequence/structure, expression pattern and biochemical interactions. However, the function of genes may actually be induced by or at least correlated with the position of its protein in the genetic regulatory and protein–protein interactions network. Here, we show that first principles on the expected dynamics of networks can help detecting genes correlated with differentiation.

*To whom correspondence should be addressed.

2 METHODS

2.1 Model organism and data processing

To study cell differentiation, we used the curated genetic regulatory network of Cui *et al*. (2007). A basic concept in Boolean networks is that each node has at least one regulator, which determines its Boolean value at each time step. Therefore, genes that have no regulators were removed from the network (recursively). Another set of genes that were removed (again recursively) are genes that do not regulate other genes. These genes have no effect on the dynamics of the network, and their removal has no effect on the results. The resulting processed network contains 435 nodes (the vast majority are genes or their products) and 1550 edges between them, 1196 of which are of positive regulation and 354 are of negative regulation. The network has seven small loops, each with two genes. There was a single large multiply connected component containing 418 genes. Some of the small loops were found to influence the large multiply connected loop, whereas others were found to be influenced by it. By definition, none of the small loops both influences and is influenced by the large multiply connected loop. The three remaining genes connected the different loops. Because the small loops that influence the large one were likely to create a bias, the analysis was performed using only the large multiply connected loop, which will be denoted as the 'human network' (HN) in the following sections. In Supplementary Section S1, we discuss the detailed effect of the removal of the small loops. In Supplementary Section S2, we present the small loops.

The HN has an average degree of 3.6 ± 3.8, and a scale-free distribution for both the in- and the out-degree (Supplementary Section S3). The typical distance between nodes in the graph is approximately a Poisson distribution with a typical distance of 6 ± 2 (Supplementary Section S3) and a relatively low clustering coefficient of 0.03. These features are similar to what is often reported in regulatory networks (Ahn *et al*., 2009; Gama-Castro *et al*., 2011; Itzhack *et al*., 2013).

2.2 Theoretical model for cellular differentiation

2.2.1 Boolean networks The edges between the genes and their signs provide only partial information for the definition of a Boolean network. The regulatory function of each gene (describing how it is influenced by its regulatory genes, e.g. NAND) must be given as well. There are multiple options for each gene, e.g. a gene with two positive regulators may have AND or OR as its regulatory function. Because setting the regulatory function for gene is independent of the other genes, the total number of Boolean networks that are consistent with the HN data is at least exponential to the number of genes that have more than a single regulator (which are the vast majority of the genes). Thus, the HN represents a family of Boolean networks, consistent with the HN data.

Here, we assign a Boolean function consistent with the number of activators and repressors each gene has. These Boolean functions are called nested canalyzing Boolean functions, as will be explained. Harris *et al*. (2002) have shown that the functions describing genetic regulation in eukaryotes are canalyzing. Kauffman *et al*. (2004) reanalyzed these data and have shown that practically all the functions of gene regulation in eukaryotes are nested canalyzing Boolean functions (NCBF) (see Supplementary Section S4 for definition and example). Thus, although there are many possible ways to reproduce a random Boolean network from observed networks, the choice of nested canalyzing functions seems to be the most appropriate from a biological point of view. Given this definition, the network produced is based on the selection of the hierarchy of the variables in the nested canalyzing function, which was random in this case, as no information was available on the hierarchy between each gene's regulators (Supplementary Section S4).

2.2.2 Computer-simulated differentiation-like events in random Boolean networks We have previously developed a general framework to model differentiation-like events in random Boolean networks (Bodaker *et al*., 2013). We investigated these differentiation-like events in systems based on random Boolean networks given a set of signaling nodes and two cycles. Basically, we studied networks that can converge to at least two different cycles. A cycle is a finite cyclic set of states of the network that a Boolean network can converge to (each cycle has a finite number of steps that evolve in time from one to the other and finally the last step evolves to the first one). Given such a network, we studied two cells (both represented by similar networks) each resting in a different cycle, and tested whether using a given subset of the nodes in one cell as inputs to the other cell can drive the second cell to be in the same cycle as the first cell. We denote the cell changing its cycle as the induced cell and the other cell as the inducing cell.

2.2.3 Unsynchronized conversions From the biological point of view most adult tissues are in steady state, with new cells being produced continuously to offset the number of dying cells. Thus, constant differentiation of tissue stem cells is observed in multiple adult organs, such as the intestine, the skin and blood (Barker *et al*., 2010; Pittenger *et al*., 1999). This led us to model differentiation as a change in a simulated cell that is independent in the initial step both of the cycle the induced cell rests in and of the cycle the inducing cell rests in. We named the differentiation-like events as unsynchronized quality conversion (UQC), which occurs if the induced cell always changes its cycle to the same cycle, following signals from the inducing cell. For UQC to occur, we require that the conversion should not depend on the phase of the two cycles. The cycle that the induced cycle converged to is referred to as the target cycle. A detailed technical definition of UQC in a network is given in Bodaker *et al*. (2013).

2.2.4 Algorithm for detecting candidate nodes to enable the differentiation-like events Based on the UQC concept, we have developed an *a priori* algorithm to detect nodes that can serve as signaling nodes between two cycles and induce UQC. Although the detailed mathematical details of the algorithm were previously explained (Bodaker *et al*., 2013), we here provide a short explanation of the basic principles driving this algorithm, see Supplementary Section S5 for the detailed steps of the algorithm. Each node in the network represents a gene or a protein. The algorithm works mostly by elimination:

(1) Eliminate nodes that behave similarly between the two cycles (basically have the same patterns between the two cycles).
(2) Eliminate nodes that neither can close a directed circle nor link between directed circles in the remaining graph.
(3) Eliminate the nodes that are not fixed in both cycles.
(4) Select the remaining nodes as candidates to induce UQC.

We have found that if nodes are selected according to the previous algorithm, in >90% of the cases, UQC will occur. Similar performances of the algorithm were also obtained for Boolean networks consistent with the HN. However, our previous work focused on the pure mathematical aspects of the algorithm on these networks with no biological meaning (Bodaker *et al*., 2013).

2.2.5 Sets of nodes derived from the algorithm As mentioned, the algorithm works mostly by elimination. After the first step of eliminating the nodes that behave similarly between the cycles, the two remaining types of nodes can be categorized: (i) nodes found in loops (which may be multiply connected) and (ii) nodes found in simple paths (paths that do not close a circle). The simple paths themselves can also be categorized: (i) simple paths linking between loops and (ii) simple paths that end with nodes that do not regulate any of the remaining nodes. These remaining loops are referred in graph theory as strongly connected components (SCCs). Categorizing the remaining strongly connected components (RSCCs) is less trivial and is beyond the scope of simple definitions of graph theory.

To further reduce the number of converting nodes, we further eliminated nodes from the signaling set, if the conversion could occur without them. Following this removal, RSCCs can be categorized into two kinds, the ones that contain nodes found in the reduced signaling sets and the ones that do not contain such nodes. The nodes found in the RSCCs and are essential for differentiation in the algorithm are the ones we believe are the most critical nodes for the conversion to occur. Therefore, through the following analysis, we rank all nodes by the fraction of simulation, where they belong to this set of RSCCs, referred to as sufficient RSCCs. For a precise definition and explanation of the logic behind these definitions, see a previously more theoretical analysis (Bodaker *et al.*, 2013, Supplementary Section S6).

2.2.6 Frequency of participation in conversions We tested whether genes in the HN associated with the theoretical model of UQC are also associated with cell differentiation in the literature. The genes of the HN were sorted according to the frequency of their participation in UQC as follows (a single simulation takes on average a couple of minutes on a single core and the analysis presented here was performed on a grid with 32 cores):

(1) A total of 10 000 Boolean networks consistent with the HN were created. This was performed by choosing randomly, for each gene, a Boolean function consistent with the number of activators and repressors the gene has, according to the HN as described in Section 2.
(2) In all, 1 000 initial random states were iterated until they converged to different cycles.
(3) If at least two cycles were observed, two cycles were randomly picked.
(4) The algorithm searched for candidate genes to induce UQC, as mentioned earlier in the text.
(5) If candidate genes were found and UQC occurred and the target cycle was native (a cycle that exists in the Boolean network in absence of the cell–cell interactions), the score of the set of genes within the sufficient remaining strongly connected components (the sufficient RSCCs) was augmented by one.
(6) The list of the genes of the HN was sorted according to their score. This list is referred to as the 'genes scored list' (GSL).

2.3 GO-based gene classification

To classify genes, the Gene Ontology (GO) annotations of each gene were retrieved using the ID converter tool (http://idconverter.bioinfo.cnio.es/) by using their Entrez serial number. At this point, it was necessary to determine if a GO annotation is related to differentiation. The genes that participate in UQC not only initiate the conversion but also maintain it. Therefore, GO annotations related to cell differentiation and the signaling pathways of the factors initiating differentiations are of interest. To detect the GO annotations related to each gene, a two-step procedure was used:

First, using the CateGOrizer (http://www.animalgenome.org/tools/catego/) tool, GO annotations related to cell differentiation were added. Then, GO annotations related to the signaling pathways of the factors initiating differentiations were sought using the key words 'growth factor' and 'signaling pathway' (all the results were then manually checked to insure that they are related to relevant processes). Taking only the genes related to cell differentiation by the CateGOrizer yielded similar results but with a smaller number of genes. Here, we only report the results for the full list. Results for the list produced directly by the CateGOrizer are given in Supplementary Section S7.

2.4 Statistical methods

Rank test. To check if the order of the GSL is correlated with cellular differentiation, we used the 'gene set enrichment analysis rank test' (Subramanian *et al.*, 2005):

(1) Let q denote the proportion of the genes in the list that are related to cellular differentiation.
(2) Create a vector of weights W, where W(i) is as follows:
   A. 1−q if the gene in that index in the list is related to cellular differentiation.
   B. −q otherwise.
(3) Compute the cumulative sum of W: $S(i) = W(1) + \ldots + W(i)$.
(4) Record the maximum of S.

For all permutation of the list of genes, the last element in S will be zero. To approximate how improbable is the null hypothesis that the two groups are ordered randomly in the list, the maximal value of S was compared with the maximal value distribution of T permuted lists. The one sided *P*-value of the null hypothesis is the fraction of permutations with maximal values equal to or greater than the maximum of the real list.

Hypergeometric distribution. Although the rank test produces a good visualization of how the genes are sorted by participation, the *P*-value retrieved from the maximum of S represents a peak that does not necessarily occur at the beginning of the list. To show that the top of the list sorted by participation is enriched with genes that are related to cell differentiation, we use a threshold that separates the beginning of the list from the rest of the list and compute the fraction of nodes belonging to the group within the upper part of the list. This number is expected to be distributed following a hypergeometric distribution. The threshold that was used here is the top 15% of the list of the genes.

Spearman correlation and Mann–Whitney test. When comparing different scores for the list of genes, a Spearman correlation was used, and when comparing the average score of genes belonging to different sets, a Mann–Whitney test was performed. Non-parametric tests were used, as the distribution of scores was not a normal distribution.

Fraction of differentiation genes—Enrichment test. In each of the previously presented statistical methods, it was assumed that the correlation among the genes is negligible. This may lead to an overestimate of the significance of the results. To test the significance of the correlation in a different way, we computed the fraction of genes with a GO annotation related to differentiation out of all the genes found in the sufficient RSCCs of the conversion. Clearly the conversions are independent observations. The population of scores was compared with the fraction of genes related to differentiation out of the entire network. We then performed a one-sample *t*-test.

2.5 Control networks

To measure the dependence of the algorithm to generic properties of the HN, two types of control networks were examined.

- The first type of networks was produced to maintain the in-degree and out-degree of each node. This type of networks was produced by permuting the edges of the HN by switching between pairs of edges (i,j) and (k,l) and replacing them by (i,l) and (k,j). In a rewired HN, each gene has the same number of activators and inhibitors as in the HN. We validated that the resulting graph is one large strongly connected component like the HN.

Ten such 'rewired-HNs' were created. Each of the 10 represents a different family of Boolean networks. For each of the 10 'rewired-HNs', the routine performed for the HN in Section 2.2.6 was performed as well, i.e. Boolean networks consistent with their wiring and signs were

created and its genes were scored according to their participation in the conversions. Then the list of genes was sorted according to the mean score the gene had in the different 10 'rewired-HNs'.

- Signed scrambled networks were produced by randomly switching the signs of the edges, while keeping the edges themselves. The fraction of activating and inhibiting edges was maintained. Clearly, this control network is more similar to the HN. Similarly, 10 such 'signed scrambled-HNs' were created, and the routine performed for the HN in Section 2.2.6 was performed as well for each of the 10 families of networks. Finally, the list of genes was sorted according to the mean score the gene had in the different 10 'signed scrambled-HNs'.

3 RESULTS

We created Boolean networks consistent with the HN by choosing, randomly for each gene, a Boolean function consistent with the activators and repressors it has according to the Cui *et al.* (2007) network. For each network, we checked whether a conversion between two random cycles could occur. The conversion is a model of differentiation-like events in random Boolean networks (Bodaker *et al.*, 2013). We investigated these simulated differentiation-like events using a set of signaling nodes. Simply stated, a conversion in this context is the transfer of a Boolean network from one attractor to another following signaling from neighboring cells. To determine which genes (nodes in the network) are the most critical for the conversion, a multistep process was performed. First, genes that behave similarly in the two cycles were eliminated (see Section 2). Then we chose within these genes those that could induce a cycle conversion. Among those, we removed the genes residing in loops that do not affect the conversion as explained in Section 2. We computed how often each gene belongs to the resulting set of converting genes in these simulations.

The result of this process is a list of genes ordered by the fraction of simulations where each gene is predicted to participate in signaling that induces cell conversion. This list is referred to as the GSL. If our assumption that conversion is related to the network structure and if our algorithm is correct, we expect the result of this purely theoretical argument to be correlated with the experimental observations of differentiations. To show that this is the case, we analyzed the order of the genes in the GSL.

Among the genes at the top 5% of the GSL are IL4R (interleukin 4 receptor), Mitogen-activated protein/extracellular signal-regulated kinase 1 (MEK1), extracellular signal-regulated kinase 2 (ERK2), epidermal growth factor receptor (EGFR) and (nuclear factor kappa-light-chain-enhancer of activated B cells (NF-kB). These genes play key roles in known important differentiation processes. IL4R is a type I cytokine receptor, that by binding to interleukin 4 promotes the differentiation of T cells to Th2 cells (Nelms *et al.*, 1999). MEK1 and ERK2 are involved in several differentiation processes including neuronal differentiation (Pang *et al.*, 1995). The EGFR is a cell surface receptor for members of the epidermal growth factor family. It is involved in many differentiation processes (Herbst, 2004). NF-kB is a protein complex that controls the transcription of DNA. NF-kB is known not only for activating B cells (Liou *et al.*, 1994) but also for osteogenic differentiation (Cho *et al.*, 2009). The top 5% of the GSL is found in Table 1. The full GSL is given in Supplementary Section S8.

Table 1. Top 20 genes (top~5%) of the 'genes scores list'

Name	Literature	GO annotations	Categorizer	Reference
PKC	+			(Lee *et al.*, 1998)
IL13RA1	+			(Nelms *et al.*, 1999)
IL4R	+	+		(Nelms *et al.*, 1999)
MEK1	+	+	+	(Pang *et al.*, 1995)
ERK2	+	+	+	(Pang *et al.*, 1995)
p53	+	+	+	(Lin *et al.*, 2004)
FAS				
cIAP				
AR				
CDC42	+	+	+	(Deng *et al.*, 2003b)
FADD				
JNK	+			(Rincon and Pedraza-Alva, 2003)
NFkB	+			(Liou *et al.*, 1994) (Cho *et al.*, 2010)
PI3K	+	+		(Okkenhaug and Vanhaesebroeck, 2003)
EGFR	+	+		(Herbst, 2004)
PIB5PA				
ATM				
Calpain1				
ABL1	+			(Era, 2002)
MAPK9	+			(Rincon and Pedraza-Alva, 2003)

Note: The full list is given in Supplementary Section S8. We have used multiple methods to determine whether a gene is related to differentiation. The second column represents genes with published literature describing them as linked to differentiation (see column 5 for typical reference). The third column represents genes with a GO annotation related to differentiation and the fourth column represents genes related to differentiation using the Categorizer automatic GO hierarchy analyzer (http://www.animalgenome.org/tools/catego/).

The top ranking genes are not the only ones related to differentiation. Actually, a clear correlation between the order of genes in the list and the definitions of genes related to differentiation based on their GO classification can be found throughout the list.

We compared the order of genes obtained from the algorithm with genes associated with differentiation using the GO annotation (see Section 2), using four statistical tests. According to the GO-based gene classification, 25% of the studied genes are associated with cellular differentiation. However, when scoring the genes according to the fraction of time they participate in differentiation in the theoretical model (which will be denoted the sorted list hitherto), 43% of the top 15% of the list are related to differentiation ($P = 0.0006$) (Table 2). The results are not sensitive to the cutoff, and a 10/20% cutoff led to $P = 0.006$ and $P = 0.002$, respectively. Similarly, throughout the entire list, genes associated with differentiation have higher scores ($P = 0.005$ for the Mann–Whitney test and $P = 0.0011$ for the gene rank test) (Fig. 1 and Table 2).

The *P*-values of these tests can be affected by correlation between neighboring genes in the network. We thus used an extra test, the enrichment test, which avoids such biases. We computed for each conversion the fraction of nodes with GO annotations related to differentiation out of the nodes in the sufficient RSCCs

Table 2. *P*-values of different lists of genes sorted by different methods

Genes sorted by	Rank test *P*-value	Hypergeometric distribution *P*-value	Mann–Whitney *P*-value
Algorithm (GSL)	0.0011	0.000616	0.0057
Degree	0.0249	0.039529	0.0118
Centrality	0.0119	0.124374	0.0456
Control Networks			
Mean of algorithm of 10 rewired	0.0226	0.198103	0.0623
Mean of algorithm of 10 signs scrambled	0.0002	0.019966	0.0089

Note: The first *P*-value is obtained from the rank test. The second *P*-value is obtained from an analysis of the fraction of positive genes (genes related to differentiation) in the upper part of the list (top 15%). The last column is the *P*-value of a two-sided Mann–Whitney test for order in the list between positive and negative genes. Each row represents another algorithm used to order the genes.

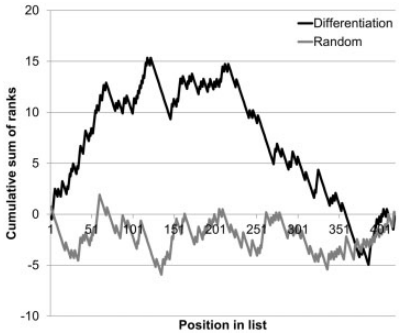

Fig. 1. Rank test for the GSL that are related to cellular differentiation. The *x*-axis denotes the number of genes in the GSL. The *y*-axis denotes the cumulative sums of the ranks for the different genes, where the genes related to differentiation are checked. The black line represents the GSL, whereas the gray line represents a random permutation of the genes. The *P*-value of the rank test of the correlation of the GSL to cellular differentiation is 0.0011

and compared it to the fraction of such nodes in the network. We observed a 26% enrichment in differentiation GO annotations in cycles with nodes inducing differentiation in the model ($P < 1.e\text{-}30$).

3.1 Effect of degree, centrality and signs

The relation between the network-based gene classification and the independent GO clearly shows that the genes found by our algorithm are highly correlated with differentiation. The proposed seemingly theoretical approach actually detects features inherent to stem cell differentiation, and that these features are correlated with properties of the network. The conversion test that we have developed can be determined by complex properties of the network, which are affected by the specific dynamics of the random Boolean network, or it can be based on generic properties such as the degree (The degree of a gene is the number of genes that regulate it plus the number of genes that it regulates.)

Table 3. Spearman correlation between the ordered obtained from the 'GSL' and other scoring methods

Scoring method to correlate	Spearman correlation
Betweenness centrality	0.605044
Degree	0.66787
Mean of algorithm of 10 rewired	0.757244
Mean of algorithm of 10 signs scrambled	0.956352

or the betweenness centrality (The betweenness centrality of a gene is the number of shortest paths from all genes to all others that pass through that gene.).

We checked for the Spearman correlation between the scores of the different genes to their degree and to their centrality (Table 3). A strong correlation exists between the degree and centrality and the UQC participation frequency. To show that the network as an ensemble determines the RSCCs and differentiation and not only the generic properties mentioned earlier in the text, we repeated the statistical tests for the lists of genes sorted according to their degree and centrality (two different lists) and saw that although the degree and centrality are linked to cellular differentiation (or vice versa), their correlation is much weaker than the full algorithm (Table 2).

To classify more precisely the elements correlating the network properties and the experimentally observed differentiation, we produced artificial networks that contain some of the properties of the real network (denoted as control networks). We checked the relation between the order in the GSL in the real HN network and in the control networks. We have produced control networks for the degree and for the effect of the edges signs (activation or inhibition). In each case, 10 networks were created (see Section 2). For each network, the algorithm scored the genes and set a mean score to each gene from these control networks. We checked for the Spearman correlation between the scores of the different genes to the mean scores obtained from the control networks (Table 3). A correlation exists between the scores in the sorted list and in the control networks, and these correlations are higher than the ones obtained for the degree and centrality. Again, when using the enrichment test, there is no enrichment in the fraction of nodes with differentiation GO annotations in cycles inducing differentiation in the control networks.

To summarize, a clear order emerges in the correlation of the theoretical predictions with the GO annotations. The highest correlation is with the real network, followed by the same network with random signs, which itself is followed by a randomized network where edges are swapped, and finally a classification based on the degree and centrality. This is in good agreement with the correlation in the order between the different lists of genes.

3.2 Comparison to other processes

A possible explanation for the correlation between UQC and differentiation is that genes in essential parts of the networks tend to induce UQC, rather than genes specific for

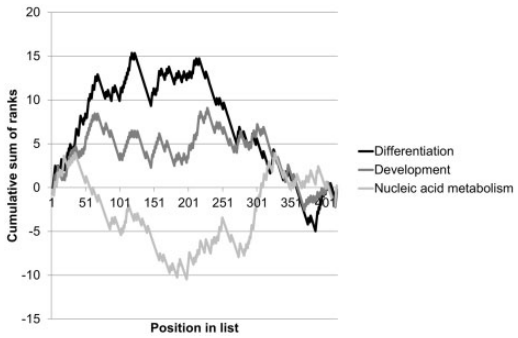

Fig. 2. Rank test for the GSL for different sets of annotations. The *x*-axis denotes the number of genes in the GSL. The *y*-axis denotes the cumulative sums of the ranks for the different genes. The dark gray line denotes the cumulative sums of the ranks for the genes in the GSL that are related to development, whereas the light gray line denotes the cumulative sums of the ranks for the genes in the GSL that are related to nucleic acid metabolism (the black line denotes the cumulative sums of the ranks for the genes in the GSL that are related to differentiation)

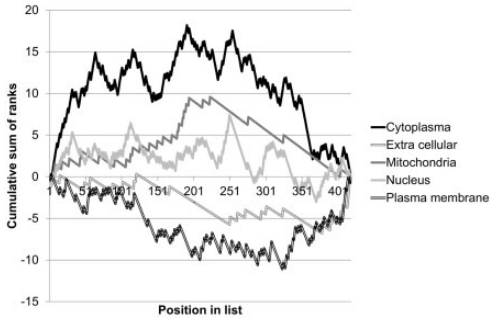

Fig. 3. Rank tests for different subcellular location for the GSL. The *x*-axis denotes the number of genes in the GSL. The *y*-axis denotes the cumulative sums of the ranks for the different genes. The different lines denote the cumulative sums of the ranks for the genes in the GSL that are present in certain subcellular location: hollow dark gray—extracellular, hollow black—plasma membrane, black—cytoplasm, light gray—nucleus and dark gray—mitochondria

differentiation. To show that this is not the case, we have compared the rank test for differentiation with a related set of GO annotations of development and with a non-related set of GO annotations of nucleic acid metabolism, based on the CateGOrizer tool (Fig. 2). The GO annotations related to nucleic acid metabolism have a limited negative correlation with a *P*-value of 0.0445 (derived from the rank test). Thus, choosing any crucial intracellular processes would not lead to similar results.

3.3 Correlation between participation in UQC and subcellular locations

There is an interesting potential correlation between the participation in the sufficient RSCC and the subcellular localization of the gene products. The cellular localization of the gene products was compared with their frequency in conversions. We only looked at the main subcellular locations: extracellular, plasma membrane, cytoplasm, nucleus and mitochondria (Fig. 3 and Table 4). Associating a gene to a subcellular location was performed using the GO annotations and the CateGOrizer tool.

Table 4. Correlations with different subcellular location for the 'GSL'

Sub cellular location	Rank test *P*-value 1	Rank test *P*-value 2	Mann–Whitney *P*-value
Mitochondria	0.0008	0.9816	0.005
Extracellular	0.9774	0.01	0.0282
Plasma membrane	0.9908	0.031	0.0171
Nucleus	0.2488	0.7769	0.4167
Cytoplasm	0.0015	0.9939	0.0004

Note: Rank test *P*-values one and two represent positive and negative correlations, respectively (measured with a rank test). The last column is the *P*-value of a two-sided Mann–Whitney test for order in the list between positive and negative genes.

Table 4 indicates that apart from the nucleus, all the subcellular locations are positively correlated (cytoplasm and mitochondria) or negatively correlated (plasma membrane and extracellular) with participation in conversions.

The positive correlation observed for mitochondria seems strange at first sight. However, all the 27 genes that have GO annotations associated with the mitochondria also have GO annotations associated with the cytoplasm, indicating they are not the standard mitochondrial genes. One possibility is that they are related to apoptosis caused by external signals, and 14 of the 27 genes have GO annotations related to apoptosis. Apoptosis is known to take place in mammalian morphogenesis to eliminate cells during organogenesis, a process which occurs parallel to differentiation (Mori *et al.*, 1995).

Along the same lines, a somewhat unexpected negative correlation was observed for extracellular proteins, which may act as signals, and for plasma membrane proteins, which may act as receptors. However, most gene products that participate in the conversion are neither extracellular signals nor membrane-bound receptors, but instead, are mostly cytoplasmic. This demonstrates that the genes detected by the UQC algorithm belong to the core genetic network rather than to signals that converge to operate it.

3.4 Edges participation in UQC and emerging subgraph

The sufficient RSCCs contain not only genes but the edges between them as well. Each edge has a score reflecting the number of times it participated in the sufficient RSCCs. It is interesting to check among the genes that participate frequently in the sufficient RSCCs what is the emerging subgraph when only edges with high scores are included.

The mean score of the genes included in the sufficient RSCCs is 0.1496 ± 0.0441. The mean score of the number of times an edge participated in sufficient RSCCs is 0.1119 ± 0.0316. A threshold of one mean plus one standard deviation was used for genes, and a threshold of one mean plus one and a half standard deviations was used for the edges, to limit the number of edges. In the emerging subgraph, there are 14 CCs (connected components) and only four small SCCs each of which includes two genes. With the exception of one, all the CCs are small and do not contain a long pathway. The large CC is presented in Figure 4 (the smaller CCs are not shown).

In Supplementary Section S9, the genes of the largest CC are presented with a short description. This large component is

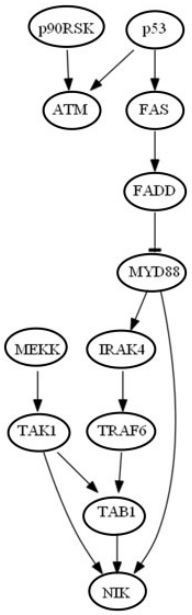

Fig. 4. Largest CC of the emerged subgraph. The sufficient RSCCs contain not only genes but the edges between them as well. Each edge has a score reflecting the number of times it participated in the sufficient RSCCs. The emerged subgraph contains only genes and edges that participated frequently in conversions. Only one large CC (connect component) was found. In Supplementary Section S9, the genes of the largest CC are presented with a short description

composed mainly of mid-degree genes showing again that the degree is not the main element driving the participation in UQC (see second column in Supplementary Table S9). The pathway inside the largest CC begins with genes that are related to apoptosis induction by external signals and continues with genes affecting the differentiation of cells in the hematopoietic system. These genes, known to affect the differentiation of cells in the hematopoietic system, not only have a mid-degree, but they also have a low centrality (the mean of degree is ~7 and the mean of centrality is ~1710). This pathway is one of the most active pathways for cell differentiation in adults. It naturally emerges from the participation frequency in UQC. In other words, the only connected subnetwork purely composed of interactions highly connected to UQC is in the core of the adult hematopoietic cell differentiation, suggesting again that differentiation is actually a global property of the genes associated with it and not purely defined by the biochemical properties of the genes.

4 DISCUSSION

The relation between a specific pathway and differentiation can be the result of the specific interactions in these pathways, and detailed molecular aspects of the component of this pathway, or this relation may be the result of the position of the pathway in a larger network. Although the two explanations are not contradictory, the first one has been classically highlighted. Here, we have used a novel approach to test the second hypothesis. We asked whether the properties of the gene regulatory network affect the genes and their products that are associated with cell differentiation.

We have shown that our methodology that was developed as a theoretical model of random Boolean networks can provide an educated guideline for the detection of genes experimentally observed to participate in cell differentiation events in humans, although the algorithm was never trained on gene classification from experimental data. Using non-parametric statistics (rank test), we have demonstrated that the correlation between the genes that participated in the conversions of the simulations (using our algorithms) to those actually known to date to be involved in cellular differentiation in human cells is statistically significant. We further validated that these results are not exclusively affected by the existing correlation between the participation of genes in conversions and their degree; by using rewired networks as a control, we have shown that the effect of degree is limited. This confirms that the seemingly theoretical approach actually uncovers features in the gene regulatory network that are inherent to stem cell differentiation. We have also tested for a possibility that our results are simply induced by the centrality (which is highly correlated with both the degree and our score), but again the centrality is much less correlated with the GO classification than our score. We did find, however, that the signs of the edges are almost meaningless, implying that the observed sign of the regulation is too partial to lead to useful conclusions at least in the current context. The absence of a sign effect may be the result of a more inherent structure of the network. A basic concept in Boolean networks introduced by Kauffman is the core of the network (Kauffman, 1993). The network's core is the set of nodes found fixed on the same Boolean values in all of the network's cycles. Kauffman studied the emergence of the core and supplied sufficient mathematical theory for its formation. However, there is no good way to estimate *a priori* the dynamics of nodes not found in the core. In the conversions studied here, the nodes change their character between the two cycles and hence are not supposed to be a part of the core. However, we did find nodes in the sufficient RSCCs that according to Kauffman were supposed to be a part of the core. Moreover, in the sufficient RSCCs, we found nodes not fixed in both of the cycles although the conversions are not dependent on the initial steps of the cycles. Thus, the structural element shaping the dynamics of cycles in Boolean networks may extend well beyond the standard definition of the core. One aspect of this modeling may explain why sign-scrambling the edges does not drastically affect the sufficient RSCCs. This may happen, as changing the signs may compensate and preserve a similar core and perhaps even similar sufficient RSCCs. A detailed analysis of the structures beyond the network core will be given in a future work.

The algorithm not only detects genes but also detects the edges between them (interaction between genes tightly linked to differentiation). We built a subgraph from the genes and edges that frequently participated in conversions. The only long pathway detected was found to be one of the most active pathways in differentiation processes of the adult human, affecting the differentiation of cells in the hematopoietic system. This long pathway is composed mainly of mid-degree genes showing again that the degree is not at all the main element driving the participation in conversions. To the best of our knowledge, this work is the first to predict genes that are related to differentiation based only on a gene regulatory network. The presented methodology can be

applied to create a novel bioinformatics tool for gene classification and can also assist experimental biologists that wish to induce differentiation to cells *in vitro* and *in vivo*.

An important result from our previous theoretical work (Bodaker *et al*., 2013) is that conversions are mostly reversible. Unlike the process of iPSCs (induced pluripotent stem cells), where fully differentiated cells become pluripotent by the ectopic expression of key transcription factors (Takahashi *et al*., 2007; Takahashi and Yamanaka, 2006), we suggested (Bodaker *et al*., 2013) that differentiated cells can be used as precursor cells to the tissue stem cells from which they originated by overexpressing relevant genes' products. The presented methodology can be applied to assist experimental biologists that wish to attempt such a process.

Boolean networks were initially developed to study theoretical concepts such as the number of possible attractors in a random network or their properties (Harris *et al*., 2002; Kauffman *et al*., 2003, 2004). Here, we have shown that these networks can have direct practical application for the classification of genes. This is the first time that such applications are presented for a concept that was considered purely theoretical. We have shown that by pure mathematical analysis of well-curated gene networks, we can predict the genes that are associated with differentiation. The algorithms that we have developed will become more and more useful as the gene networks will grow and expand in the future.

ACKNOWLEDGEMENTS

The authors would like to thank Prof. N. Linial from the School of Computer Science and Engineering and Prof. U. Motro from the Institute of Life Sciences, The Hebrew University of Jerusalem, Jerusalem, Israel, for a fruitful discussion and the MOSIX team for providing cluster computing support.

Funding: E.M. would like to acknowledge the European Research Council (281781) and the Israel Science Foundation (657/12).

Conflict of Interest: none declared.

REFERENCES

Ahn,S. *et al*. (2009) Directed mammalian gene regulatory networks using expression and comparative genomic hybridization microarray data from radiation hybrids. *PLoS Comput. Biol*., **5**, e1000407.

Alonso,L. and Fuchs,E. (2003) Stem cells of the skin epithelium. *Proc. Natl Acad. Sci. USA*, **100**, 11830–11835.

Barker,N. *et al*. (2010) Tissue-resident adult stem cell populations of rapidly self-renewing organs. *Cell Stem Cell*, **7**, 656–670.

Bodaker,M. *et al*. (2013) Mathematical conditions for induced cell differentiation and trans-differentiation in adult cells. *Bull. Math. Biol*., **75**, 819–844.

Cho,H.H. *et al*. (2009) NF-kappaB activation stimulates osteogenic differentiation of mesenchymal stem cells derived from human adipose tissue by increasing TAZ expression. *J. Cell. Physiol*., **223**, 168–177.

Cho,H.H. *et al*. (2010) NF-κB activation stimulates osteogenic differentiation of mesenchymal stem cells derived from human adipose tissue by increasing TAZ expression. *J. Cell. Physiol*., **223**, 168–177.

Cui,Q. *et al*. (2007) A map of human cancer signaling, *Mol. Syst. Biol*., **3**, 152.

Deng,M. *et al*. (2003a) Prediction of protein function using protein-protein interaction data. *J. Comput. Biol*., **10**, 947–960.

Deng,X. *et al*. (2003b) Mirk/dyrk1B is a Rho-induced kinase active in skeletal muscle differentiation. *J. Biol. Chem*., **278**, 41347–41354.

Eisen,M.B. *et al*. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.

Era,T. (2002) Bcr-Abl is a "molecular switch" for the decision for growth and differentiation in hematopoietic stem cells. *Int. J. Hematol*., **76**, 35–43.

Gama-Castro,S. *et al*. (2011) RegulonDB version 7.0: transcriptional regulation of *Escherichia coli* K-12 integrated within genetic sensory response units (Gensor Units). *Nucleic Acids Res*., **39**, D98–D105.

Gibbons,F.D. and Roth,F.P. (2002) Judging the quality of gene expression-based clustering methods using gene annotation. *Genome Res*., **12**, 1574–1581.

Harris,S.E. *et al*. (2002) A model of transcriptional regulatory networks based on biases in the observed regulation rules. *Complexity*, **7**, 23–40.

Herbst,R.S. (2004) Review of epidermal growth factor receptor biology. *Int. J. Radiat. Oncol. Biol. Phys*., **59**, S21–S26.

Hishigaki,H. *et al*. (2001) Assessment of prediction accuracy of protein function from protein–protein interaction data. *Yeast*, **18**, 523–531.

Itzhack,R. *et al*. (2013) Long loops of information flow in genetic networks highlight an inherent directionality. *Syst. Biomed*., **1**, 35–34.

Karaoz,U. *et al*. (2004) Whole-genome annotation by using evidence integration in functional-linkage networks. *Proc. Natl Acad. Sci. USA*, **101**, 2888–2893.

Kauffman,S. (1993) *The Origins of Order: Self Organization and Selection in Evolution*. Oxford University Press, New York, NY, USA.

Kauffman,S. *et al*. (2003) Random Boolean network models and the yeast transcriptional network. *Proc. Natl Acad. Sci. USA*, **100**, 14796–14799.

Kauffman,S. *et al*. (2004) Genetic networks with canalyzing Boolean rules are always stable. *Proc. Natl Acad. Sci. USA*, **101**, 17102–17107.

Klesse,L.J. *et al*. (1999) Nerve growth factor induces survival and differentiation through two distinct signaling cascades in PC12 cells. *Oncogene*, **18**, 2055–2068.

Krause,D.S. *et al*. (2001) Multi-organ, multi-lineage engraftment by a single bone marrow-derived stem cell. *Cell*, **105**, 369–377.

Lee,Y.S. *et al*. (1998) Differentiation of cultured human epidermal keratinocytes at high cell densities is mediated by endogenous activation of the protein kinase C signaling pathway. *J. Invest. Dermatol*., **111**, 762–766.

Lin,T. *et al*. (2004) p53 induces differentiation of mouse embryonic stem cells by suppressing Nanog expression. *Nat. Cell Biol*., **7**, 165–171.

Liou,H.C. *et al*. (1994) Sequential induction of NF-kappa B/Rel family proteins during B-cell terminal differentiation. *Mol. Cell. Biol*., **14**, 5349–5359.

Martin,S. *et al*. (2007) Boolean dynamics of genetic regulatory networks inferred from microarray time series data. *Bioinformatics*, **23**, 866–874.

Mori,C. *et al*. (1995) Programmed cell death in the interdigital tissue of the fetal mouse limb is apoptosis with DNA fragmentation. *Anat. Rec*., **242**, 103–110.

Moustakas,A. *et al*. (2002) Mechanisms of TGF-β signaling in regulation of cell growth and differentiation. *Immunol. Lett*., **82**, 85–91.

Nelms,K. *et al*. (1999) The IL-4 receptor: signaling mechanisms and biologic functions. *Annu. Rev. Immunol*., **17**, 701–738.

Okkenhaug,K. and Vanhaesebroeck,B. (2003) PI3K in lymphocyte development, differentiation and activation. *Nat. Rev. Immunol*., **3**, 317–330.

Pang,L. *et al*. (1995) Inhibition of MAP kinase kinase blocks the differentiation of PC-12 cells induced by nerve growth factor. *J. Biol. Chem*., **270**, 13585–13588.

Pittenger,M.F. *et al*. (1999) Multilineage potential of adult human mesenchymal stem cells. *Science*, **284**, 143.

Rincon,M. and Pedraza-Alva,G. (2003) JNK and p38 MAP kinases in CD4+ and CD8+ T cells. *Immunol. Rev*., **192**, 131–142.

Schwikowski,B. *et al*. (2000) A network of protein–protein interactions in yeast. *Nat. Biotechnol*., **18**, 1257–1261.

Serra,R. *et al*. (2007) Why a simple model of genetic regulatory networks describes the distribution of avalanches in gene expression data. *J. Theor. Biol*., **246**, 449–460.

Serra,R. *et al*. (2004) Genetic network models and statistical properties of gene expression data in knock-out experiments. *J. Theor. Biol*., **227**, 149–157.

Sharan,R. *et al*. (2007) Network-based prediction of protein function. *Mol. Syst. Biol*., **3**, 88.

Subramanian,A. *et al*. (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*., **102**, 15545–15550.

Takahashi,K. *et al*. (2007) Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell*, **131**, 861–872.

Takahashi,K. and Yamanaka,S. (2006) Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, **126**, 663–676.

Vazquez,A. *et al*. (2003) Global protein function prediction from protein-protein interaction networks. *Nat. Biotechnol*., **21**, 697–700.

Zhou,X. *et al*. (2002) Transitive functional annotation by shortest-path analysis of gene expression data. *Proc. Natl Acad. Sci. USA*, **99**, 12783–12788.