Two approaches to reverse engineering of the activating-repressive Boolean gene networks

Evgeny Pshenichnyy, Dmitry Romanov, Natalia Ponomareva and Olga Lyangasova

***Abstract*—** This paper considers two approaches to the problem of reverse engineering of synchronous Boolean gene regulatory networks. Under this model, a gene network is a directed graph with each vertices corresponding to a gene. Each gene is associated with a Boolean variable enoting gene activity state, and with a Boolean function that determines the dependence of the state of a gene in the next moment of time on the state of genes connected with this one in the current moment of time. In this paper we makes an assumption about the nature of connections between nodes in the network based on the biological nature of modeled objects and consider gene networks of activating-repressive type. We propose two algorithms (A-Reverse and N-Reverse) for reverse engineering this specific type of Boolean networks.

***Keywords*—** Bioinformatics, Boolean networks, Reverse engineering.

I. Introduction

Genes function in an ensemble and form gene networks, the coordinated work of which regulates all the processes in the organism and stipulates his phenotypical characteristics. Boolean networks are the simplest models of networks, which, nevertheless, possess system properties similar to those, which are possessed by the real biological networks [1]. Firstly these networks were adopted by Kauffman S. as a model of genetic regulation of the biological processes, proceeding in a cell. Boolean networks as a model of genetic regulation of biological processes is based on hypothesis, such that every gene through the proceeded product under its control may influence the expression of any other genes [11]. Thus, the expression of every other gene depends on the pattern of other genes expression at any time.

To study the behavior of the biological gene networks using the model of boolean networks the Boolean network is created, which is in line with visible data. This task is also known as a task of reconstruction (Inferring) of gene network [13]. The structure recovery of the graph and the Boolean functions is included in the task of inferring of the Boolean networks, related to each of the nodes, conformed to the experimental data about the genes expression of genes. The data about genes expression is given in the form of pair sets of the successive in times states of a boolean network. Unlike these researches[2-4], in the paper it is suggested that the nodes connection properties in the network regards the biological nature of the modeling objects.

The algorithm of reconstruction of the networks is presented in the paper, considering the fact that genes in the gene network can have an activating or suppress impact on each other. The suggested algorithm appeared to be more effective than the universal one for the Boolean gene networks of the activating-repressive type.

II. Boolean Networks And Inference Problem

The Boolean network G is an oriented graph $(X,E),|X|=n$, where each top node consists of a Boolean variable. *A* Boolean function $f_i(x_1,x_2,\ldots x_l)$ is connected to each top node $x \in X$, where $x_i \in X$ are the parameters of those nodes, which have outgoing arcs, ended at X [5].

Values of Boolean variables in the nodes of the graph generate the sequences $x_i(t), t \in Z$, t – discrete time, i – index of the top node in the network. Consider a state of the Boolean network at time t as a set of values of all Boolean variables at the time t $S(t)=(x_1(t),\ldots, x_n(t))$.

The Boolean network passes from the condition $S(t)$ to $S(t+1)$ synchronously, i.e. all the parameter values in the nodes regenerate consistently, regards to the Boolean functions in the nodes:

$$x_i(t+1) = f_i\left(x_{i_1}(t), x_{i_2}(t), \ldots, x_{i_l}(t)\right)$$

The table of states T consists of the pair sets of the input and output states $\{I_i / O_i\}_{i=1}^m$. Such set of the input-output states is considered to be a table of visualized states, or briefly – a table o states 1.

The table of states T can be divided into two parts – inputs and outputs, table 1. Each part consists of n columns. The input columns are to be denoted as i^k, while the output columns - o^k.

The task of the Boolean gene network reconstruction is a task of searching of the Boolean network G, coordinated with the given table of states $T = \{I / O\}_{i=1}^m$.

This work was supported by project of the Ministry of Education and Science of Russia №6.703.2014/K.

Evgeny A. Pshenichnyy is with the Southern Federal University, Research Institute of Biology, Rostov-on-Don, Russia, (corresponding author to provide phone: 007-863-2975070; address: 344090 pr. Stachki 194/1 Rostov-on-Don, Russia; e-mail: pshenichniy.eugene@gmail.com).

Dmitry Romanov is with the Southern Federal University, Research Institute of Biology, Rostov-on-Don, Russia(e-mail: rdme@yandex.ru).

Olga Lyangasova is with the Southern Federal University, Research Institute of Biology, Rostov-on-Don, Russia(e-mail: oll@sfedu.ru).

Natalia Ponomareva is with the Southern Federal University, Research Institute of Biology, Rostov-on-Don, Russia(e-mail: nsponomareva@sfedu.ru).

Table 1. The structure of the table of states.

3	i^2	$\cdots$	i^n	o^1	o^2	$\cdots$	o^n
0	1	0	1	1	0	0	1
1	0	0	1	0	0	0	1
1	0	1	0	1	0	1	0
1	1	0	0	1	0	1	1
0	1	1	0	1	1	0	0
1	0	1	0	1	0	1	1
…	…	…	…	…	…	…	…

III. AR-Boolean Networks

Consider construction of the Boolean network for the cooperation of genes in the process of regulation the transcription. The mechanism of regulation the transcription was widely studied in prokaryotic and eukaryotic organisms. In many cases initiation of the gene transcription happens under control of promoter and many other regulator elements. DNA – sites of coupling proteins recognize these regulatory sequences and have an activating or repressive impact on the expression through the interaction with the promoter and RNA polymerase [6,7].

Consider the gene networks with two types of interactions between genes: activating interaction and repressive interaction. Taking into account this constraint, consider the task of reconstructing for the Boolean networks, where in the nodes of which are only activating – repressive Boolean functions of the type $f(a_1,\ldots a_s,r_1,\ldots,r_t)=(a_1,\ldots a_{s)(}r_1,\ldots,r_t)$. Here, parameters-activators are defined as a, and parameters-repressors are deduced as r. In this case, $s+t=d$, $0<s\leq.$, $t\geq 0$.

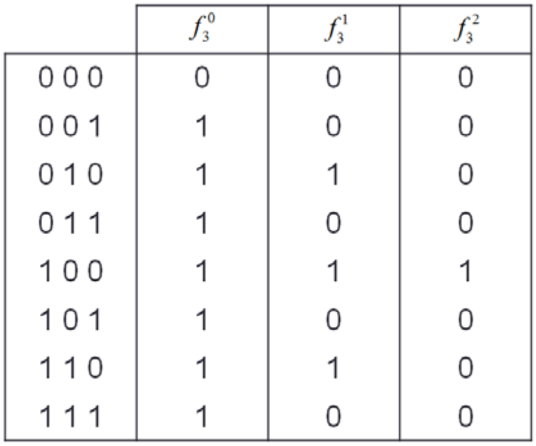
It should be mentioned that the function of the type activator-repressor should have at least one activator and, at the same time it could not have no parameters-repressors. In brief, the function of type activator-repressor is defined as AR-function, and the network with respective functions in the nodes – AR-networks.

In spite of this, activating-repressive Boolean functions satisfy the constraint of the "number of channels" according to Kaufman [1]. This constraint requires, that the definition of the Boolean function can be defined by any of the variables, despite the values of other Boolean functions. Parameters-repressors possess this property. Namely, if at least one repressor is active, it means that the gene would be repressive and the value of the Boolean function would be equal to zero.

The set of parameters-activators of the AR-function f define as *A(f)*, and the set of parameters-repressors - *R(f)*.

In the table 2 is presented the truth table for three Boolean AR functions of three variables with different number of repressors. The first function has no repressors, while the second has one repressor (the first two parameters are active, the third is a repressor), and the third function has only one activator as an argument (the first one) and two repressors (the second and third arguments).

Table 2. The truth table for three AR-functions of three variables



	f_3^0	f_3^1	f_3^2
0 0 0	0	0	0
0 0 1	1	0	0
0 1 0	1	1	0
0 1 1	1	0	0
1 0 0	1	1	1
1 0 1	1	0	0
1 1 0	1	1	0
1 1 1	1	0	0

IV. A-Reverse Algorithm

Provide an algorithm for solving the task of exact reconstruction of the Boolean gene activating-repressor network. Assume that it is given the complete table of states T generated by the unknown activating-repressor gene network of the size n.

A-Reverse:

1. For each input o^j do the step 2.
2. Find the set of activators and repressors:

1.1 Construct the down-sized table Ψ_j from the rows of the table T, for which the value in o^j column is equal to one. The down-sized table for the inputs is created by crossing out all the rows, the output of which has the zero value.

1.2 For each column of Ψ_j compute the number of zeroes and ones.

1.3 Partition the column sets of Ψ_j into three sets:

1) Indexes of the columns with respect to the inputs, from which does not depend the output. The input column in the down-sized table would have equal number of zeroes and ones, so the frequency of occurrence of 1 in these columns would equal to ½.

$$U_j=\left\{i\in\{1,n\},\ fr(i,\Psi_j)=\frac{1}{2}\right\}$$

2) Indexes of the columns of repressors inputs. According to the definition, the inputs of the repressors cannot have 1 in the rows, where the input is equal to 1. In the other words, the columns of the down-sized table with zeroes would hit to the set of parameters-repressors with the sequence of occurrence of 1 in such columns equal to zero.

$$R_j=\left\{i\in\{1,n\},\ fr(i,\Psi_j)=0\right\}$$

3) The columns, which have not hit in the previous sets, are the columns equal to the inputs of the activators.

$$A_j = \{1,n\} \setminus R_j \cup U_j$$

Thus, for each output o^j there would be corresponded two disjoint sets *Aj*,*Rj*. Such range of sets uniquely assigns the activating-repressor gene network.

Evaluate the complexity of the algorithm of the gene A-Reverse network reconstruction. The algorithm consists of the cycle, on the outputs of which the sets of activators and repressors are determined. The cycle on outputs consists of n steps.

At the second stage of the A-Reverse algorithm a table Ψ_j is created and the frequency setting $fr(i^1, \Psi_j)...fr(i^n, \Psi_j)$ is computed. Deduce the number of ones in the column o^j as *q*. For one iteration of the cycle the time proportional to *qn* would be required. In the worst case, the number of ones in the column o^j is equal to 2^n*-1*, that is why in the worst case it is required to calculate the number of ones for n-columns with the length of $2^n - 1$. Hence, for the algorithm execution the time proportional to $n \cdot qn \le n^2 \cdot (2^n - 1)$ would be required. Write down the computational cost of the algorithm:

$$O(n \cdot qn) \approx O(n^2 \cdot (2^n - 1)) = O(n^2 \cdot 2^n)$$

V. NUMERICAL TEST AND COMPARISON WITH REVEAL

The number of numerical test has been carried out to evaluate the difference in productivity of the REVEAL [7] and A-Reverse algorithms. Accomplish this, the reconstruction of the gene networks with different number of genes and various levels of the tope nodes has been obtained. In all the tests the results appeared to be equal and the networks have been reconstructed correctly.

The time results for the corresponding algorithms REVEAL and A-Reverse are presented in the table 3.

Table 3. Time in seconds, taken for reconstruction of the networks using the algorithms A-Reverse and REVEAL

N\D	2		3		4	
	A-Reverse	REVEAL	A-Reverse	REVEAL	A-Reverse	REVEAL
10	0,02	0,33	0,02	1,26	0,01	5,95
12	0,14	2,76	0,12	18,21	0,10	23,69
14	0,48	19,41	0,54	32,22	0,63	149,13
16	2,57	107,98	2,79	420,33	2,98	1352,33
18	14,36	678,38	14,10	2989,08	14,57	9773,00

Both of the algorithms have exponential complexity. However, it is observed that the A-Reverse algorithm runs significantly faster.

VI. RECONSTRUCTION FROM INCOMPLETE TABLE OF STATES

In the previous section the algorithm of reconstruction of the Boolean gene AR-network for the complete table of states was provided. The significant disadvantage of such task formulation was the requirement of large input data and noise sensitivity.

Assume that it is given the incomplete table of states *T* generated by the unknown activating-repressor gene network of the size *n*. Among all of the Boolean AR-networks of the size *n* the AR-network G with maximum input degree D and which is in the best way in accordance with table of states T is to be found:

$$\sum_{j\in[1..n]} d(f^{|R|}_{|A|+|R|}, o^j) \to \min$$

Here *j* denotes index of gene in the network, $f^{|R|}_{|A|+|R|}$ is the Boolean AR-function with arity | A_j|+| R_j|≤*D*, sets A,R consisting of indices of genes and determining the set of activating and repressive arguments; *d* denotes the Hamming distance between two boolean vectors.

It is clear, that the boolean function of every gene can be chosen independently, that is the task can be rewritten in the form:

$$\forall j \in [1..n], d(f^{|R|}_{|A|+|R|}, o^j) \to \min$$

Provide an algorithm for solving of this task. Assume that it is given noisy table of states *T* , generated by the unknown activating-repressor gene network of the size *n*.

N-Reverse

1. For each input o^j do the step 2.
2. Find the set of activators and repressors for output o^j.
    1. Construct the searching space S_p, containing combinations of the sets of activators A_j and repressors R_j,

    $$|A_j|+|R_j|=D$$

    2. Among the elements of S_p the element that minimizes expression $d(f^{|R|}_{|A|+|R|}, o^j) \to \min_{s\in Sp}$ is to be found.

It is clear, that by means of exhaustive search through all of the possible combination of activating and repressive indices we are able to find the best one. This two sets of indices unambiguously define AR-function.

However, the size of searching space rapidly grows with growth of *D* and *n*, that's why instead of examination of all searching space S_p we examine *E* of the best-fitting combinations of activating and repressive indices. The selection of this combinations is based on the quantity of zeros and ones in the truncated tables Ψ derived from complete tables *T* of all possible activating – repressive functions with rank less than *D*.

Provide the algorithm for selection of the E best-fitting combinations of activating and repressive indices.

Heuristics for N-Reverse algorithm

For every output o^j construct S_p and select from it first E elements:

1. Calculate $f_{o^j} = fr(j,O)$, where O - output part of table T;
2. Construct truncated table Ψ_j from rows of table T, for which the value in o^j column equals one.
3. For every column of table Ψ_j calculate the frequency of entries of ones in this column. Construct vector
$$F = (fr(1,\Psi_j),..., fr(n,\Psi_j))$$
4. Construct the set of permissible activating-repressive frequencies:
$$OFR(D) = \left\{ f_i^j,\ i = 1..D,\ j = 1..i-1 \right\},\ f_d^r = 2^{d-r} - 1$$
For every element f_d^r of set OFR(D) generate sequence Sp_d^r of possible combinations of activating and repressive indices based on the set of indices of table Ψ_j. Than compose full sequence S, which will include all elements from all sequences Sp_d^r.
5. Order the sequence S_p The sequence consist of combinations k_1,k_2, k_3 of pairs of sets of activating and repressive indices. Enumerate elements of the sequence S_p in accordance to the growth of the next expression:
$$k_1 \sum_{i \in s[A]} \left| fr\left(A^{s[R]}_{s[A]+s[R]}\right) - fr(i,\Psi_j \right| + k_2 \sum_{i \in s[R]} \left| fr(i,\Psi_j \right| + k_3 \left| fr\left(A^{s[R]}_{s[A]+s[R]}\right) - f_{o^j} \right|$$
6. Retain in the set S_p first E elements and remove the rest.

Parameters k_1,k_2,k_3 influence the way of ordering of the searching space. In the numerical experiments described in the next section this coefficients were chosen to be equal to one.

Thus, with the help of this heuristics the task of reconstruction from the incomplete table may be approximately solved, the table consisting of randomly chosen rows from the complete table of states, generated by activating-repressive network.

VII. NUMERICAL EXPERIMENT FOR ALGORITHM N-REVERSE FOR INCOMPLETE TABLES

To verify the algorithm the series of numerical experiments to reconstruct of gene networks from incomplete tables by the use of N-Reverse algorithm were conducted. During the experiments the time in seconds taken for reconstruction of the networks was measured. Two series of numerical experiments for networks with maximum input degree 2 and 3 were conducted. The size of the network varied from 20 to 50 genes with step 5. The quantity of rows in the incomplete table varied from 100 to 500 with step equal to 100 rows. The results of two series are given in tables 4 and 5 respectively.

From charts above one can see that time taken to reconstruct the network from the same data grows together with the growth of the dimension of the network on equal rate. Moreover, this is true for both the networks with maximum degree 2 and maximum degree 3.

Table 4. Time in seconds, taken for reconstruction of the networks using the algorithms N-Reverse for networks with maximum input degree 3

D=2	**100**	**200**	**300**	**400**	**500**
20	0,31	0,25	0,42	0,49	0,44
25	0,46	0,73	0,81	0,95	0,82
30	1,08	1,16	2,15	2,19	1,93
35	1,92	1,85	2,9	3,15	3,94
40	3,77	4,90	3,94	5,41	7,30
45	4,88	6,46	7,07	8,16	9,65
50	8,76	10,87	11,06	10,15	13,39

Table 5. Time in seconds, taken for reconstruction of the networks using the algorithms N-Reverse for networks with maximum input degree 3

D=3	**100**	**200**	**300**	**400**	**500**
20	4,39	5,42	5,10	4,89	5,86
25	17,24	18,81	19,14	20,72	22,59
30	60,73	59,40	63,52	53,97	53,82
35	173,76	180,78	168,45	130,31	169,41
40	421,54	410,11	396,53	403,35	409,30
45	892,19	660,55	686,57	839,45	647,30
50	1418,16	1353,44	1753,37	1544,11	1562,38

VIII. CONCLUSION

In the paper there have been suggested the method for the Boolean gene networks identification, optimized for reconstruction of the networks with properties appropriate to the biological gene networks. Limitation of the Boolean function in the nodes was based on the works [1, 10], in which there have been presented properties of the biological networks – such that, small maximum level of inputs in the network and activating-repressive relationships between the nodes.

That is why, unlike the universal algorithms, as REVEAL, the A-Reverse and N-Reverse one reconstructs the network, satisfying to the above-described biological properties. It is important to note, that algorithm N-Reverse is capable to reconstruct the network from incomplete expression data [12].

It appears that the suggested methodology allows not only obtaining more accurate networks, in the context of biology, but also accelerating the solution of the reconstruction problem.

References

[1] Kauffman, S. A. Metabolic stability and epigenesis in randomly constructed genetic nets. Journal of Theoretical Biology, 22:437-467.,1969

[2] T. Akutsu, S. Kuhara, O. Maruyama, and S. Miyano. Identification of gene regulatory networks by strategic gene disruptions and gene overexpressions under a Boolean model. Theor. Comput. Sci., 298:235–51, 2003

[3] Martin S. et al. Boolean dynamics of genetic regulatory networks inferred from microarray time series data //Bioinformatics. – 2007. – Т. 23. – №. 7. – С. 866-874.

[4] S. Liang, S. Fuhrman, and R. Somogyi. REVEAL, a general reverse engineering algorithm for inference of genetic network architectures. In Pac. Symp. Biocomputing, volume 3, pages 18–29, 1998

[5] Aluru S., Handbook of Computational Molecular Biology Handbook Taylor & Francis Group, LLC, 2006

[6] Ptashne M. A. Genetic Switch. Oxford: Cell Press, 1986

[7] Ratner VA Genetics, molecular cybernetics // Personalities and Problems, Moscow, Nauka, 2002

[8] Shmulevich I., Lahdesmaki H.The role of certain Post classes in Boolean network models of genetic networks, PNAS September 16, 2003 vol. 100 no. 19 10734-10739

[9] Milo R. et al. Network motifs: simple building blocks of complex networks //Science. – 2002. – Т. 298. – №. 5594. – С. 824-827.

[10] Shmulevich I., Lahdesmaki H. The role of certain Post classes in Boolean network models of genetic networks, PNAS September 16, 2003 vol. 100 no. 19 10734-10739

[11] Lubovac Z., Olsson B., Jonsson P., Laurio K., Andersson M. Biological and statistical evaluation of clusterings of gene expression profiles, Proceedings of the WSES International Conference on Mathematics and Computers in Biology and Chemistry 2001 (MCBC 2001), September 26-30, 2001

[12] Gamalielsson J., Olsson B. On the (lack of) robustness of gene expression data clustering, Proceedings of the WSEAS International Conference on Mathematical Biology and Ecology 2004, August 17-19, 2004

[13] Lei Wang, Gen Qi Xu, Mastorakis N. Inverse Problem of networks–reconstruction of graph, Proceedings of the WSEAS International Conference on New Aspects of Systems Theory and Scientific Computation, 2010