International Journal of Computing Academic Research (IJCAR)
ISSN 2305-9184 Volume 1, Number 2 (December 2012), pp. 51-65

http://www.meacse.org/ijcar

Binary Based Biological Network Model: A Review

Ugege Peter Edemewe[1], Agbaegbu Johnbosco[2] and Akee Sakirat Adedayo[3]

[1]Computer Unit, Forestry Research Institute of Nigeria, Ibadan
[2]School of Computer Science, Mathematics and Information Technology,Houdegbe North American University, Republic of Benin
[3]College of Information and Communication Technology, CrescentUniversity, Abeokuta, Nigeria

Abstract
This article reviews some binary based models, techniques and methods used for clustering biological datasets. Models presented include the deterministic **Boolean Network** and its stochastic extension; **Probabilistic Boolean Network**. Some of the problems associated with these models as well as attempts at resolving them were presented. Their application to the modeling and study of the gene regulatory network and cell differentiation, and evolutionary changes were reviewed.

Keywords: Boolean, Gene, Clustering, Biological Network.

Introduction

Clustering is the task of breaking a set of data into groups in such a way that each group represent proximate collections of data elements based on a distance or similarity function. The clustering of biological data is not only helpful for exploring the data but also to help in discovering hidden links (relationships) between the data. For instance; Biologists often use clustering techniques to identify sets of genes that have similar expression profiles [14]. The performance of a classification model is invariably affected by the characteristics of measurement data it is built upon. If the quality of the data is generally poor, then the classification model will demonstrate poor performance. The amount of noisy instances present in a given dataset is a good reflection of the quality of the data. Detection and removal of noisy data instances will improve quality of the data and consequently the performance of the classification model [15]. A major advantage of using a binary approach is noise resilience and computational efficiency [16, 17]. Hence the binary based models have become one of the most extensively used for clustering and similarity measures. This paper reviews some of these binary based models, techniques and methods, and their application to modeling and analysis of gene regulatory network.

1. The Boolean Model

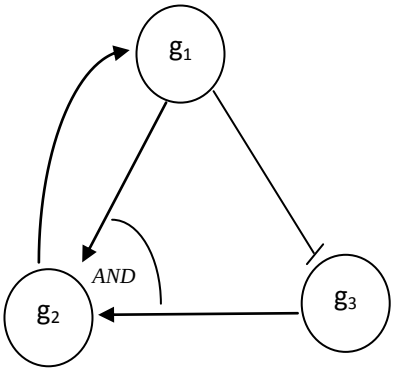
The Boolean model was first introduced by Kauffman as way of modeling the dynamics and evolution of the complex genetic regulatory network [7]. Within this model, there is a link between two genes if the product protein of one gene influences the expression of the other gene. It simplified a gene expression with two levels ON and OFF. A Boolean network is defined by a set of

n elements (genes) $\{g_1, g_2, \ldots, g_n\}$, each $g_i \in \{0,1\}, i = 1, \ldots, n$.The value of g_i at time t+1is determined by the value of its k_i controlling elements $g_{j1}(i),\ g_{j2}(i), \ldots, g_{jk_i}(i)$ at time t. This is represented symbolically as:

$$g_i(t+1) = f_i(g_{j1}(i)^{(t)}, g_{j2}(i)^{(t)}, \ldots, g_{jk_i}(i)^{(t)}),$$

where f_i is a Boolean function associated with the i^{th} element that depends on K_i arguments. Each g_i represents the state (expression) of gene i where $g_i = 1$ means that the gene i is expressed and $g_i = 0$ means it is not expressed. The state of a gene regulatory network containing n entities is then naturally represented as a Boolean vector $(g_i \ldots\ldots\ldots\ldots g_n)$ and this gives us a state space containing 2^n states [11]. The example below throws more light on this definition. Consider the Boolean network in Fig. 1 (a) [12] which contains three entities g_1, g_2, and g_3 where the next state of g_i of each entity is defined by the following Boolean functions:

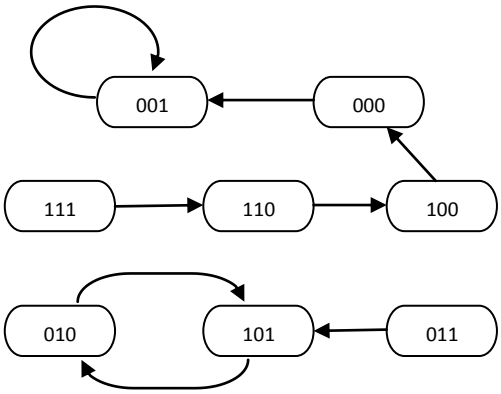
$g'_1 = g_2 \qquad g'_2 = g_1 g_3, \qquad g'_3 = \overline{g_1}$



(a) Boolean network

Current			Next		
g_1	g_2	g_3	g'_1	g'_2	g'_3
0	0	0	0	0	1
0	0	1	0	0	1
0	1	0	1	0	1
0	1	1	1	0	1
1	0	0	0	0	0
1	0	1	0	1	0
1	1	0	1	0	0
1	1	1	1	1	0

(b) Truth table



(c) State transition graph

Fig. 1: An example of a Boolean network for three entities g_1, g_2 and g_3.

The dynamic behavior of a Boolean network can be semantically interpreted in two distinct ways [13]: asynchronously where genes update their state independently; and synchronously where all genes update their state together. Our interest is on the synchronous semantics which appears to be widely used in literatures [13, 21]. The truth table in Fig. 1(b) and a state transition graph [11] in Fig. 1(c) depicts the synchronous behavior of the network. Boolean model has provided conceptual insights into the behavior of genetic regulatory networks [22–24], hence have been used extensively for the analysis of gene expression data. Majority of such approaches use real-valued expression data produced by high throughput screening technologies such as microarrays [4]. Obtaining meaningful information from the observed data, frequently require some measure of similarity. A major difficulty is that the manner in which we quantify the notion of similarity has a most profound effect on the resulting clustering. Furthermore, no theory exists on how to choose the best similarity measure. To address this issue, [4] proposed the analysis of gene expression data entirely in the binary domain, using Hamming Distance as the measure of similarity since it reflects the notion of similarity used by biologist when comparing gene expressions from different tissue samples. That is it reflects the number of genes that significantly disagree in their expression levels. The need to choose a threshold is important for all binary based models like the Boolean model, so that all genes that exceed the threshold get assigned a label 1 and the rest of the genes a label 0.

Since inferring results on the basis of the Boolean model can be affected by the threshold for the binarization of the continuous values, an appropriate threshold need to be selected for inferring an accurate result [5]. Most works never gave a clear explanation on the method of determining this threshold; however few gave such needed explanations. [4] Selected a threshold based on the basic idea that the location of the threshold should be where the separation between low and high expression values is greatest. Referred to as the location, in which the first 'big jump' occurs, that is the point where the finite differences between successive sorted values first exceed some predefined values. The task of choosing a threshold however became confusing by the fact that some arrays may have different overall (average) intensities due to various conditions such as photomultipliers gain or other parameter settings, amount of exposure etc. It is commonly assumed that the sources of error are multiplicative and thus the true expression levels are modified by a multiplicative factor [26]. As a result, any gene belonging to an array with an overall higher intensity has a greater chance of getting set to 1. A data dependent optimization-based normalization procedure has been used to solve this problem [4]. Normalization is a crucial preprocessing procedure that is common to gene expression studies in which data from one array must be compared to data from another array. A number of approaches can be used to normalize the data. The data can be normalized with respect to statistics such as mean, median, maximum and standard deviation [27]. The data could also be normalized with respect to some set of 'house keeping' gene like GAPDH [28]. However, there is equally no currently accepted standard [29], so the chosen method should be motivated by the application at hand and the goals for data analysis. For instance [4] used statistical method based only on the gene expression difference and [5] proposed that genetic interactions be inferred by optimizing the threshold using biological knowledge, to facilitate discovering of new interactions. A common question is whether certain genes, when quantized as binary switches, can be informative in separating phenotype classes such as tumors and normal tissue, as well as different stages of tumor development, depending on the bimodality of their behavior [50]. The efficiency for binary discrimination was revealed for clustering and classification by [4] and [8] respectively.

1.1.Some Boolean Model Application Areas

The Boolean model has found applications in many areas including: modeling gene regulatory network and cell differentiation, and evolutionary changes.

1.1.1. Gene regulatory networks and cell differentiation

The different cells that make up a multi-cellular organism have the same genetic information. The difference between the cells can only be noticed in the genes which are expressed by the cells. For each cell, some genes are turned ON (expressed) while others are turned OFF (not expressed). Gene regulation and control can occur at every stage in the gene expression pathway right from the genetic information contained in the DNA to the translation of this information to protein. However, most of the gene regulation and control appears to be at the level of transcription of the genetic information. At this level, one gene of DNA is transcribed into a molecule of mRNA only if the conditions for such transcription are present. For prokaryotic cells, the transcription of one gene into a molecule of mRNA can only occur if a necessary protein known as "Activator" is present. It attaches to the beginning of the gene indicating that the gene is ready to be transcribed (Fig. 2 (a)). For a gene not ready for transcription, another kind of protein known as "repressor" attaches to the beginning of such gene to inhibit its transcription therefore turning the gene off (Fig. 2(b))

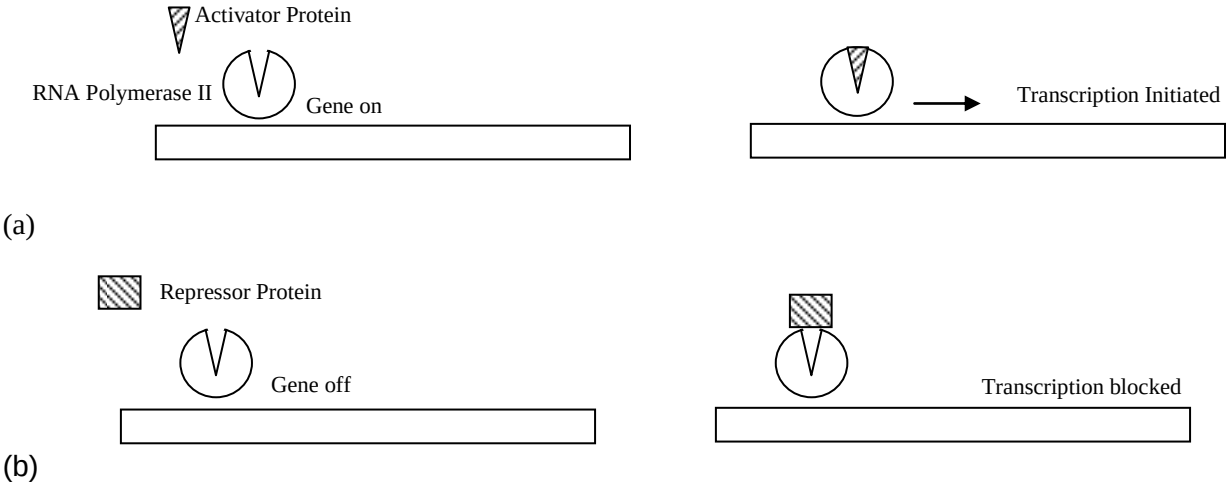


Fig. 2: Diagrammatic representation of gene regulatory proteins. (a) An activator protein attaches to the gene at a specific binding site, activating the polymerase II which then can transcript the gene into a molecule of mRNA. (b) When a repressor protein is attached to the gene, the polymerase II is blocked up and therefore unable to transcript the information contained in the gene.

A more complex situation is noticed in a eukaryotic cell in that many activators or repressors could be required to activate or block the expression of a single gene. The human β-globin gene for instance is regulated by more than 20 different proteins. Hence the expression of one gene is determined by the protein specified by other genes whose expression is in turn controlled by other proteins specified by other genes and so on. The protein specified by the activation of one gene can influence the activation or deactivation of other genes. So also is the absence of the protein that would have been specified by a deactivated gene. So genes are in continuous interaction with each other through the protein they specify. This could be visualized as a complex network of interacting elements (genes) often referred to as **gene regulatory network**. There is rather a wide range of approaches for modeling this network. This include: linear models, Bayesian networks, neural networks, nonlinear ordinary differential equations, stochastic models, Boolean models, logical networks, Petri nets, graph-based models, grammars, and process algebras [1]. Quite a number of publications on modeling and simulation of genetic regulatory network also exist [18 - 20].

1.1.1.1. Obtaining regulatory relationships between genes

Using truth tables to define the Boolean behavior of all entities (genes) in a genetic network, it is possible to extract a compact representation of the regulatory relationships between the entities. Known Boolean logic techniques [11, 34] could be employed to achieve this since they allow us to derive Boolean terms describing the functional behavior of each entity. The idea is to consider the truth table for each entity and to list all the states which result in a next state in which the entity is active (i.e. in state 1) [10]. For instance considering the truth table for g_1 in Fig. 1(b), it can be seen that states 010, 011, 110, and 111 result in g_1 being 1 in its next state. We can represent each state as a product term, called a minterm [34], using the AND Boolean operator, where the variable g_i represents that an entity g_i is in state 1, and the negated variable $\overline{g_i}$ represents that an entity g_i is 0. So the state 010 for g_1 is represented by the minterm $\overline{g_1} g_2 \overline{g_3}$. Applying this approach and then summing the derived minterms using the OR Boolean operator allows us to derive a Boolean term in disjunctive normal form (Sum of Products) [10] that defines the functional behaviour of an entity. Continuing with our example, we derive the following Boolean term for gene g_1:

$\overline{g_1}g_2\overline{g_3} + \overline{g_1}g_2g_3 + g_1g_2\overline{g_3} + g_1g_2g_3$

This term completely defines the functional behavior of g_1, meaning whenever the term above evaluates to 1 in a state we know g_1 will be active in the next state, and whenever the term is 0 we know g_1 will be inactive. Using this technique we can construct a Boolean network that completely specifies the functional behavior of a genetic network. From our example, we derive the following terms defining the behavior of g_1, g_2 and g_3:

$$\overline{g_1}' = \overline{g_1}g_2\overline{g_3} + \overline{g_1}g_2g_3 + g_1g_2\overline{g_3} + g_1g_2g_3$$
$$\overline{g_2}' = g_1\overline{g_2}g_3 + g_1g_2g_3$$
$$\overline{g_3}' = \overline{g_1}g_2\overline{g_3} + \overline{g_1}\overline{g_2}g_3 + \overline{g_1}g_2\overline{g_3} + \overline{g_1}g_2g_3$$

These Boolean terms are often unnecessarily complex therefore requiring some kind of simplification which can be easily achieved using logic minimization [11, 34]. This simplification is of great importance viewing it from the biological point of view as it helps to identify the underlying existing regulatory relationships between entities in the genetic network. The idea behind logic minimization is to simplify Boolean terms by merging minterms that differ by only one variable [10]. Let's consider the term $\overline{g_1}g_2\overline{g_3} + \overline{g_1}g_2g_3$ as an example. The term contains two minterms that differ by only one variable g_3. We can simplify this term by merging the two minterms to produce a simpler term $\overline{g_1}g_2$ which is logically equivalent [11, 34]. Illustrating this idea with the simplification of the terms from our running example we have:

$$\overline{g_1}' = \overline{g_1}g_2\overline{g_3} + \overline{g_1}g_2g_3 + g_1g_2\overline{g_3} + g_1g_2g_3 = \overline{g_1}g_2 + g_1g_2 = g_2$$
$$\overline{g_2}' = g_1\overline{g_2}g_3 + g_1g_2g_3 = g_1g_3$$
$$\overline{g_3}' = \overline{g_1}g_2\overline{g_3} + \overline{g_1}\overline{g_2}g_3 + \overline{g_1}g_2\overline{g_3} + \overline{g_1}g_2g_3 = \overline{g_1}\overline{g_2} + \overline{g_1}g_2 = \overline{g_1}$$

1.1.2 EVOLUTIONARY CHANGES

Variations in the phenotype of organism gradually accumulate, to yield a gradual increase in complexity of form and function [38 – 40]. Simple structures slowly assemble together to form more complex structures, which in-turn assemble to build up even more elaborate systems, and so on. At every stage in the formation of a complex system (organism) out of simpler elements, many sub-systems are created, which represent temporal stable states along the way in the construction of the whole system. Genomes of complex organisms are made up of functional modules of genes, each module encoding the solution of a given biological problem encountered by the species at some point through evolution [37]. New mechanisms for evolutionary processes have been suggested through the study of Boolean networks. [41 – 42] Emphasized, that the numerous changes that occur evolutionary involve reorganizing the same genetic material rather than making it more complex. As pointed out earlier in section 1.0, genes are organized in dynamic complex genetic networks. The search for evolution therefore consists of a search for the most stable organization of genes, which is a search for stable cycles. This means that the evolving systems are not systems with increasing complexity rather they are systems with more complex organizational dynamics which settles down in a finite number of stable attractors. The role of evolution therefore is to search for more stable attractors which the system can fall into [37]. A Boolean network made up of N genes has 2^N states, but the system organizes itself into a much smaller number of attractors. Attractors play very important role in Boolean networks. Given a starting state, within a

limited number of steps, the network will transit into a cycle of states, called an *attractor*, and perturbation will continue to cycle thereafter. Each attractor is a subset of a *basin* comprising those states that lead to the attractor if chosen as starting states. The basins form a partition of the state space for the network. Non-attractor states are transient. They are visited at most once on any network trajectory [50]. When modeling genetic regulatory networks, attractors are often identified with phenotypes [41]. Stability is defined according to the response of the network to perturbations, which can be of three different kinds [37]:

- *Chaotic*, the Hamming distance grows exponentially with time (changes in the states of a few genes by flipping the value of some randomly chosen ones);
- *Frozen*, the Hamming distance decays exponentially with time (permanent changes in the linkages of some genes);
- *Critical*, the Hamming distance evolves temporally (permanent changes in the values of the evolution functions associated with some genes).

The reader is referred to [37] and [43] for more on attractors, cycles and the chaotic, frozen and critical phases of the Boolean network. Only networks in the critical phase have the stability needed to form evolvable systems, since they are able to recover to most of the described mutations. These networks show a very high homeostatic stability, which means that after some perturbation; they are very likely to fall again in the same attractor. As a result, evolution is also interpreted as a sort of search for gene networks possessing stable dynamics and not merely as a searching for stable sub-systems out of which more complicated systems can be built up. The attractors in the chaotic phase are very unstable as any kind of perturbation would shift the system to another attractor. In the frozen phase, though majority of the genes are motionless, little changes in the linkages or in the evolution function would make the system jump to a very different attractor if damages are done to the relevant elements of the network. The fact that more complex living organisms have bigger amounts of genetic material, together with the fact that real genetic networks actually exhibit high homeostatic stability, suggest that evolutionary processes consist of both kinds of "searching," complex hierarchical-modular and dynamical stability[37].

1.1.3. DRAWBACKS OF THE BOOLEAN NETWORK MODEL

The successful use of Boolean networks in modeling real world regulatory networks notwithstanding, [35, 36], a number of shortcomings can be identified with their use: analysis can be problematic due to the exponential growth in Boolean states and the lack of tool support; and they do not cope well with the inconsistent and incomplete data that often occurs in practice. Also, the time complexity problem that arises when trying to construct reliable networks structures with high number of elements (genes) and larger in-degree for the genes is a major drawback. Many attempts have been made to solve these problems. The approaches include: the use of chi square test, Petri nets and Probabilistic Boolean Network [1 - 3].

The Chi Square Test: As already mentioned, one of the major problems with a Boolean network is that it incurs extremely high computation time to construct a reliable network structure. As a result, most Boolean Network algorithms can only be used with a small number of gene (*n)* and low in-degree value *(k)*. All Boolean models still exhibit an exponential increase in computation time for both the parameters *n* and *k* [3]. With emphasis in recent studies on the need to simultaneously consider thousands of genes to be able to construct gene regulatory networks in an organism [44-45], there must be a way to solve the limitation imposed by the use of the Boolean network. This led to [3] proposing the use of Chi Square Test as the variable selection method for the two-way and three-way contingency tables of Boolean count observations. This method showed a significant

reduction in the computation time of the original Boolean model. For example, the computation time of the Chi Square Test is approximately 70.8 times faster than that of the original Boolean network method for a 120 gene network. With an increase in the values of n and k, the improvement is expectantly significantly greater.

Petri nets: Following the problems of analysis as a result of the exponential growth in Boolean states, lack of tool support, data inconsistency and incompleteness that occurs in practice, [10] proposed Petri nets based modeling of gene regulatory networks. The idea was to use logic minimization [11] to extract Boolean terms representing the genetic network's behavior and to then directly translate these into Petri net control structures [10]. The result is a compact Petri net model that correctly captures the dynamic behavior of the original regulatory network and which is agreeable to detailed analysis through existing Petri net tools [49]. The reader is referred to [46 – 49] for details on Petri nets.

2. Probabilistic Boolean Network

The deterministic Boolean-network model has recently been extended to a stochastic framework that allows for different functional relationships and perturbations between states. These stochastic networks are called "probabilistic Boolean networks" [30]. A probabilistic Boolean network maintains the rule-based structure of a Boolean network, and also allows uncertainty [8]. Some studies have demonstrated that intervention can be addressed within the context of probabilistic Boolean networks [31, 32], and has considered optimal time-dependent external control based on dynamic programming [33]. Suppose the necessary technologies for diagnosis, monitoring, and analysis for these kinds of model-based strategies become practically feasible, their successful use would still depend on appropriate binarization of continuous data [8]. In a Boolean network, each (target) gene is "predicted" by several other genes by means of a Boolean function (predictor). Having inferred such function from gene expression data, we could easily decide on the value for the target gene based on the observed values of the predictive genes. However, such deterministic prediction in an environment controlled by a number of biological uncertainties appears problematic as the data used for the inference show uncertainty on several levels [1]. First, genetic regulation exhibits uncertainty at the biological level. Second, microarray data for inferring the model may have errors due to experimental noise in the complex measurement processes. The development of a Probabilistic Boolean Network model resulted from the need to develop a model that could incorporate the stochastic nature of gene expression data. The basic idea in this model is to bring together several promising predictors or Boolean functions with each contributing to the prediction of a target gene. It is the generalization of a Boolean network such that instead of having one Boolean function for each target gene, there are a number of Boolean functions with corresponding prediction abilities; that can be selected randomly with some probabilities. Mathematically, it is a network composed of a set of n elements (genes) $g_1, g_2, \ldots, g_n$, each taking values in a finite set V having d values and a set of vector valued network functions, $f_1, f_2, \ldots, f_r$, governing the state of the genes. There exist a set of state vectors $S = \{x^1, x^2, \ldots, x^m\}$ with $m = d^n$ and $x^k = (x_{k1}, x_{k2}, \ldots, x_{kn})$, where x_{ki} is the value of gene g_i in state k. Each function f_j has n functions β_{ji} $(1 \leq i \leq n)$ and the value of gene g_i at time t+1 given as g_i $(t$+1).

The state space S of the network together with the set of functions, in conjunction with transitions between the states and network functions, determine a Markov chain, the states of the Markov chain being of the form (x^i, f_j). The random perturbation makes the Markov chain ergodic, meaning that it has the possibility of reaching any state from another state and that it possesses a steady-state distribution [50].

3. Shadow Clustering

Shadow Clustering is a rule generation method, based on monotone Boolean function reconstruction, which is able to achieve performances comparable to those of best machine learning techniques [51, 52]. In Shadow clustering, binary strings belonging to the same class and having a closeness determined by a properly defined distance are grouped together. Being binary based, every sequence of bases is first converted to Boolean form before generating set of rules. The standard basis conversion: 'A' = '0111', 'C'
= '1011', 'G' = '1101', 'T' = '1110' is used. This produces a training set for Shadow Clustering containing binary strings with length 4*n*. In a situation whereby the training set have a huge collection of negative sequences, the execution of Shadow Clustering generates a high number of rules, many of which are obtained through specializations of a general consensus pattern. To determine these relationships a proper hierarchical clustering technique is adopted; which can be viewed as a modification of the single linkage algorithm, that takes into account the presence of an ordering among the elements to be clustered, given by the relevance associated with each rule. This technique has been used to classify splice sites of human exons.

4. Data Discretization

Discretization of real data into binary form is an important pre-processing task for the construction of binary based models. It represents the transition of continuous data into discrete form. Problems abound in subsequent steps if this transition is inappropriately implemented. The proper discretization of experimental data into binary form remains the wheel on which any binary based model will successfully run, hence the need for a quick review of this crucial pre-processing step.

Considering three discretization techniques that have been used for encoding the over-expression of genes in [53]: Mid-Ranged, Max – X% Max, and X% Max; [25] defined a new pre-processing technique that supports the evaluation and assessment of different discretization techniques for a given gene expression dataset based on the comparison of dendrograms obtained by clustering various derived Boolean matrices with the one obtained on the raw matrix while using the same clustering algorithm.

Mid-Ranged: The highest and lowest expression values are identified for each gene and the mid-range value is defined. All expression values that are strictly above the mid-range are assigned to value 1 and 0 otherwise.

Max – X% Max: The threshold is fixed with respect to the maximal expression value observed for each gene. From this value, we remove a percentage X of this value. All expression values that are greater than the (100 - X)% of the Max are assigned to value 1 and 0 otherwise

X% Max: For each gene, we consider the situations in which its level of expression is in X% of the highest values. These genes are assigned to value 1, 0 otherwise.

This technique enables choosing the best discretization method and parameters for a given data set. However the best choice of a discretization threshold could be a trade-off between the value for which we get the best similarity score and the value for which the data mining task remain tractable [25]. Choosing the best discretization threshold facilitate the control of trade-off between extraction completeness and noise impact. A major concern with the binary discretization techniques like X% Max and Max – X% Max is that the representation of real-valued data in 0s and 1s (Present/Absent) generally leads to loss of information. This led to the development of discretization techniques that allows multiple states [57-58] since such models can also handle data represented in multiple states [60].

Comparatively not much systematic works has been done to ease the discretization of data for discrete multi-state model based methods [56]. Existing works including [54 -55] assumes larger data sets than those practically available. Recent work by [56] included the development of a new discretization method meant to specifically handle the more experimentally realizable short time data. The method known as short series discretization (SSD) was derived from single-link clustering (SLC) [59] by modifying the algorithm to also address the issues of not having a definite way for selecting the number of discrete classes to be obtained and the choice of an appropriate threshold value.

5. Usefulness of Binary Based Models

Outstanding results from the use of kinetic data for the modeling and analysis of Networks, has placed skepticism on the usefulness of discrete models especially by believers in differential equations. It becomes pertinent therefore to consider a case that showed explicitly the usefulness of such models. Though Network analysis based on quantitative kinetic data has proved powerful and extrapolative [61-64] it requires differential equations with detailed information and high computational power. Also there are often no kinetic data for individual steps in the network thus sometimes presenting an infeasible situation in generating larger scale network simulations. [65-66] demonstrated that kinetic parameters are not required in many networks to describe the overall dynamics of the network. Hence discrete network modeling could provide a solution to the problem. Since Boolean models (an instance of discrete model) characterize nodes with two qualitative states and use just three operators (OR, AND, NOT) there is an extensive reduction in the computational power required to run Boolean simulation when compared to the requirement in kinetic network models [67 -68].

5.1.Modeling the Ca^{2+} signaling network

The Ca^{2+} signaling network, even within a single cell, may contain hundreds or even thousands of different nodes (proteins, small molecules, lipids, ions, etc). Successful attempts have been made to model Ca^{2+} signaling networks using kinetic data from biochemical experimentation [61-62, 70-71, 9]. For example [6] demonstrated that by building a library of small signaling network modules (e.g. PLC, MAP kinase, PKC, PLA_2, etc), a network could be built to simulate multiple functional outputs such as intracellular Ca^{2+} concentration, kinase activity, phosphatase activity, and so on. In an attempt to establish the usefulness of discrete models, [69] investigated whether Boolean modeling can be used to study phospholipase C-coupled calcium (PLC-coupled Ca^{2+}) signaling pathways. The results they generated where very convincing about the usefulness of discrete models. Their results demonstrate that:

(i) Ca^{2+} networks can be constructed from legacy knowledge
(ii) Experimental data from these multiple sources can be used to define network directionality and Boolean update rules for each node in network, and
(iii) Network modeling based on these Boolean rules provided a descriptive and predictive model of PLC-mediated Ca^{2+} signaling.

Based on these results, it can be said that with sufficient legacy knowledge and/or computational biology predictions, Boolean networks provide a robust method for predictive-modeling of any biological system.

Conclusions and Future Research

In this paper, we reviewed binary based models, techniques and methods used for biological interaction network. Some of the drawbacks of these models and techniques; as well as attempts at

overcoming them were equally reviewed. Also we tried to convince further on the usefulness of discrete models by presenting a recent work by [69] that showed that Boolean modeling can be used to study phospholipase C-coupled calcium (PLC-coupled Ca^{2+}) signaling pathways. From the review, Binary based models were discovered to be noise resilient hence making them suitable for integrative analysis of biological data scattered over different storage sources and in different formats. They handle quite well the high degree of noise in biological datasets. The use of binary based models has also found good application in the modeling of host – pathogen interactions which had led to the discovery of some hidden knowledge about some disease causing pathogens like Mycobacterium tuberculosis. In the nearest future we intend to study these models more with the intention of developing new techniques for using them to predict Proteins and genes functions from Binary based Protein interaction/gene interaction networks. We certainly could not have covered all binary based models and methods in this short review; we are recommending a more encompassing review to cover such models, techniques and methods not included in this review.

References

[1] I. Shmulevich, E.R. Dougherty, W. Mang. "From Boolean to probabilistic Boolean networks as models of genetic regulatory networks". Proceedings of the IEEE Volume: 90 Issue: 11 pp. 1778-1792 NOV 2002

[2] L. J. Steggles, R. Banks and A. Wipat. "Modelling and Analysing Genetic Networks: From Boolean Networks to Petri Nets". University of Newcastle upon Tyne Technical Report Series No. CS-TR-962 May, 2006

[3] H. Kim, J. K Lee and T. Park. "Boolean networks using the chi-square test for inferring large-scale gene regulatory networks". *BMC Bioinformatics* 2007, pp.8:37

[4] I. Shmulevich and W. Zhang. "Binary analysis and optimization-based normalization of gene expression data". Bioinformatics, 18,pp. 555–565, 2002.

[5] K. Hakamada, T. Hanai, H. Honda, and T. Kobayashi. "A preprocessing Method for Inferring Genetic Interaction from Gene Expression Data Using Boolean Algorithm". Journal of Bioscience and Bioengineering Vol. 98, No. 6, pp.457–463. 2004.

[6] U. S. Bhalla and R. Iyengar, "Emergent properties of networks of biological signaling pathways," Science, vol. 283, no. 5400, pp. 381-387, Jan.1999.

[7] S. A. Kauffman, "Metabolic stability and epigenesis in randomly constructed genetic nets," J. Theoret. *Biol.*, vol. 22, pp. 437–467, 1969.

[8] X. Zhou, X. Wang, and E. R. Dougherty: "Binarization of Microarray Data on the Basis of a Mixture Model". Molecular Cancer Therapeutics Vol. 2, 679–684, July 2003

[9] Y. Tang, "Simplification and analysis of models of calcium dynamics based on IP3-sensitive calcium channel kinetics," Biophys. J., vol. 70, no. 1, pp. 246, 1996.

[10] L. J. Steggles, R. Banks and A. Wipat. "Modelling and Analysing Genetic Networks: From Boolean Networks to Petri Nets". University of Newcastle upon Tyne, Computing Science Technical Report Series. No. CS-TR-962 May, 2006

[11] K. J. Breeding. "Digital Design Fundamentals". Prentice Hall, 1992.

[12] T. Akutsu, S. Miyano and S. Kuhara. "Identification of genetic networks from small number of gene expression patterns under the Boolean network model". Proceedings of Pacific Symp. on Biocomputing, 4,pp.17-28, 1999.

[13] C. Gershenson. "Classification of random boolean networks". In: R. K. Standish et al (eds), Proc. of the 8th Int. Conf. on Artificial Life, p.1–8, MIT Press, 2002.

[14] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. "Cluster analysis and display of genome-wide expression patterns". Proc. Natl. Acad. Sci. USA, 95:14863-14868, December 1998.

[15] T. M. Khoshgoftaar, N. Seliya, and K. Gao. "Detecting noisy instances with the rule- based classification model". Intelligent Data Analysis Volume 9, Number 4/2005 pages 347-364

[16] P. M. Bowers, B. D. O'Connor, S. J. Cokus, E. Sprinzak, T. O. Yeates, and D. Eisenberg. 2005. "Utilizing logical relationships in genomic data to decipher cellular processes". FEBS J. 272:5110-5118.ou

[17] I. Shmulevich, and W. Zhang. 2002. "Binary analysis and optimization-based normalization of gene expression data". Bioinformatics. 18:555-565.

[18] P. Smolen, D. Baxter, and J. Byrne, "Mathematical modeling of gene networks," *Neuron*, vol. 26, pp. 567–580, 2000.

[19] J. Hasty, D. McMillen, F. Isaacs, and J. J. Collins, "Computational studies of gene regulatory networks: *In numero* molecular biology,"*Nature Reviews Genetics*, vol. 2, pp. 268–279, 2001.

[20] H. de Jong, "Modeling and simulation of genetic regulatory systems: A literature review," *J. Comput. Biol.*, vol. 9, no. 1, pp. 69–103, 2002.

[21] J. M. Bower and H. Bolouri. "Computational Modelling of Genetic and Biochemical Networks". MIT Press, 2001.

[22]. R. Somogyi, and C. Sniegoski. "Modeling the complexity of gene networks: understanding multigenic and pleiotropic regulation". Complexity, *1:* 45–63, 1996.

[23]. A. Wuensche. "Genomic regulation modeled as a network with basis of attractions". Pacific Symp. Biocomput., *3:* 89–102, 1998.

[24]. R. Thomas, D. Thieffry, and M. Kaufman. "Dynamical behavior of biological regulatory networks - 1. Biological role of feedback loops and practical use of the concept of the loop-characteristic state". Bull. Math. Biol., 57, pp.257–276, 1995.

[25] R. G. Pensa, C. Leschi, J. Besson and J. Boulicaut. 2004. "Assessment of discretization techniques for relevant pattern discovery from gene expression data". BIOKDD04: 4th Workshop on Data Mining in Bioinformatics (with SIGKDD Conference) pp.24-30.

[26] D. G. Hartemink, I. Jaakkola and R. Young (2001). "Maximum likelihood estimation of optimal scaling factors for expression of array normalization". Microarrays: Optical Technologies and Informatics (Proceedings of SPIE, pp. 4266

[27] T. R. Golub, D. K Slonium, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. Loh, J. R. Downing, M. A. Calinguiri, C. D. Bloomfield, and E. S. Lander (1999). "Molecular Classification of cancer" Class discovery and Class prediction by gene expression monitoring Science, 286, pp.531 - 537

[28] S. Kim, E. R. Dougherty, Y. Chen, K. Sivakumar, P. Meltzer, J. M. Trent, and M. Bittner. (2000) "Multivariant measurement of gene expression relationships". Genomics, 67, pp.201 - 209

[29] J. E. Celis, M. Kruhoffer, I. Gromova, C. Frederiksen, M. Ostergaard, T. Thykjaer, P. Gromova, J. Yu. H. Palsdottir, N. Magnusson, and O. orntoft (2000) "Monitoring transcription and translation products using DNA Microarrays and Proteomics". FEBS Lett., 480, pp.2 – 16.

[30] I. Shmulevich, E.R. Dougherty, S. Kim, and W. Zhang. "Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks". Bioinformatics, 18,pp. 261–274, 2002.

[31] I. Shmulevich, E.R. Dougherty, and W. Zhang. "Gene perturbation and intervention in probabilistic Boolean networks". Bioinformatics, 18.pp.1319–1331, 2002.

[32] I. Shmulevich, E.R. Dougherty, and W. Zhang. "Control of stationary behavior in probabilistic Boolean networks by means of structural intervention". Biol. Systems, *10,pp.* 431–446, 2002.

[33] A. Datta, A. Choudhary, M. L. Bittner, and E. R. Dougherty. "External control in Markovian genetic regulatory networks". Machine Learning, 52,pp.169–181, 2003.

[34] P. Grossman. "Discrete Mathematics for Computing". Palgrave MacMillan, Second Edition, 2002.

[35] S. Huang. "Gene expression profiling, genetic networks, and cellular states: an integrating concept for tumorigenesis and drug discovery". Journal of Molecular Medicine, Vol. 77,pp.469-480, 1999.

[36] Z. Szallasi and S. Liang. "Modeling the Normal and Neoplastic Cell Cycle with "Realistic Boolean Genetic Networks": Their Application for Understanding Carcinogenesis and Assessing Therapeutic Strategies". Pacific Symposium on Biocomputing Vol. 3, pp. 66–76, 1998.

[37] M. Aldana, S. Coppersmith and L. P. Kadanoff. "Boolean Dynamics with Random Couplings". arXiv:nlin/0204062v2 [nlin.AO] 29 April 2002

[38] H. A.Simon, 1969. "The Sciences of the Artificial". The MIT Press, Cambridge, MA.

[39] R. Dawkins, 1986. "The Blind Watchmaker". W.W. Norton and Company, USA.

[40] R. Dawkins, 1989. "The Selfish Gene". Oxford University Press, Oxford, second edition.

[41] S. A. Kauffman, 1993. "The Origins of Order: Self-Organization and Selection in Evolution". Oxford University Press, Oxford.

[42] S. A. Kauffman, 1995. "At Home in the Universe: the Search for Laws of Self-Organization and Complexity". Oxford University Press, Oxford.

[43] F. Karlsson, and M. Ho¨rnquist. "Order or chaos in Boolean gene networks depends on the mean fraction of canalizing functions". Physica A 384 (2007) pp.747–757

[44] J. D. J. Han, N. Bertin, T. Hao D. S. Goldberg, G. F. Berriz, L. V. Zhang, D. Dupuy, A. J. M. Walhout, M. E. Cusic, F. P. Roth M. Vidal. "Evidence for dynamically organized modularity in the yeast protein-protein interaction network". Nature 2004, 430,pp.88-93.

[45] H. Jeong, B. Tombor , R. Albert , Z. N. Oltvai, A. L. Barabasi. "The large scale organization of metabolic networks". Nature 2000, 407,pp.651-654.

[46] T. Murata. "Petri nets: properties, analysis and applications". Proceedings of the IEEE, 77(4),pp.541–580, 1989.

[47] W. Reisig. "Petri Nets, An Introduction. EATCS Monographs on Theoretical Computer Science, W.Brauer et al (Eds.)" Springer–Verlag, Berlin, 1985.

[48] W. Reisig and G. Rozenberg. "Lectures on Petri Nets I: Basic Models. Advances in Petri Nets, Lecture Notes in Computer Science 1491", Springer-Verlag, 1998.

[49] Petri nets World, http://www.informatik.uni-hamburg.de/TGI/PetriNets/, 2006.

[50] E. R. Dougherty and Y. Xiao. "Design of Probabilistic Boolean Networks Under the Requirement of Contextual Data Consistency". IEEE Transaction on Signal Processing, VOL. 54, NO. 9, SEPTEMBER 2006

[51] M. Muselli, A. Quarati "Reconstructing positive Boolean functions with Shadow Clustering". In *Proceedings of the 17th European Conference on Circuit Theory and Design (ECCTD 2005)*, (Cork, Ireland, August 2005).

[52] M. Muselli "Switching neural networks: A new connectionist model for classification". In *WIRN/NAIS 2005*, vol. **3931** of *Lecture Notes in Computer Science* (2006) Eds. B. Apolloni, M. Marinaro, G. Nicosia, R. Tagliaferri, Berlin: Springer-Verlag, pp.23–30.

[53] 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, 12, November 2002.

[54] B. Di Camillo, and F. Sanchez-Cabo. 2005. "A quantization method based on threshold optimization for microarray short time series". BMC Bioinform. Suppl 6, S11

[55] J. Ernst, and Z. Bar-Joseph. 2006. "STEM: a tool for the analysis of short time series gene expression data". BMC Bioinform. 7, 191.

[56] E. S. Dimitrova, M. P. V. Licona, J. Mcgee, and R. Laubenbacher. 2010. "Discretization of Time Series Data". Journal of Computational Biology Volume 17

[57] D. Thieffry and R.Thomas 1998. "Qualitative analysis of gene networks". Proc. Pac. Symp. Biocomput,pp.77–88.

[58] R. Laubenbacher, and B. Stigler 2004. "A computational algebra approach to the reverse engineering of gene regulatory networks". J. Theor. Biol. 229, pp.523–537

[59] A. Jain, and R. Dubes. 1988. "Algorithms for Clustering Data". Prentice Hall, Englewood Cliffs, NJ.

[60] G. Glazko, M. Coleman and A. Mushegian. "Similarity searches in genome-wide numerical data sets". *Biology Direct* 2006, 1:13

[61] M. R. Maurya and S. Subramaniam, "A kinetic model for calcium dynamics in RAW 264.7 cells: 1. Mechanisms, parameters, and subpopulational variability," Biophys. J., vol. 93, no. 3, pp. 709-728, Aug.2007.

[62] M. R. Maurya and S. Subramaniam, "A kinetic model for calcium dynamics in RAW 264.7 cells: 2. Knockdown response and long-term response," Biophys. J., vol. 93, no. 3, pp. 729 - 740, Aug.2007.

[63] G. von Dassow, "The segment polarity network is a robust developmental module," Nature, vol. 406, no. 6792, p. 188, 2000.

[64] D. H. Sharp and J. Reinitz, "Prediction of mutant expression patterns using gene circuits," Biosystems, vol. 47, no. 1-2, pp. 79-90, June1998.

[65] M. Chaves, E. D. Sontag, and R. Albert, "Methods of robustness analysis for Boolean models of gene control networks," Syst. Biol (Stevenage.), vol. 153, no. 4, pp. 154-167, July2006.

[66] R. Albert and H. G. Othmer, "The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in Drosophila melanogaster," Journal of theoretical biology, vol. 223, no. 1, pp. 1-18, July2003.

[67] R. Albert, "Scale-free networks in cell biology," J. Cell Sci, vol. 118, no. Pt 21, pp. 4947-4957, Nov.2005.

[68] M. Chaves, R. Albert, and E. D. Sontag, "Robustness and fragility of Boolean models for genetic regulatory networks," J. Theor. Biol., vol. 235, no. 3, pp. 431-449, Aug.2005.

[69] G. Bhardwaj, C. P. Wells, R. Albert, D. B. van Rossum, and R. L. Patterson, "Mapping Complex Networks: Exploring Boolean Modeling of Signal Transduction Pathways" Physics Archives, November 2009.

[70] Y. X. Li, "Equations for InsP3 receptor-mediated [Ca^{2+}]i oscillations derived from a detailed kinetic model: a Hodgkin-Huxley like formalism," Journal of theoretical biology, vol. 166, no. 4, pp. 461, 1994.

[71] R. H. Chow, "Cadmium block of squid calcium currents. Macroscopic data and a kinetic model," The Journal of General Physiology, vol. 98, no. 4, pp. 751-770, Oct.1991.