Journal of
Computer Science & Systems Biology
ISSN: 0974-7230

Mandal et al., J Comput Sci Syst Biol 2013, 6:5
DOI: 10.4172/jcsb.1000121

Review Article Open Access

Reconstruction of Dominant Gene Regulatory Network from Microarray Data Using Rough Set and Bayesian Approach

Sudip Mandal[1]*, Goutam Saha[2] and Rajat Kr Pal[3]

[1]*ECE Department, Global Institute of Management and Technology, Krishna Nagar, Nadia, West Bengal, India*
[2]*IT Department, North-Eastern Hill University, Shillong, India*
[3]*CSE Department, University of Calcutta, Kolkata, India*

Abstract

Biological databases, containing genetic information of patients, are undergoing tremendous growth beyond our analysing capability. However such analysis can reveal new findings about the cause and subsequent treatment of any disease. Interactions between genes and the proteins they synthesize shape Genetic Regulatory Networks (GRN). In this context, it has been developed a model capable of representing small dominant GRN, combining characteristics from the Rough Set and Bayesian Network. The investigation has been carried out on the publicly available microarray dataset for Lung Adenocarcinoma, obtained from the National Center for Biotechnology Information (NCBI) website. The analysis revealed that Rough Set Theory (RST) is able to extract the various dominant genes in term of reducts which play an important role in causing the disease and also able to provide a unique simplified rule set for building expert systems in medical sciences with high accuracy and coverage factor. The next part of this work is based on reconstruction of GRN using Bayesian network, which is a mathematical tool for modelling conditional independences between stochastic variables like different gene expression. This proposed Bayesian approach using scaled mutual information for scoring is applied to the dataset corresponding to most dominant responsible genes for Adenocarcinoma to uncover, gene/protein interactions and key biological features of the cellular system. Finally different interacting regulatory path which are the gene signature for a particular disease, between dominating genes are inferred from the probability distribution table and Bayesian Graph. Such reconstructed regulatory network is attractive for their ability to describe complex stochastic processes like gene transcription, classification of biological sequencing and intuitive model of causal influence successfully. This may serve as a signature pattern of the disease Adenocarcinoma, which has been extracted from huge microarray dataset. Extraction of this signature pattern is very useful for diagnosis of this disease.

Keywords: Cancer diagnosis; Micro array data; Reduct; Rough set; Rule reduction; Gene regulatory network; Bayesian network

Introduction

Genes act as blue print of every living object's activity. Genes produces proteins, this protein or a set of proteins produced by other genes switches on or off or regulates the protein formation activity of other genes. Thus genes form a gene regulatory network, study of which appears to be very important to find the cause of a disease and the solution thereof i.e., 'Drug design'. This is a sort of reverse engineering activity where end result is given i.e. dataset of a diseased person is given. From there we need to go back to normal person's gene network configuration by rectifying the erring network pathway.

Gene regulation is a general name for a number of sequential processes, the most well known and understood being transcription and translation, which control the level of a gene's expression, and ultimately result with specific quantity of a target protein. A gene regulation system consists of genes, cis-elements, and regulators. The regulators are most often proteins, called transcription factors, but small molecules, like RNAs and metabolites, sometimes also participate in the overall regulation. A GRN is a collection of DNA segments of chromosome in a cell which interact with each other indirectly (through their protein products) and with other substances in the cell, thereby governing the expression levels of mRNA and proteins.

In computational point of view, a gene regulatory network is represented by a model or graph which represents regulations or interactions amongst genes using a directed graph. In gene networks, nodes represent genes and edges represent relations or interaction amongst genes (e.g., activation or suppression or regulation). DNA Microarray is an experimental procedure which indicates whether a gene is active or not, and if active, how much are their activation profile [1]. They are represented as a dataset in public domain websites. Microarrays are used in the medical domain to represent genetic profile of diseased and normal tissues of patients. Such profiles are useful as an aid in more accurate diagnosis, prognosis, treatment planning, as well as drug discovery for a particular disease.

Main drawback for the gene network construction from microarray data is that it is often constrained by limited number of samples (patients), and the data contains a large number of genes. This imparts inaccuracy in the network design. Also the information contained in gene expression data, is limited by their quality, the experimental design, noise, and measurement errors. Therefore, estimated gene networks may represent some incorrect gene regulations, which may infer wrong conclusion on biological viewpoint. As the microarray data contain thousands of genes, the construction of GRN by using such a large number of genes will result into huge network, which is almost impossible to analyze and computationally excessive time-consuming.

***Corresponding author:** Sudip Mandal, Head of the Department, ECE Department, Global Institute of Management and Technology, Krishna Nagar, India, E-mail: sudip.mandal007@gmail.com

Received August 22, 2013; **Accepted** September 24, 2013; **Published** September 30, 2013

Citation: Mandal S, Saha G, Pal RK (2013) Reconstruction of Dominant Gene Regulatory Network from Microarray Data Using Rough Set and Bayesian Approach. J Comput Sci Syst Biol 6: 262-270. doi:10.4172/jcsb.1000121

Therefore, efforts have been exerted to find smaller number of relevant genes from the vast Microarray data set, keeping the biological relevance intact to its maximum level.

Next, efforts are made to try to find out the hidden dependency, i.e., regulatory interaction amongst the responsible smaller number of genes as derived. The direct or indirect interactions amongst the genes can be derived by constructing the GRN from this reduced information table. This network will act as a signature pattern of a particular disease extracted from the huge microarray dataset, which can be used for disease diagnosis purpose.

The most common intelligent techniques used for the above mentioned analysis are based on soft computing tools like Data Mining, Neural Network, Genetic Algorithms, Decision Trees and fuzzy theory etc [2-9]. Several methods have been proposed for estimating gene networks from microarray data using mathematical models such as Boolean networks differential equations and Bayesian networks [10-23]. Although these methods succeed in constructing networks where genes known to be biologically related come close together, it is difficult to determine the correct direction of the edges, as well as whether or not the relation of genes is direct or indirect.

In this investigation, Rough Set theory has been proposed to find dependence relationship among Microarray data, where genes have been considered as attributes and patients as objects [24-29]. Since RST can work in an environment where some of the data may be inexact or superfluous, it better suits for the present analysis where probability of having superfluous data in Microarray is there. Using RST, rule sets are extracted and from there important attributes or genes are identified. Then, a pattern based on the interrelation amongst the related genes is to be assumed as signature pattern for a probable type of disease. These decision-making rules eliminate all redundant objects i.e. patients and attributes i.e. genes, and thereby resulting into finding minimum subset of attributes to be used for attaining a satisfactory classification rules [30,31]. Moreover, the rough set reduction algorithms help to approximate the decision classes using possibly large and simplified patterns.

On the other hand, Bayesian Networks represent the dependence structure between multiple interacting quantities (e.g., expression levels of different genes). The present approach, which is probabilistic in nature, is capable of handling noise and estimating the confidence in the different features of the network. Therefore, it is possible to focus on interactions whose signal in the data is strong. Here, using this approach a network has been derived, which is referred as 'signature pattern' of the particular disease.

Lung Adenocarcinoma often begins in the outer parts of the lungs and shows well-known symptoms of Lung Cancer such as a chronic cough and coughing up blood may be less common until later stages in the disease. Early symptoms of Adenocarcinoma which include fatigue, mild shortness of breath, or pain in our back, shoulder, or chest are generally overlooked, which could have been curable if detected at early stage.

In this investigation, a novel methodology to find out the most dominant regulatory network for the automated diagnosis of Lung Adenocarcinoma using the microarray dataset of Adenocarcinoma on the basis of probability dependency among dominant genes has been proposed. This proposed dominant network model has been derived using RST and Bayesian network analysis methodology.

The RST has been applied on microarray dataset of Adenocarcinoma involving around 22,000-30,000 human genes and 115 patients. This reduction technique is applied to find all possible reducts from the microarray dataset which contains the minimal subset of attributes those are associated with a class label for classification. Total numbers of rule sets generated are 2187, which is further reduced to 15 rules without sacrificing accuracy. These 15 genes are stronger contender for the cause of Adenocarcinoma. Using the gene expressions corresponding to 15 numbers of dominant genes, a Bayesian Network has been constructed based on conditional independencies, which in turn depicts the dominant molecular regulatory network from gene expressions profiles. This network is very useful to find out the direct and indirect causal influence between different responsible genes. This network pattern can act as 'signature pattern' for the disease 'Adenocarcinoma'. This network profile will be helpful for diagnosis of the disease & may be helpful information for drug design of the disease.

The rest of the paper is organized as follows. Section II depicts the theoretical aspect of Rough set theory and the Bayesian Network which are the basic tools for this work. The evolutionary process of rule reduction and to find out the most dominant genes responsible for Adenocarcinoma is described in Section III. In Section IV, properties, learning structure algorithm and search algorithm of Bayesian network have been discussed. In next section, the inferred result corresponding to dominant Molecular Regulatory Network is discussed. The discussion and reference for this study is given in Section VI and VII respectively.

Preliminaries

In this section, we briefly discuss the basic concepts of Rough Set theory and Bayesian Network expression. A.

Rough set theory

Rough sets constitute a major mathematical tool for managing uncertainty that arises from granularity in the domain of discourse due to incomplete information about the objects of the domain. The granularity is represented formally in terms of an indiscernibility relation that partitions the domain. If there is a given set of attributes ascribed to the objects of the domain, objects having the same attribute values would be indiscernible and would belong to the same block of the partition. The intention is to approximate a rough (imprecise) concept in the domain by a pair of exact concepts. These exact concepts are called the lower and upper approximations and are determined by the indiscernibility relation. The lower approximation is a set of objects definitely belonging to the rough concept, whereas the upper approximation is a set of objects possibly belonging. The formal definitions of the aforementioned notions and others required for the present work are given as follows.

Definition 1: An information system $A = (U,A)$ consists of a nonempty, finite set U of objects (cases, observations, etc.) and a non-empty, finite set A of attributes a (features, variables), such that $a : U \rightarrow Va$, where Va is a value set. We shall deal with information systems called decision tables, in which the attribute set has two parts ($A = C \cup D$) consisting of the condition and decision attributes (in the subsets C, D of A, respectively). In particular, the decision tables we take will have a single decision attribute d and will be consistent, i.e., whenever objects x, y are such that for each condition attribute a, $a(x) = a(y)$, then $d(x) = d(y)$.

Definition 2: Let $B \subset A$. Then a B-indiscernibility relation IND (B) is defined as

$$\text{IND}(B) = \{(x, y) \in U: a(x) = a(y), \forall a \in B\}. \tag{1}$$

It is clear that IND (B) partitions the universe U into equivalence classes

$$[xi]B = \{xj \in U: (xi, xj) \in \text{IND}(B)\},\ xi \in U. \tag{2}$$

The equivalence class of IND (B) is called the elementary set in B because it represents the smallest discernible objects.

Definition 3: The *B*-lower and *B*-upper approximations of a given set $X(\subseteq U)$ are defined, respectively, as follows:

$$\overline{BX} = \{x \in U : [x]B \subseteq X\} \tag{3}$$

$$BNB(X) = BX \,/\, \underline{BX} \tag{4}$$

The *B*-boundary region is given by

$$BNB(X) = \overline{BX} \,/\, \underline{BX} \tag{5}$$

Assuming B and C are equivalence relation in U, the important concept of positive region

$$POS_B(C) = \cup_{[X \in C]} \underline{BX} \tag{6}$$

Definition 4: In an information system there often exist some condition attributes that do not provide any additional information about the objects in *U*. So, we should remove those attributes since the complexity and cost of decision process can be reduced if those condition attributes are eliminated. Given a classification task mapping a set of variables *C* to a set of labeling *D*, a reduct is defined as any $R \subseteq C$, such that $\gamma(C,D) = \gamma(R,D)$ and a reduct set is defined with respect to the power set $\boldsymbol{P}(C)$ as the set $R \subseteq \boldsymbol{P}(C)$ such that

$$R = \{A \in P(C) : \gamma(A,D) = \gamma(C,D)\} \tag{7}$$

That is, the reduct set is the set of all possible reducts of the equivalence relation denoted by C and D. a minimal reduct is defined as any reduct *R* such that $|R| \le |A|, \forall A \in R$. That is, the minimal reduct is the reduct of least cardinality for the equivalence relation denoted by *C* and *D*.

Definition 5: The set of attributes which are common to all reduct is called core. The core is the set of attributes which is possessed by every legitimate reduct, and therefore consists of attributes which cannot be removed from the information system without causing collapse of the equivalence-class structure. It is possible for the core to be empty, which means that there is no indispensable attribute.

In Rough Set Theory, the datasets are represented with the help of decision tables. The decision table contains attributes i.e. condition and objects for different cases of samples. The decision table describes the decision in terms of conditions that must be satisfied in order to obtain the decisions specified in the decision table. In this paper, different genes are considered as attributes and gene‘s expression value are considered as object of a decision table. Based on different condition or different value of attribute the decision of the sample may be either normal or cancerous. This decision table is used as a training dataset which is used to know hidden dependency between different genes which are responsible for Adenocarcinoma.

In this section, we describe how decision rules are generated based on the reduct system obtained from previous section. If we distinguish in information system two disjoint classes of attributes, called condition and decision attributes respectively, then the system will be called a decision table and will be denoted by $S = (U,C,D)$, where *C* and *D* are disjoint sets of condition and decision attributes, respectively.

Let $S = (U, C, D)$ be a reduced decision table where C denotes the reduced no. of attributes i.e. reduct. Every $x \in U$ determines a sequence $c_1(x) \ldots c_n(x)$; $d_1(x) \ldots d_m(x)$, where $\{c_1, \ldots, c_n\} = C$ and $\{d_1, \ldots, d_m\} = D$. The sequence will be called a decision rule induced by x (in *S*) and will be denoted by $c_1(x) \ldots c_n(x) \to d_1(x) \ldots d_m(x)$, or in short $C \to_x D$.

The number $\text{supp}_x(C,D)$ will be called a support of the decision rule $C \to_x D$ and the number is given by

$$\text{supp}_x(C,D) = |A(x)| = |C(x) \cap D(x)| \tag{8}$$

The strength of the decision rule $C \to_x D$ can be written as the following equation where denotes the cardinality of *X*.

$$\sigma_x(C,D) = \frac{\sup p_x(C,D)}{|U|} \tag{9}$$

With every decision rule $C \to_x D$ we associate the coverage factor of the decision rule, denoted $cov_x(C,D)$ and defined as follows:

$$\text{co}\nu_x(C,D) = \frac{|C(x) \cap D(x)|}{|D(x)|} \tag{10}$$

Bayesian network

In this paper, Bayesian approach has been introduced to extract the inter dependencies of the relevant genes by studying their expression patterns that uncovers their transcriptional patterns by examining statistical properties of dependence and conditional independence in the given data. Bayesian networks are particularly useful for describing processes composed of locally interacting components; that is, the value of each component directly depends on the values of a relatively small number of components. Secondly, the statistical foundations for learning Bayesian networks from observations, and computational algorithms are well understood and have been used successfully in many applications. Finally, Bayesian networks provide models of causal influence: although Bayesian networks are mathematically defined strictly in terms of probabilities and conditional independence statements, a connection can be made between this characterization and the notion of direct causal influence.

This stage involves the construction of a Genetic Network from the set of dominant gene expressions which are obtained from the preliminary stage of rules reduction process by using Rough set theory. The Bayesian belief network is a kind of probabilistic model for the construction of genetic network. It uses Direct Acyclic Graph which consists of nodes representing attributes and directed acyclic edges between them according to their independency and each node is attached with a joint probability distribution table to represent dependency relationships between variables. Since every independent statement in belief networks satisfies a group of axioms, we can construct belief networks from data by analyzing conditional independence relationships. The Conditional Independence (CI) test based method is used by all the algorithms of the second category which analyze relations of different quantities based on their dependency relationships.

Let‘s review the concept of d-separation or independency between nodes. For any three disjoint node sets *X, Y,* and *Z* in a belief network, *X* is said to be d-separated from *Y by Z* if there is no active undirected path between *X* and *Y*. A path between *X* and *Y* is active if:

i) Every node in the path having head-to-head arrows is in *Z* or has a descendant in *Z*;

ii) Every other node in the path is outside *Z*.

The amount of information flow between two nodes can be measured by using mutual information, when no nodes are instantiated, or conditional mutual information, when some other nodes are instantiated. In information theory, the mutual information of two nodes, is defined as

$$I(X_i, X_j) = \sum_{X_i, X_j} P(X_i, X_j) \log \frac{P(X_i, X_j)}{P(X_i)P(X_j)} \quad (11)$$

and conditional mutual information is defined as

$$I(X_i, X_j / C) = \sum_{X_i, X_j, C} P(X_i, X_j, C) \log \frac{P(X_i, X_j / C)}{P(X_i / C)P(X_j / C)} \quad (12)$$

where, X_i, X_j are two nodes and C is a set of nodes. Conditional mutual information is used as CI tests to measure the average information between two nodes when the status of some valves is changed by the condition-set C. When $I(X_i, X_j/C)$ is smaller than a certain threshold value, we say that X_i, X_j are d-separated by the condition-set C, and they are conditionally independent.

A gene network, or a gene regulatory network, is a graphical model that represents the regulatory relationships between genes. In a gene network, if there is an edge from gene *a* to gene *b*, then the edge represents that gene *a* regulates gene *b*, or the expression of gene *b* depends on the expression of gene *a*. We model a gene network *G* as a Bayesian network, where genes are represented by random variables and the structure is described as a directed graph with the random variables as its nodes. Let $X = \{X_1, X_2, \ldots, X_p\}$ be a set of random variables (genes) in the network *G*, where *p* is the number of nodes. In the context of Bayesian networks, the joint probability of *X* conditional on *G* can be decomposed as a product of conditional probabilities,

$$P(X|G) = \prod_{j=1}^{p} P(X_j | Pa(X_j)) \quad (13)$$

where $Pa(X_j)$ is a set of parent variables of X_j in *G*. Suppose that $\boldsymbol{X}$ is a gene expression data matrix whose element x_{ij} corresponds to the expression value of the *j*th gene in the *i*th array, where $i = 1, \ldots, n$, $j = 1, \ldots, p$. Here, *n* and *p* represent the number of microarrays and genes, respectively. Since microarray data take continuous variables, the probabilistic measures in Eq. (3) are replaced by densities and the likelihood of $\boldsymbol{X}$ conditional on *G* is given by

$$P(X|G) = \prod_{i=1}^{n} \prod_{j=1}^{p} P(X_{ij} | Pa(X_{ij})) \quad (14)$$

where $Pa(X_{ij})$is a set of expression values of the parent genes of *j*-th gene at *i*-th experiment. The joint probability distribution table for each node is known as inference.

Identification of Responsible Genes by Rule Reduction Process

Here, the genes are considered as attributes and about 82 patient's data as object. These are used to generate the required decision table. Based on different conditions or different values of attributes the decision of the sample may be either normal or cancerous. This decision table is used as a training dataset which is used to calculate hidden dependency amongst different genes which are responsible for Adenocarcinoma. The microarray data for Adenocarcinoma and normal human being (Series Geo_accession No.: GSE10072) has been collected from the NCBI website [http://www.ncbi.nlm.nih.gov] for the present context. Here, number of genes / attributes in the dataset are 22284 and the no of patients / objects are 82. The values of the different attributes are real but for RST analysis it is considered as integers. The following table is the resultant decision table corresponding to the Adenocarcinoma Microarray data where the no of column is 22285 and no of rows are 82.

The training data set consist of 42 cases of tumor and 40 cases of normal data set. Another 15 different random data set of different cases has not been used for generating rule sets. They are treated as test data set to test the validity of generated rules. As shown in the Table 1, 1007_s_at, 1053_at… AFFX-Trpnx-M _at are found to be the genes, depending upon which, decision is taken.

Reduct calculation is a crucial task in RST system. Here, the minimal numbers of reducts have been calculated from decision table using the software package "Rough Set Exploration Systems (RSES 2.2.2)"{[http://logic.mimuw.edu.pl/rses/]. Calculation of all reducts is very exhaustive and complex in nature. Therefore, for the calculation of all minimal reducts, Genetic Algorithm with full indiscernibility and modulo decision technique has been used. From the huge database or decision table, 37 no of reducts of various size are generated, each of which have the positive region 1 and Stability Coefficient equal to 1.

It is interesting to find that there is no core in the reduct set which means that there is no indispensable attribute and there are huge dependencies between different attributes of minimal reduct sets. In other words, there is a huge inhomogeneity among the attributes and there are many possibilities of substitution. These hidden dependencies, responsible for Adenocarcinoma needs to be calculated out. Thus only 195 numbers of different attributes has been figured out of 22284 numbers of total attributes in the Microarray dataset. Thus the dependency and complexity of the decision has been reduced by a factor 195/22284 after the calculation of reduct sets. Using these generated reducts (Table 2) and decision table, the decision rules have been generated with the help of RSES2.2 package. Due to the huge dependency with each other in the attributes of reduct set, 2187 no of rules have been generated. Using these rules, it is possible to classify an unknown Microarray data set of any human either into Adenocarcinoma affected or normal lungs.

Considering the first reduct i.e., {201591_s_at, 202295_s_at, 209613_s_at}, which consist of three attributes, it is found that, for the first reduct the number rules generated are 35 which are quite large in number and can be used for classification.

Table 3 shows a portion of the generated rules (first 35 rules only for the first reduct set) where each generated rules are given first and at rightmost column 'Match/support' denotes the number of case in the decision table which are supported by this particular rule.

Due to large number of rules, it is almost impossible to understand the hidden data dependency on each other manually. That's why the

CASES				ATTRIBUTES OR GENE EXPRESSION		
	1007_s_at	1053_at	..	AFFX- pnX-5_at	AFFX-TrpnX-M_at	Decision
X1	10	7	..	4	4	TUMOR
X2	10	6	..	4	4	NORMAL
X3	10	6	..	4	4	TUMOR

Table 1: Sample Decision table (truncated).

Reduct Set
{ 201591_s_at, 202295_s_at, 209613_s_at }
{ 201456_s_at, 203065_s_at, 203217_s_at, 215972_at }
{ 201413_at, 201969_at, 204987_at,205022_s_at,206550_s_at,208429_x_at}

Table 2: Reduct Set for the data set of Adenocarcinoma after using RSES (partially shown).

generated rule needs to be reduced in number without effecting overall accuracy and coverage of the rules. In this rules reduction approach, basically there are following 3 steps.

Step 1-Shortening: Shortening is the first step where large rules are simplified. The process by which the maximum numbers of condition attributes are removed without losing essential information is called value reduction and the resulting rule is called maximally general or minimal length. Computing maximally general rules is of particular importance in knowledge discovery since they represent general patterns existing in the data.

The simplification rule algorithm initialize general rules GRULE to empty set and copies one rule $r_1 \in$ RULE to rule *r*. A condition is dropped from rule *r*, and then rule *r* is checked for decision consistency with every rule $r_j \in$ RULE. If rule *r* is inconsistent, then the dropped condition is restored. This step is repeated until every condition of the rule has been dropped once. The resulting rule is the simplified rule.

Consider the 35 rules which are generated for the first reduct set. Using above algorithm, it is found that that the condition which are imposed by the attribute 201591_s_at and 202295_s_at can be removed without affecting the accuracy of the rules generated by the first reduct set as 209613_s_at is the most dominant attribute in the first reduct set. Using RSES2.2, shortening is applied to the 2187 no of rule with shortening ratio 0.9. The user provides a coefficient between 0 and 1, which determines how 'aggressive' the shortening procedure should be. The coefficient is equal to 1.0 means that no shortening occurs. If shortening ratio is near zero, the algorithm attempts to maximally shorten reducts. This shortening ratio is in fact a threshold imposed on the relative size of positive region after shortening Applying the above mentioned algorithm the number of rules for the first reduct set is reduced to 7 (5 for the tumor, 2 for the normal) with only a single attribute 209163_s_at.The following shortened 7 rules are the rules only for the first reduct set.

(209613_s_at=8)=>(CLASS=TUMOR[7])

(209613_s_at=4)=>(CLASS=TUMOR[11])

(209613_s_at=5)=>(CLASS=TUMOR[4])

(209613_s_at=7)=>(CLASS=TUMOR[11])

(209613_s_at=6)=>(CLASS=TUMOR[7])

(209613_s_at=11)=>(CLASS=NORMAL[7])

(209613_s_at=10)=>(CLASS=NORMAL[28])

After shortening overall number rules is reduced to only 770 rules. So reduction factor after shortening is=(2187/770) =2.84.

Step 2-Generalization: Though it is found that numbers of rules are reduced after shortening but due to different value there exist different rules for the same decision. Using shortening we minimize the dependent attribute in a rule whereas Generalization is the process by which the values of the reduced attribute are reduced into conjunctive form. For the first reduct set, the number of shortened rule is only 7 but there are the different values of a same attribute 209613_s_at for which the decision is either tumor or normal. As an example, if the value is '4 or 5 or 6 or 7 or 8' then decision will be tumor and if the value is either 10 or 11 then samples will be considered as the normal. Therefore after generalization of the shortened rules for the first reduct set can be written

(209613_s_at=8|4|7|6|5)=> (CLASS=TUMOR [40])

(209613_s_at=11|10)=> (CLASS=NORMAL [35])

Interestingly it is found that the Match/support of this generalized rule is increased as the generalized & shortened rule support all individual shortened rules for normal and tumor. So, total 'support' for a particular decision is the sum of the individual rule for that decision.

Obviously generalized rules signify stronger rule than shortened rules. The 'strength' of the rule is defined as the ratio of the number supported by a rule for a particular condition to the total no of cases in the universe of decision table. So the strength of the above described generalized rule for tumor is = 40/82=0.488 which is quite good. Overall after generalization of shortened rule in this case, number of rules is reduced to 707.

Step 3-Filter: In previous section it is found that after generalization the strength of the rule increases as the support is increased. Moreover stronger rules are better for classification of unknown dataset. Therefore we filter the shortened & generalized rule with the condition of removing the rules up to 35 supports which is chosen randomly without sacrificing the accuracy of classification.

After this filter, we get only 15 numbers of strongest rules (Table 5) from 707 no of generalized rules which can be used for diagnosis of a human being from the microarray dataset which are given below. The no. of rules must be reduced in such a way that it should not affect the accuracy of classification.

Table 4 shows how the number & length of the rule are decreased but strength of the rule is increased with the different stage of rule reduction method which guarantee the accuracy of the rule will also increased in spite of the reduction of rules.

The attributes or the genes that consist of different 15 genes are stronger contender for the cause of Adenocarcinoma because they remain in reduced rule set. Table 6 depicts the most responsible genes for Adenocarcinoma.

(201591_s_at=10)&(202295_s_at=11)&(209613_s_at=8)=>(CLASS=TUMOR[2])	(201591_s_at=10)&(202295_s_at=11)&(209613_s_at=11)=>(CLASS=NORMAL[3])
(201591_s_at=10)&(202295_s_at=10)&(209613_s_at=4)=>(CLASS=TUMOR[4])	(201591_s_at=10)&(202295_s_at=12)&(209613_s_at=11)=>(CLASS=NORMAL[4])

Table 3: Generated rules set from reduct sets (partially shown).

Different stage of rules reduction	No of rules	Support of the rules		Length of the rules premises		Strength of the rules	
		Minimum	Maximum	Minimum	Maximum	Minimum	Maximum
Rules form reduct set	2187	1	18	3	11	0.0120	0.219
After Shortening	770	1	39	1	8	0.0121	0.476
After Generalization	707	1	41	1	8	0.0121	0.500
After filter	15	37	41	1	3	0.4512	0.500

Table 4: Comparative study of different stages of rule reduction.

(209613_s_at=8|4|7|6|5)=>(CLASS=TUMOR[40]) 40
(203065_s_at=10|9|8|7)=>(CLASS=TUMOR[41]) 41
(206068_s_at=6|5)&(206757_at=4)=>(CLASS=TUMOR[41]) 41
(206068_s_at=6|5)&(212760_at=8|9)=>(CLASS=TUMOR[41]) 41
(201772_at=9|8|10)&(206068_s_at=6|5)=>(CLASS=TUMOR[41]) 41
(49452_at=6)=>(CLASS=TUMOR[39])39 (208056_s_at=7|6|5)&(218918_at=9|7|8|6)&(49452_at=6)=>(CLASS=TUMOR[39])39
(209072_at=7|6)&(218982_s_at=8|9|10|11)&(49452_at=6)=>(CLASS=TUMOR[38])38
(201591_s_at=10|11)&(202295_s_at=11|12|10)&(209613_s_at=9|10|11)=>(CLASS=NORMAL[38])38 (203065_s_at=12|11)=>(CLASS=NORMAL[39]) 39
(203091_at=8|7)&(203249_at=9)=>(CLASS=NORMAL[37]) 37
(206068_s_at=7|8)=>(CLASS=NORMAL[38])38 (209613_s_at=9|10|11)&(222313_at=7|6|5)=>(CLASS=NORMAL[39]) 39
(49452_at=8|7)=>(CLASS=NORMAL[37]) 37
(201938_at=10|9)&(205261_at=10|11|12)=>(CLASS=NORMAL[36]) 36

Table 5: Most dominant 15 number of rules for classification of Adenocarcinoma.

201591_s_at	202295_s_at	203249_at	208056_s_at	218918_at
201772_at	203065_s_at	205261_at	209072_at	222313_at
201938_at	203091_at	206068_s_at	209613_s_at	49452_at

Table 6: Most dominant genes responsible for Adenocarcinoma.

Bayesian Network Model

This stage involves the construction of a Genetic Network from the set of dominant gene expressions which are obtained from the preliminary stage of rules reduction process by using Rough set theory. The dual nature of a Bayesian network makes learning a Bayesian network from an unknown dataset as a two stage process a natural division. First learn a network structure, and then learn the probability tables. There are various approaches to structure learning one of which is conditional independence test: These methods mainly stem from the goal of uncovering causal structure. The assumption is that there is a network structure that exactly represents the independencies in the distribution that generated the data. Then it follows that if a (conditional) independency can be identified in the data between two variables, there is no arrow between those two variables. Once locations of edges are identified, the direction of the edges is assigned such that conditional independencies in the data are properly represented.

At the moment, only the ICS algorithm is implemented. The algorithm makes two steps, first, find a skeleton (the undirected graph with edges if there is an arrow in network structure) and second, direct all the edges in the skeleton to get a DAG. Starting with a complete undirected graph, we try to find conditional independencies *P[x, y|Z]* in the data. For each pair of nodes x, y, we consider sets Z starting with cardinality 0, then 1 up to a user defined maximum. Furthermore, the set Z is a subset of nodes that are neighbors of both x and y. If an independency is identified, the edge between x and y is removed from the skeleton. A test is performed by using any of the score metrics to test whether variables x and y are conditionally independent given a set of variables Z.

The first step in directing arrows is to check for every configuration x—*y*—*z* where x and y not connected in the skeleton whether z is in the set Z of variables that justified removing the link between x and y (cached in the first step). If z is not in Z, we can assign direction x→ z ←y.

Finally, a set of graphical rules is applied to direct the remaining arrows.

Rule 1: i→j—k & i-/-k => j→k

Rule 2: i→j→k & i—k => i→k

Rule 3 : i→j←k & m—i—j & m—k—j & m—j => m→j

Rule 4 : i→j & m—i—j & m—k—j & i—k => k→m & i→m

Rule 5: if no edges are directed then take a random one (first we can find).

The problem of learning a Bayesian network can be stated as follows. Given a training set $D=\{X^1, X^2 \ldots X^N\}$ of independent instances of *X*, find a network B= *[G, C]* that best matches *D*. The common approach to this problem is to introduce a statistically motivated scoring function that evaluates each network with respect to the training data, and to search for the optimal network according to this score.

A commonly used scoring function is the Bayesian scoring metric (Cooper & Herskovits 1992, Heckerman et al. 1995) $Score(G:D) = logP(G|D)=logP(D|G)+logP(G)+C_1$ where C_1 is a constant independent of *G* and *P(D|G)* the *marginal likelihood* which averages the probability of the data over all possible parameter assignments to *G* . The particular choice of priors *P (G)* and $P(C_1|G)$ for each *G* determines the exact Bayesian score. Under mild assumptions on the prior probabilities, this scoring metric is asymptotically consistent. Given a sufficiently large number of samples, graph structures that exactly capture all dependencies in the distribution, will receive, with high probability, a higher score than all other graphs. This means, that given a sufficiently large number of instances in large data sets, learning procedures can pinpoint the exact network structure up to the correct equivalence class.

This algorithm also makes the following assumptions:

1. The database attributes have discrete values and there are no missing values in all the records.
2. The volume of data is large enough for reliable CI tests.

A truncated decision table, containing 82 different cases of gene expressions corresponding to 15 most dominant responsible genes which we already collected by rule reduction method using Rough Set theory from the same microarray dataset (GSE10072), is used to reconstruct the Bayesian Network. To build the regulatory path between the responsible genes i.e. Bayesian Network, "WEKA -3.6" software tool [http://www.cs.waikato.ac.nz/ml/weka/] with Bayes Net classifier is used. We have consider state of the lungs either normal or cancer (named as class) as a virtual gene which is also included in the table to find out the regulatory path between other genes and the state. ICS algorithm is used as the learning structure search algorithm and simple estimator option can be used to select the method for estimating the conditional probability distributions of different genes or attributes. The maxCardinality option determines the largest subset of Z to be

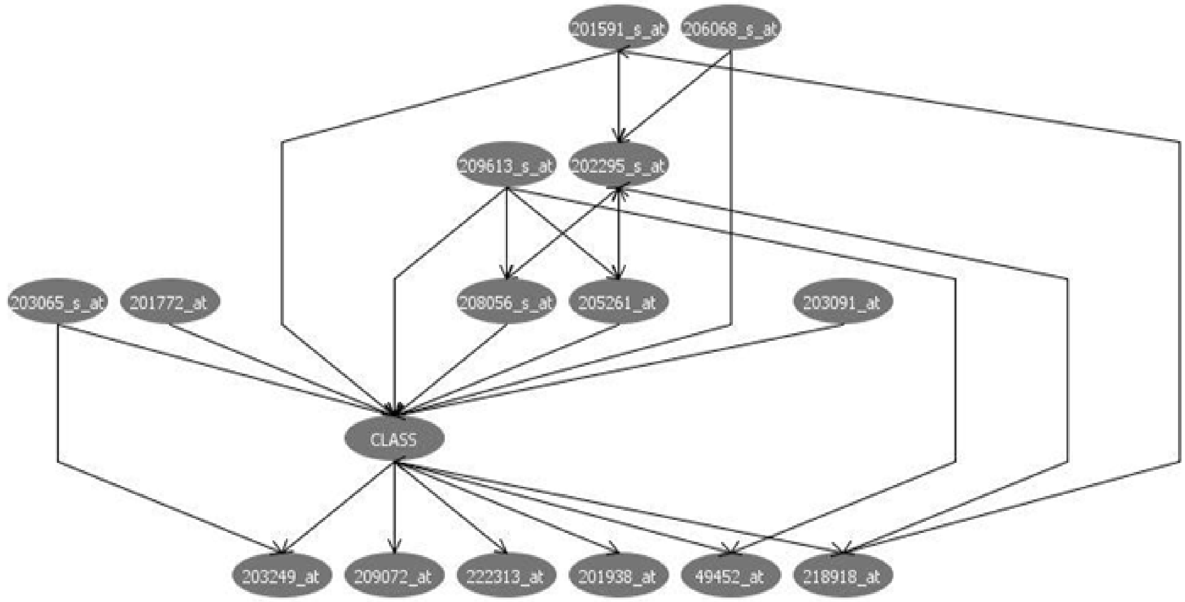

Figure 1: Dominant Gene Regulatory Network for the disease Adenocarcinoma using Bayesian Model.

considered in conditional independence tests *P[x, y|Z]* and Bayes type score is used for this purpose. The Bayesian Network & probability distribution table corresponding to the 15 dominant genes is shown in Figure 1. This acts as 'signature' of the disease 'Adenocarcinoma' extracted from its Microarray dataset.

Result Analysis

Here, this study has been carried out for automated human disease diagnosis. The input to the process is genetic information of the person under investigation in the form of Microarray dataset. The extracted rules are cross validated against the training data set and also unknown persons Microarray dataset. The results show that these rules can predict the 42 tumor & 40 normal which is 100% accurate. Then again these reduced rules are verified against the 15 number of unknown test data sets which consist of 8 tumor & 7 normal cases. After classification using the rules, it can be seen that the rules can classify the new data set with same accuracy and coverage factor 1. The genes that remain in different 15 rules are stronger contender for the cause of Adenocarcinoma. By analyzing the attribute values of responsible genes we can easily identify whether a person has normal or cancerous lung, from his microarray data. The above results are verified using publicly available website Gene-ontology, named DAVID [http://david.abcc.ncifcrf.gov/] which provides a comprehensive set of functional annotation tools for investigators to understand biological meaning behind large list of genes. Among the 15 no of genes, all genes directly or indirectly related to Lung Adenocarcinoma, proves the fact that RST can extract biological relevant information also. Thus it can be concluded that using this technique, accurate diagnosis of Adenocarcinoma can be done, if the microarray data of the concern patient is available.

From the above Bayesian network and inference table, is following observation can be made.

1. *201591_s_at, 206068_s_at, 203065_s_at, 20172_at, 209613_s_at & 203091_at* these genes are *parent genes* which have *direct causal influence* on Adenocarcinoma (*class*). Therefore depending upon expression value of these genes, the state of the lungs of human is directly influenced. So if we want to design a drug for Adenocarcinoma, these parent genes must be druggable as they are directly responsible for changing the state from normal to cancerous lungs.
2. On the other hand, *202295_s_at, 208065_s_at, 205261_at* these genes are *intermediate gene* which are indirectly regulate the state of the lungs.
3. The intermediate genes are particularly depending on the *parent genes*. As an example *209613_s_at* gene can directly influenced the state. But *20805_s_at* is also triggered by *209613_s_at* which is parent node of the 20805_s_at. In other words, the effect of gene *209613_s_at* on state of the lungs _ is *mediated* through gene *20805_s_at* _. Once we know the expression level of *209613_s_at* gene _, the expression of gene _ *20805_s_at* does not give new information about the state. So *209613_s_at* can directly or indirectly infer the class or state. Same statement can be concluding for other intermediate genes.
4. There is a different regulatory path existing in the gene network based on dependency of different gene expression value. These dominant regulatory path are described below:

- 201591_s_at→*202295_s_at→205261_at→class*
- 201591_s_at→*202295_s_at→208065_s_at →class*
- *206068_s_at→202295_s_at→205261_at→class*
- *206068_s_at→202295_s_at→208065_s_at →class*
- *209613_s_at→208065_s_at →class*
- *209613_s_at→205261_at →class*
- *201591_s_at→class*
- *206068_s_at→class*

- *203065_s_at→class*
- *20172_at→class*
- *209613_s_at→class*
- *203091_at→class*

5. On the other hand *203249_at, 49452_at & 21898_at* these three genes depend on the class or state of the lung as well as on other parent genes but they don't infer the state. So these genes are less of concern during drug design as they are the *children node* of the *class*.

6. *209072_at, 2213_at and 201938_at* these genes neither directly nor indirectly cause cancer. Moreover these genes are not affected according the current state of the lungs which can be concluded by observing probability tables of therm. No dependency is found either with other genes or the states. So these genes which are at the low layer than the Class of the Bayesian Network have also no practical significance during drug design.

So, using Bayesian Network, a GRN for Adenocarcinoma is constructed using these 15 genes which can be treated as 'signature patter' for a particular disease from microarray data which can be used for classification as well as to find out regulatory path which is helpful for drug design in future.

Discussion

In this paper, a new and novel approach for reconstruction of dominant molecular regulatory network from gene expressions profiles using Rough set theory & Bayesian network analysis method has been proposed. Rough set theory has been successfully implemented to find out the hidden dependency between huge imperfect dataset by calculating reduct and decision rules from it. Using shortening, generalization & filter process we can easily get a simple & few numbers of rules, by which we can predict the status of a human with 100% accuracy and moreover find the most dominant genes which are responsible for Adenocarcinoma. The result also predicted 15 responsible / affected genes for causing the disease Adenocarcinoma. The validation for this can be carried out in the Gene Ontology website-David. It has been assumed that the set of genes will more or less act in the same general way in a particular species e.g. human beings in the present investigation scenario.

We also presented a new approach for finding out the gene network by analyzing gene expression data of dominant gene that builds on theory and algorithms for learning Bayesian networks. The approach includes two techniques that were motivated by the challenges posed by this domain: a novel search algorithm (CI Test-ICS algorithm) and an approach for estimating statistical Bayes score. We applied our methods to the real expression data of Adenocarcinoma. Although we did not use any prior knowledge, we managed to extract many biologically plausible conclusions & regulatory path for drug design in future. The dominant Gene Regulatory Network can be considered as 'Genetic Signature' of the disease. The results needed to be verified in wet-lab but we hope this process have future potential in medicine.

References

1. Masys DR (2001) Linking microarray data to the literature. Nat Genet 28: 9-10.
2. Mohammadi A, Saraee MH, Salehi M (2011) Identification of disease-causing genes using microarray data mining and Gene Ontology. BMC Med Genomics 4: 12.
3. Wang Z, Palade V, Xu Y (2006) Neuro-fuzzy ensemble approach for microarray cancer gene expression data analysis. Evolving Fuzzy Systems, 2006 International Symposium 241-246.
4. Huerta EB, Duval B, Hao J (2006) A hybrid GA/SVM approach for gene selection and classification of microarray data. Appl Evol Comp 3907: 34-44.
5. David JM, Balakrishnan K (2010) Machine learning approach for prediction of learning disabilities in school-age children. Int J Comp Appl 9: 7-12.
6. David JM, Balakrishnan K (2011) Prediction of key symptoms of learning disabilities in school-age children using rough sets. Int J Comp Electrical Eng 3: 163-168.
7. Bezdek JC (1993) Editorial: Fuzzy models-What are they and why? Fuzzy Systems, IEEE Transactions 1: 1-6.
8. Vinterbo SA, Kim EY, Ohno-Machado L (2005) Small, fuzzy and interpretable gene expression based classifiers. Bioinformatics 21: 1964-1970.
9. Ho SY, Hsieh CH, Chen HM, Huang HL (2006) Interpretable gene expression classifier with an accurate and compact fuzzy rule base for microarray data analysis. Biosystems 85: 165-176.
10. Akutsu T, Miyano S, Kuhara S (1999) Identification of genetic networks from a small number of gene expression patterns under the Boolean network model? Pac Sym Biocomput 4: 17-28.
11. Akutsu T, Miyano S, Kuhara S (2000) Inferring qualitative relations in genetic networks and metabolic pathways. Bioinformatics 16: 727-734.
12. Maki Y, Tominaga D, Okamoto M, Watanabe S, Eguchi Y (2001) Development of a system for the inference of large scale genetic networks. Pac Symp Biocomput 2001: 446-458.
13. Shmulevich I, Dougherty ER, Kim S, Zhang W (2002) Probabilistic Boolean Networks: a rule-based uncertainty model for gene regulatory networks. Bioinformatics 18: 261-274.
14. Chen T, He HL, Church GM (1999) Modeling gene expression with differential equations. Pac Symp Biocomput 1999: 29-40.
15. de Hoon M, Imoto S, Miyano S (2002) Inferring gene regulatory networks from time-ordered gene expression data using differential equations. Discovery Science 2534: 267-274.
16. de Hoon MJ, Imoto S, Kobayashi K, Ogasawara N, Miyano S (2003) Inferring gene regulatory networks from time-ordered gene expression data of Bacillus subtilis using differential equations. Pac Symp Biocomput 2003: 17-28.
17. Friedman N, Goldszmidt M (1998) Learning Bayesian networks with local structure. 252-262.
18. Friedman N, Linial M, Nachman I, Pe'er D (2000) Using Bayesian networks to analyze expression data. J Comput Biol 7: 601-620.
19. Hartemink AJ, Gifford DK, Jaakkola TS, Young RA (2001) Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. Pac Symp Biocomput 2001: 422-433.
20. Hartemink AJ, Gifford DK, Jaakkola TS, Young RA (2002) Combining location and expression data for principled discovery of genetic regulatory network models. Pac Symp Biocomput 2002: 437-449.
21. Imoto S, Goto T, Miyano S (2002) Estimation of genetic networks and functional structures between genes by using Bayesian networks and nonparametric regression. Pac Symp Biocomput 2002: 175-186.
22. Imoto S, Sunyong K, Goto T, Aburatani S, Tashiro K, et al. (2002) Bayesian network and nonparametric heteroscedastic regression for nonlinear modeling of genetic network. Proc IEEE Comput Soc Bioinform Conf 1: 219-227.
23. Darwiche A (2010) Bayesian networks. Communication of the ACM 53: 80-90.
24. Walczak B, Massart DL (1999) Rough sets theory. Chemometrics and Intelligent Laboratory Systems 47: 1-16.
25. Tsumoto S (2001) Medical diagnostic rules as upper approximation of rough sets. Fuzzy Systems, 2001. The 10th IEEE International Conference 1551-1554.
26. Pawlak Z (2002) Rough set theory and its applications. Journal of Telecommunications and Information Technology 7-10.
27. Midelfart H, Komorowsk J, Norsett K, Yadetie F, Sandvik AK, et al. (2002) Learning rough set classifiers from gene expression and clinical data. Fundamenta Informaticae 53: 155-183.

28. Banerjee M, Mitra S, Banka H (2007) Evolutionary rough feature selection in gene expression data. Systems, Man, and Cybernetics, Part C: Applications and reviews 37: 622-632.

29. Hassanien AE, Ali JMH (2004) Rough set approach for generation of classification rules of breast cancer data. Informatica 15: 23-38.

30. Du Y, Hu Q, Zhu P, Ma P (2011) Rule learning for classification based on neighborhood covering reduction. Information Sciences 181: 1-12.

31. Midelfart H, Laegreid A, Komorowski J (2001) Classification of gene expression data in ontology. Medical Data Analysis 2199: 186-194.