BOOLEAN NETWORKS: A STUDY ON MICROARRAY DATA DISCRETIZATION

Cyntia Velarde[1] **Cristina Rubio-Escudero**[2] **Rocío Romero-Zaliz**[3]

[1] Dpto. Computación, FCEyN, Universidad de Buenos Aires, Argentina, cvelarde@dc.uba.ar
[2] Dpto. Lenguajes y Sistemas Informaticos, Universidad de Sevilla, España, crubioescudero@us.es
[3] DECSAI, Universidad de Granada, España, rocio@decsai.ugr.es

Abstract

Biomedical research has been revolutionized by high-throughput techniques and the enormous amount of biological data they are able to generate. Genetic networks arise as an essential task to mine these data since they explain the function of genes in terms of how they influence other genes. Genetic networks based on discrete states, such as boolean networks, have been widely used and have shown abilities to model some of the complex dynamics of gene expression networks. In this work we propose a new method for the discretization of gene expression data based on the fuzzification of already proposed techniques. The proposal is applied to the microarray data obtained from a problem on the inflammation and host response to injury in human beings.

Keywords: Microarray, Boolean network, Discretization process.

1 INTRODUCTION

Microarray technology has revolutionized modern biomedical research by its capacity to monitor the behavior of thousands of genes simultaneously [2]. In particular, time-series expression experiments are an increasingly popular method for studying a wide range of biological phenomena. The reconstruction of genetic networks is becoming an essential task to understand the enormous amount of information generated by this high-throughput technique data [4].

Systems biology research arises at this point as the field to explore the life regulation processes in a cohesive way making use of the new technologies. In regulatory networks, proteins have a main role in the regulation of genes [13], but unfortunately, for the vast majority or biological datasets available, there is no information about the level of protein activity. Therefore, we use the expression level of the genes as an indicator of the activity of proteins they generate. Gene networks represent these gene interactions. A gene network can be described as a set of nodes which usually represent genes, proteins or other biochemical entities and the transcription factors that regulate the system.

Given the complexity of this problem, most approaches work on discretized versions of the gene expression data. In fact, discrete state, time and models have been extensively used to model biological networks, and have shown to be able to model, at least in part, the complex dynamics of these networks.

In this paper, we describe a number of methods for the discretization of gene expression data and we propose a new method based on the fuzzification of the already described methods capable to reduce the rate of misdiscretized data. We apply these methods to the microarray data obtained from a problem on the inflammation and host response to injury in human beings.

2 BACKGROUND

In this section we will introduce the discretization methods used in the experimental section and the inference process developed to extract a boolean network from a set of microarray time-series data.

2.1 DISCRETIZATION PROCESS

Several discretization techniques have been used in expression data analysis. These techniques can be grouped in two high level categories: (1) discretization using expression absolute values, and (2) discretization using expression variations between time points [9].

2.1.1 Notation

Let A' be an n row by m column gene expression matrix, where A'_{ij} represents the expression level of gene i under condition j. The matrix A is defined by its set of rows, I, and its set of columns, J. Moreover, let A'_{IJ} denote the average value in the expression matrix A' and A'_{iJ} and A'_{Ij} denote the mean of row i and condition j, respectively. Let let H_{IJ} denote the maximum (high) value in the expression matrix A and H_{iJ} and H_{Ij} denote the maximum value of row i and column j, respectively. In the same way, let M_{IJ} denote the median value in the expression matrix A and M_{iJ} and M_{Ij} denote the median value of row i and column j, respectively.

In this work, we are concerned with applications that use a discretized version of the matrix, where each element in A' is mapped to one element of an alphabet, Σ. Each different symbol represents a distinct activation level. In the simpler case, Σ may contain only two symbols, one used for *regulation* and other for *non-regulation*. In this case, the expression matrix is usually transformed into a binary matrix, where 1 means regulation and 0 means no regulation. After the discretization process, matrix A' is transformed in matrix A and A_{ij} represents the discretized value of the expression level of gene i under condition j.

We define the *gene expression pattern* for gene i as the discretized expression level of gene i under all conditions $j = 1...m$ in matrix A. Note that for gene a and gene b, the expression levels for conditions $j = 1...m$ in matrix A' might result in the same discretized values for conditions $j = 1...m$ in matrix A (Figure 1). Then we say that gene a and gene b *share the same gene expression pattern*.

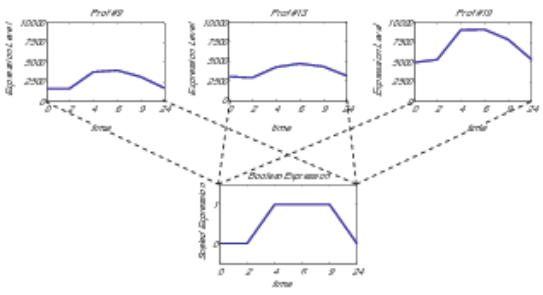


Figure 1: Genes with different expression values at each condition might end up sharing a common expression pattern after the discretization process.

2.1.2 Discretization methods using expression absolute values

Several discretization methods that use expression absolute values [1, 12] will be used in our experiments:

- *A_{IJ} method*: Using the *average* expression value alone, either computed using all the values in the matrix (Equation 1), by row, or by column.

$$A_{ij} = \begin{cases} 1 & if\ A'_{ij} \geq A'_{IJ} \\ 0 & otherwise \end{cases} \quad (1)$$

- *M_{IJ} method*: Using the highest and lowest expression values for each gene i and computing its *median* using all the values in the matrix (Equation 2), by row, or by column.

$$A_{ij} = \begin{cases} 1 & if\ A'_{ij} \geq M_{IJ} \\ 0 & otherwise \end{cases} \quad (2)$$

- *H_{IJ} method*: Using the maximal expression value as the cut-off, observed for the whole matrix, for each gene i or for each condition j. If X is a fixed percentage of this maximal expression value, it can either be computed using all the values in the matrix (Equation 3), by row, or by column.

$$A_{ij} = \begin{cases} 1 & if\ A'_{ij} \geq H_{IJ}(1 - X\%) \\ 0 & otherwise \end{cases} \quad (3)$$

- *TOP X% method*: Using a percentage $X\%$ of the highest values, the expresssion values inside this percentage are discretized to 1 and the remaining to 0. Again, this procedure can be computed using all the values in the matrix (Equation 4), by row , or by column.

$$A_{ij} = \begin{cases} 1 & if\ rank(A'_{ij}, sort(A')) \geq \frac{X \times n \times m}{100} \\ 0 & otherwise \end{cases} \quad (4)$$

where $sort(x)$ returns all elements of x in ascending order, and $rank(x, y)$ returns the position of element x in the list y.

2.1.3 Discretization methods using expression variations between time points

- *Transitional state discrimination (TSD) method* [10]. For this discretization process all values need to be standarized to z-scores (Equation 5):

$$z = \frac{x - \mu}{\sigma} \quad (5)$$

NOT A = A NAND A
A AND B = (A NAND B) NAND (A NAND B)
A OR B = (A NAND A) NAND (B NAND B)

Table 1: Boolean functions obtained only using the $NAND$ function.

where x is a raw score to be standardized, σ is the standard deviation of the population and μ is the mean of the population. The quantity z represents the distance between the raw score and the population mean in units of the standard deviation. If z is negative, then the raw score is below the mean and positive when above.

After this standardization of A' data, each gene expression profile is discretized using two state transitions (Equation 6, $\alpha = 0$).

$$A_{ij} = \begin{cases} 1 & if\ |A'_{ij} - A'_{i(j-1)}| \geq t \\ 0 & otherwise \end{cases} \quad (6)$$

- Another alternative is to first compute the standard deviation for time point 0, $std(0)$ and providing a parameter α to compute a discretization threshold $t = std(0) \times \alpha$ [6] (Equation 6, $\alpha > 0$).

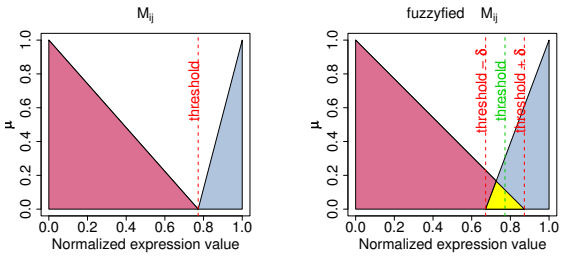
2.2 BOOLEAN NETWORK CONSTRUCTION

A boolean network is composed by a set of nodes n which represent genes, proteins or other biochemical entities. These nodes can take on/off values. The net is determined by a set of at maximum n boolean functions, each of them having the state of k specific nodes as input, where k depends on each node. Therefore, each node has its own boolean function which determines the next state (state at time t_{+1}) based on the actual state (state at time t) of the input nodes. The changes in the net are assumed to occur at discrete time intervals.

The algorithm applied to build the boolean network with our data is the GeneYapay [5]. It performs an exhaustive search of boolean functions over the data, where a number of nodes, less or equal than k, univocally determine the output of some other gene. All possible subsets of $1, 2, ..., k$ elements are visited calculating the number of inconsistencies of the boolean functions in relation to the output value of each gene. The algorithm stops the search for each node when a subset of nodes is found which defines the expression profile. The implementation applied [15] only uses the $NAND$ function since all other boolean function -AND, OR, NOT- can be expressed using $NAND$ (Table 1).

3 PROPOSAL

Crisp discretization methods are widely used in microarray experiments [1, 12, 8, 10]. However, there is a drawback in these methodologies when the expression level is near the threshold value specified by the discretization method. Therefore, several points in the microarray may be misdiscretizied. This is mainly due to the crisp characteristics of the discretization methods (Figure 2 (a)). In this work, we propose an alternative methodology to avoid potential misdiscretization errors by using fuzzy sets (Figure 2 (b)).



(a) (b)

Figure 2: Example of a fuzzyfied discretization method M_{IJ}. (a) Original method. (b) Fuzzified version, where δ is a given small value.

The basic idea is to extend the crisp set (Figure 2 (a)) to a fuzzy set by overlapping both crisp sets using a small value δ as a deviation from the calculated threshold (e.g., $threshold \pm \delta$) (Figure 2 (b)). The data in the intersection of both fuzzy sets (i.e., zero set $ZERO$ and one set ONE) can be discretize to 0 or to 1 as they have almost the same membership μ to both fuzzy sets (Figure 2 (b)). We suggest that expression values with $\mu(ZERO) > 0$ and $\mu(ONE) > 0$ to both fuzzy sets are to be discretized to 0 or 1 depending on their correlation with the elements belonging to the 0 or 1 set. That is, using the Pearson correlation coefficient [11] we calculate if a given expression value is more likely to be similar to the elements in the 0 set (i.e., $\mu(ZERO) > 0$ and $\mu(ONE) = 0$) or to the 1 set (i.e., $\mu(ONE) > 0$ and $\mu(ZERO) = 0$), an discretize it accordingly.

Then, we can redefine the discretization methods introduced in Section 2.1 using this new methodology. We will notate them with an extra tilde (e.g., M'_{IJ}, TSD') to differentiate them from the original versions.

4 EXPERIMENTS

Experiments using all the discretization functions described in Section 2.1 and the new proposed versions

were applied to an immuno-inflammatory response problem. This study, in part carried out at the Cellular Injury and Adaptation Laboratory, Washington University School of Medicine, is a piece of a large-scale research project devoted to profile leukocyte gene expression and plasma proteins of burn and trauma patients [3]. The host response to trauma and burns is a collection of biological and pathological processes that depends critically upon the regulation of the human immuno-inflammatory response. The objective of this study is to identify significant relationships that regulate the integration of the human complex biological system. For this purpose, 48 GeneChips®HG-U133A v2.0 from Affymetrix Inc. were analyzed, derived from samples taken from human blood of eight patients: four treated with intravenous endotoxin and four with a placebo and expression retrieved over time at hours 0, 2, 4, 6, 9 and 24.

The microarray data has been normalized by scaling (Invariant Set Approach) and normalizing (Model-Based Expression Index) [7].

4.1 Results for the original versions

Several interesting results were obtained after discretizing the microarray information with the different methods (Tables 2 and 3). We analyze patient 1 as an example of a treatment patient and patient 5 as an example of a control patient. Also, we analyze the results obtained by the all patient expression patterns.

We can extract from Table 3 that all A_{IJ}, M_{IJ} and H_{IJ} obtain the same amount of patterns. M_{IJ} and H_{IJ} discretization methods have a very similar behavior since they discretize almost all genes in the same fashion since the threshold for each method is almost the same (Table 3). On the other hand, the comparison between the other methods (Table 2) show a high dissimilarity in the discretized patterns. The union of all patients shows similar results than the patients taken independently (Table 2).

Thresholds calculated as a result of the different discretization process are very close to each other (Table 3). Nevertheless, a small difference in the threshold value can affect the discretization of many genes.

The results obtained by the $TOP\ X\%$ discretization methods (Table 3) produce different patterns, nevertheless most of the genes are clustered in two major groups for all $X = 10, 20, 30$, therefore providing a very poor discretization.

Analyzing patterns I and II from Table 4, we can see that $TSD\ \alpha = 0, 1$ discretize them equally, while $TSD\ \alpha = 2$ discretize them differently. We can observe in Figure 3 that their expression levels are not so different. Something similar happens with patterns III and IV. An interesting result is obtained by the discretization of patterns V and VI. We can observe in Figure 3, that they are almost opposite patterns, but $TSD\ \alpha = 1, 2$ discretize them equally, while $TSD\ \alpha = 0$ discretize them correctly.

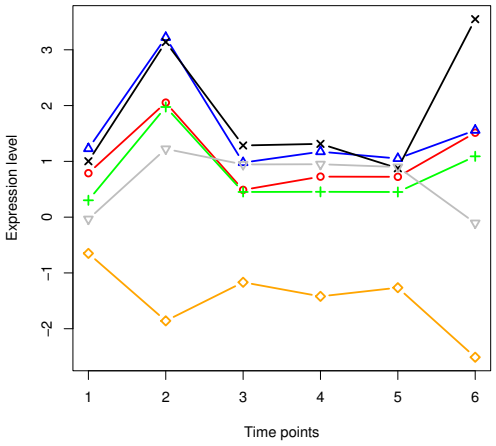


Figure 3: Expression levels for the patterns in Table 4. Color-codes: expression pattern I (red, circle), expression pattern II (blue, triangle), expression pattern III (green, plus sign), expression pattern IV (black, cross), expression pattern V (orange, diamond) and expression pattern VI (gray, upside-down triangle).

Our experiments reveal that the studied discretization methods are sensible to the threshold used for each of them, therefore, data located near this threshold would have high probability of misdiscretization.

4.2 Results for the new versions

We applied the modified versions of the discretization methods introduced in Section 2.1. Due to space restrictions, we will only show the results obtained for method M'_{IJ} for patient 1 as an example. We decided to use a small threshold ($\delta = 0.01$) since greater values will force to unecessarily re-analyze a great amount of discretizations.

The correlation between values inside the threshold and a cluster prototype for each group (i.e., zero set $ZERO$ and one set ONE) are calculated. The discretization of these values is determined by the best correlated group. Each cluster prototype is selected between patters with expression level values outside of the threshold.

Results show that correlation values of the new discretization method is much better (Table 5). We can also see in Figure 4 that the expression levels are more

Table 2: Comparison between discretization methods A_{IJ} (Equation 1), M_{IJ} (Equation 2) and H_{IJ} (Equation 3) described in Section 2.1.2. Each cell shows the number of genes, and its percentage of the total dataset, discretized differently by each pair of methods. All methods applied to the eight patients, here we show only patient 1 (treatment) and patient 5 (control).

METHODS	TREATMENT PATIENT	CONTROL PATIENT	ALL PATIENTS
A_{IJ} vs. M_{IJ}	13752 (62%)	9650 (43%)	78725 (50%)
A_{IJ} vs. H_{IJ}	13769 (62%)	9677 (43.5%)	78878 (50.7%)
M_{IJ} vs. H_{IJ}	106 (0.47%)	126 (0.56%)	826 (0.53%)

Table 3: Discretization results. Discretization methods A_{IJ} (Equation 1), M_{IJ} (Equation 2) and H_{IJ} (Equation 3) described in Section 2.1.2. Discretization method $TOP\ X\%$ (Equation 4) described in Section 2.1.2 with $X = 10, 20, 30$. Discretization method TSD (Equation 6) described in Section 2.1.3 with $\alpha = 0, 1, 2$. All methods applied to the eight patients, here we show only patient 1 (treatment) and patient 5 (control).

	#PATTERNS		
METHOD	TREATMENT	CONTROL	THRESHOLD
A_{IJ}	64	64	0.22242430
M_{IJ}	64	64	0.21305742
H_{IJ}	64	64	0.21303658
TOP 10%	63	64	0.22853785
TOP 20%	55	54	0.23952316
TOP 30%	47	43	0.24580431
$TSD\ \alpha = 0$	64	64	-
$TSD\ \alpha = 1$	34	31	-
$TSD\ \alpha = 2$	29	13	-

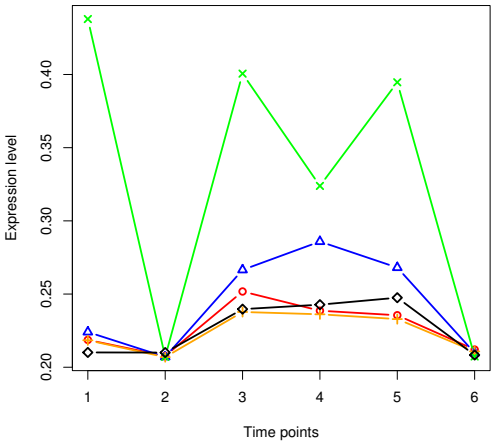


Figure 4: Expression levels for the patterns in Table 5 against M_{IJ} and M'_{IJ} prototypes. Color-codes: expression pattern I (red, circle), expression pattern II (blue, triangle), expression pattern III (orange, plus sign), M_{IJ} prototype (green, cross) and M'_{IJ} prototype (black, diamond).

similar to the new method's prototype than the original method's prototype. Thus, the discretization values will be more appropriate.

As an example of the results obtained by the new methodology proposed, Table 6 shows some genes whose M_{IJ} discretization is corrected by the new M'_{IJ}.

We then calculate the boolean networks for the original M_{IJ} (Figure 5) and the modified M'_{IJ} method (Figure 6). The networks obtained have some portions in common, but also some other protions are quite different. Due to the discretization changes between M_{IJ} and M'_{IJ}, several genes belonging to a pattern in the original method, may now belong to a different pattern in the new method. We studied the biological annotation of every gene inside each of these patterns for both discretization methods using the Onto-CC software [14]. We could see that the new discretization method M'_{IJ} produces more homogeneous sets, that is, the genes in each set have a very simmilar annotation. For instance, pattern G33 (`100001`) have 360 genes in the M_{IJ} method, while the same pattern share the same information plus five extra genes. These five genes are annotated with several terms, like "ubiquitin cycle" and "protein targeting" wich are also associated with the other 360 genes. Therefore, the boolean networks calculated using the new methodology are more reliable since they are constructed using information with

Table 4: Example of genes discretized differently by $TSD\ \alpha = 0$, $TSD\ \alpha = 1$ and $TSD\ \alpha = 2$.

ID	TSD $\alpha = 0$	TSD $\alpha = 1$	TSD $\alpha = 2$	**EXPRESSION VALUES**					
I	010101	011000	000000	0.788009	2.053057	0.488498	0.726193	0.723284	1.513306
II	010101	011000	001000	1.228279	3.221180	0.977528	1.173404	1.050649	1.557317
III	010101	011000	001000	0.302274	1.974590	0.451306	0.454151	0.450831	1.090290
IV	010101	011001	010001	1.002203	3.13927	1.284922	1.314039	0.876443	3.548847
V	101010	010001	000000	−0.649956	−1.859078	−1.167513	−1.42009	−1.26441	−2.511316
VI	110100	010001	000000	−0.035346	1.223311	0.944796	0.948552	0.899278	−0.105929

Table 5: Correlation between genes to M_{IJ} and M'_{IJ}.

ID	**EXPRESSION VALUES**						M_{IJ}	M'_{IJ}
I	0.2186793	0.2076971	0.2517480	0.2386211	0.2355112	0.2120486	0.63664744	0.88130940
II	0.2240480	0.2070605	0.2665070	0.285858p	0.2680730	0.2097774	0.55217325	0.95665902
III	0.2184243	0.2069158	0.2377598	0.2361701	0.2328981	0.2108157	0.67215942	0.93229356

less missclassification errors.

5 CONCLUSIONS

Different discretization methods produce significant differences in the analysis of microarray experiments. Thus, selecting an appropriate method will directly affect the quality of the results. Classical discretization methods show major differences in their results when applied to the same dataset. Moreover, expression levels near the threshold value in each method are questionably discretized and produce potential misdiscretizations. We have proposed new fuzzy based methods that uses correlation to calculate the appropriate discretization values for these dubious expression levels. Thus, the boolean networks are calculated using discretized expression values with a smaller missclassification errors. Further studies with different datasets will provide more detailed analysis.

References

[1] C. Becquet, S. Blachon, B. Jeudy, J-F. Boulicaut, and O. Gandrillon. Strong-association-rule mining for large-scale gene-expression data analysis: a case study on human sage data. *Genome Biology*, 3(12):Electronic publication, 2002.

[2] P. Brown and D. Botstein. Exploring the new world of the genome with dna microarrays. *Nature Genetics*, 21:33–37, 1999.

[3] S.E. Calvano et. al. A network-based analysis of systemic inflammation in humans. *Nature*, 437(7061):1032–7, 2005.

[4] G.W. Carter. Inferring network interactions within a cell. *Bioinformatics*, 6(4):380–389, 2005.

[5] D. D'Onia, L. Tam, J.P. Cobb, and I. Zwir. A hierarchical reverse-forward methodology for learning complex genetic networks. In *Proceedings of the 3rd International Conference on Systems Biology (ICSB), Stockholm, Sweden*, 2003.

[6] S. Erdal, O. Ozturk, D. Armbruster, H. Ferhatosmanoglu, and W. C. Ray. A time series analysis of microarray data. In *Proceeding of the 4rd IEEE Symposium on Bioinformatics and Bioengineering*, pages 366–374, 2004.

[7] C. Li and W. H. Wong. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl. Acad. Sci. USA*, 98(1):31–6, 2001.

[8] S. Lonardi, W. Szpankowski, and Q. Yang. Finding biclusters by random projections. In *Proceedings of the 15th Annual Symposium on Combinatorial Pattern Matching*, pages 102–116. Springer, 2004.

[9] S.C. Madeira and A.L. Oliveira. An evaluation of discretization methods for non-supervised analysis of time-series gene expression data. Technical Report 42, INESC-ID, 2005.

[10] C. Mollr-Levet, S. Cho, and O. Wolkenhauer. Microarray data clustering based on temporal variation: Fcv and tsd preclustering. *Applied Bioinformatics*, 2(1):35–45, 2003.

[11] K. Pearson. Mathematical contributions to the theory of evolution. iii. regression, heredity and

Table 6: Example of genes discretized differently by M_{IJ} and M'_{IJ}.

M_{IJ}	M'_{IJ}	**EXPRESSION VALUES**					
111111	101111	0.273675	0.225677	0.246043	0.250797	0.254966	0.248701
001111	001110	0.212099	0.211830	0.245788	0.245442	0.254923	0.214511
001000	000000	0.212132	0.209505	0.214234	0.212397	0.211857	0.209414

panmixia. *Philos. Trans. Royal Soc. London Ser. A*, 187:253–318, 1896.

[12] R.G. Pensa, C. Leschi, J. Besson, and J. Boulicaut. Assessment of discretization techniques for relevant pattern discovery from gene expression data. In *4th Workshop on Data Mining in Bioinformatics*, 2004.

[13] J. Rice and G. Stolovitzky. Making the most of it: Pathway reconstruction and integrative simulation using the data at hand. *Biosilico*, 2(2):70–77, 2004, 2(2), 70.

[14] R. Romero-Zaliz, C. del Val, J.P. Cobb, and I. Zwir. Onto-cc: a web server for identifying gene ontology conceptual clusters. *Nucleic Acids Res*, 2008.

[15] C. Velarde. Una metodologia para la construccion de redes booleanas basadas en microarrays. Master's thesis, University of Buenos Aires, 2008. In preparation.

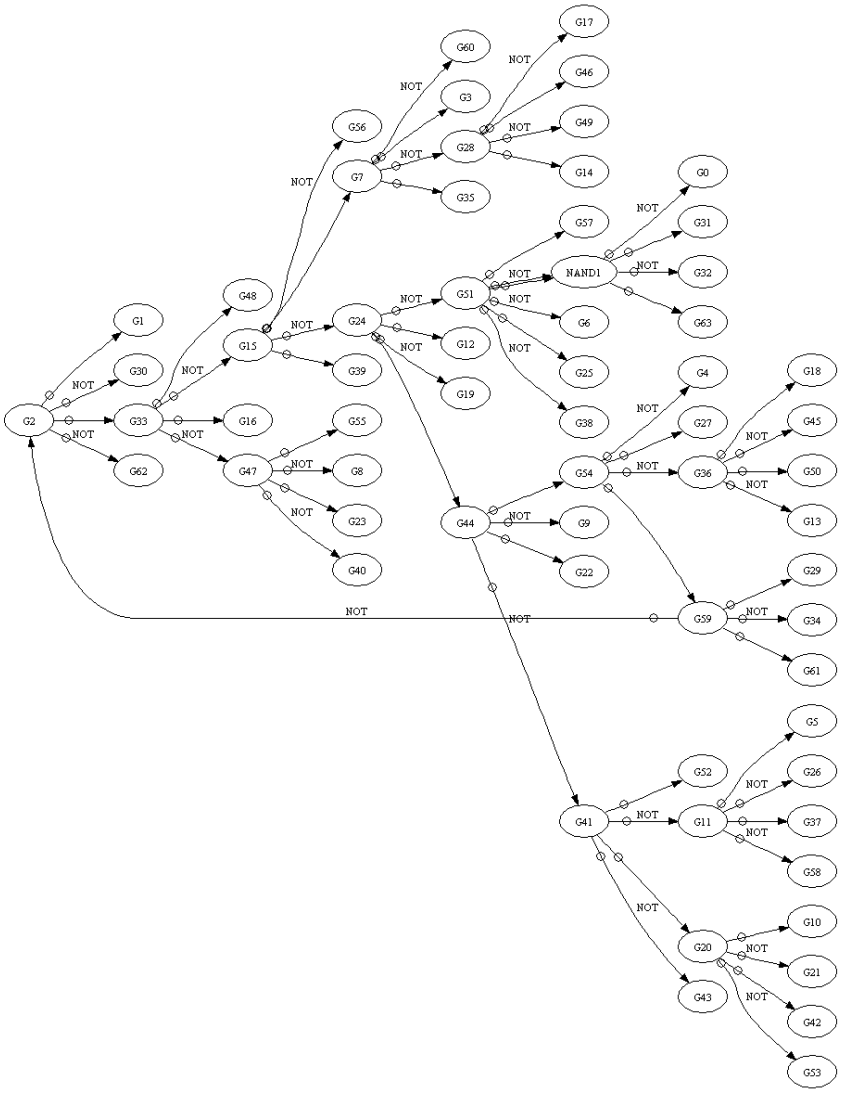


Figure 5: Boolean network obtained for patient 1 using M_{IJ} discretization method. Each node correspond to a an expression pattern.

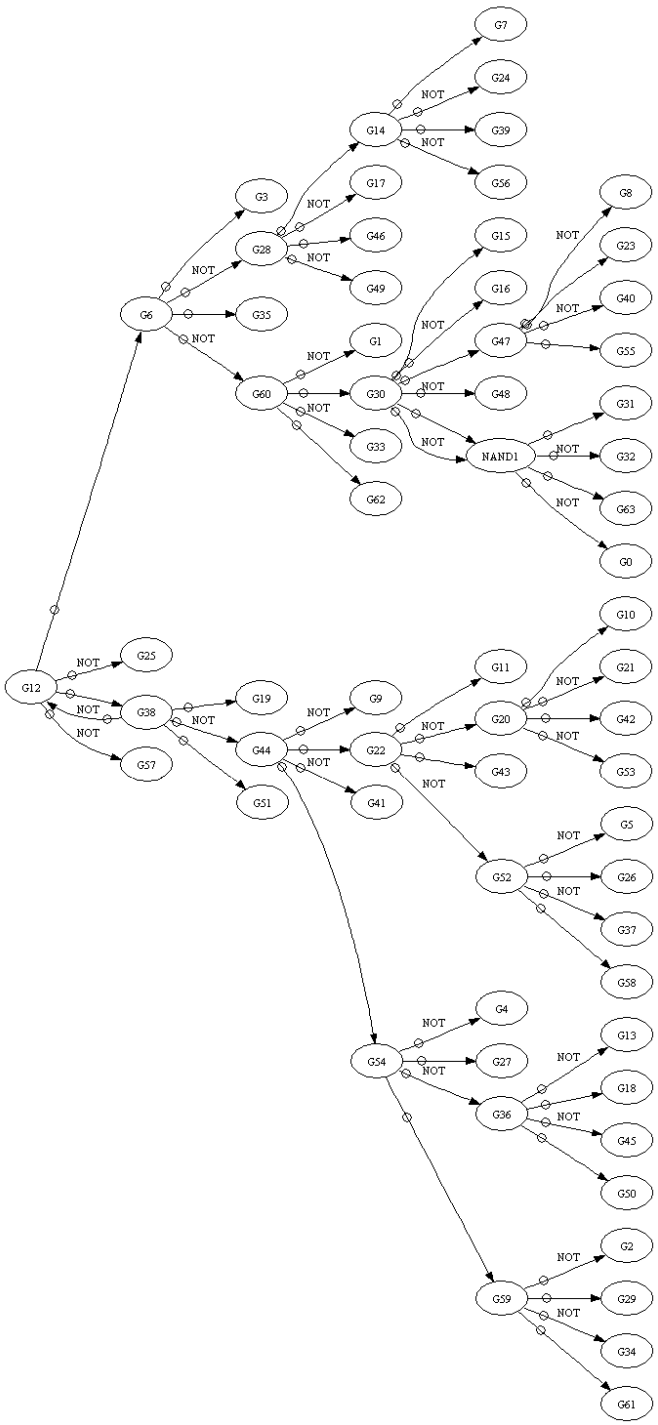


Figure 6: Boolean network obtained for patient 1 using M'_{IJ} discretization method. Each node correspond to a an expression pattern.