

# A novel knowledge-discovering approach from massive data

H. Bouhamed<sup>1</sup>, A. Rebai<sup>2</sup>, T. Lecroq<sup>1</sup> and M. Jaoua<sup>3</sup>

<sup>1</sup>Department of Computer Science, University of Rouen, Rouen, France

<sup>2</sup>Department of Bioinformatics, Biotechnologies Centre of Sfax, Sfax, Tunisia

<sup>3</sup>Department of Computer Science, University of Sfax, Sfax, Tunisia

**Abstract** - *The objective of our study lies in developing a new data-reducing approach whose useful application is crucial as a prerequisite to learning Bayesian Networks (BNs) structures. The application of our approach may, in some cases, turn out to be significantly effective in reducing the computational complexity of the BNs structures learning. Firstly, it is essential to define BNs and recall its widely-common relevant problem of learning structure from massive data. Secondly, we suggest a solution for optimizing the computational complexity by means of data organizational and optimization methods. As a matter of fact we have applied our approach to biological facts concerning hereditary complex illness where the literatures in biology identify the responsible variables for those diseases. Finally, we conclude by highlighting the limits arched by this work and proposing suggestions for further research.*

**Keywords:** optimization, automatic knowledge discovery, Bayesian Network (BN), score fusion, selection of clusters.

## 1 Introduction

It is worth noting that the immense amounts of diverse data, made recently available, pertaining to different research fields, has increasingly kindled interest for the training techniques, skills and efficiency to monitor and deal with the sophisticated data interdependences. Owing to their flexibility, comprehensive mathematical formulations, easy handling and manipulation, the BNs are most often regarded as the favourably chosen models to be applied to various fields and a wide array of applications: astronomy areas, web-mining as well as bioinformatics applications. Yet, training and handling the BNs structures within a large number of diverse variables remains a great challenge to retrieve in the contexts of powerful high-speed processing calculations, algorithmic complexities as well as application execution time [1]. In this respect, several algorithms have recently been devised and developed for the sake of applying and monitoring BNs structures from data [20] and [3]. In fact, a wide array of these algorithms rests on metric scoring methods, the widely compared and most frequently applied scoring methods [2] and [26].

Nevertheless these algorithms and scoring methods remain still insufficient and limited in scope as regards those cases in which the number of variables exceed some hundreds of thousands [18]. Moreover, they do not implement upstream processing treatment of cases where variables are either not

sufficiently and entirely implicated or irrelevant and redundant with respect to a certain information system, in such a way as to exclude them from all-considerations in the modeling process. Most frequently, however, the fact of excluding non-implicated or irrelevant variables, or those whose implications, might considerably decrease the algorithmic complexities as well as the execution time. Thus, providing the possibility of extending the variable-modeling capacity during the initial step of the information system processing stage.

Noteworthy, a more developed and highly promoted type of algorithm has often been applied, using the hierarchical class of latent models (HCLM) [1], along with the double layer BN [25]. These types of algorithms are promising in so far as data reduction capacities are concerned. Yet, they turn out to be incapable of processing a quite large number of variables exceeding the range of about one thousand [18]. As for [10], they have set up a special method allowing to process quite a large amount of data (up to 6000 variables). This has been made possible by means of reducing the HCLM research and retrieving space to some possible connecting relationships among brother nodes. Nevertheless, the restriction imposed by this method is likely to deviate the model from actual and realistic facts [18].

As regards our research study, a novel approach has been devised. Designed to achieve a maximum reduction of the variables number prior to a BN structure learning implementation, this proposed model's effective usefulness lies in the fact that during its execution, neither data flow nor information loss could be engendered during its implementation. At this junction, it is worth noting that this new approach is to be tested and estimated on an intricate data base of genetics' variables pertaining to a complex genetic illness.

As a matter of fact, the present work turns out to be crucially important for a number of reasons. First, it helps determine and extract relevant information from a large set of data variables. Second, it enables to exploit the extracted information and reduce the scope of data ahead of BN modelization. Actually, this modelization is targeted to lessen the level and effects of algorithmic complexity as well as reduce the processing execution time without any loss of information flow nor any resulting imposed restriction as a prerequisite for the structure initiation or apprenticeship skills, as can be noticed in the elaborated works of [10].

In addition, on combining pioneering approach with other BN apprenticeship algorithms, one might well manage to modelize once non-modelizable information systems, thanks to the larges number of variables made available.

As for the remaining constituent sections of the present research work, they are organized as follows: the next section is allotted to the introductory exposition of the BN structure learning problem. As for the following section, a new data reduction approach is going to be presented, which is going to be applied and tested on a special biological data base. As regards the last section, it depicts our conclusion along with the perspectives for further future researches.

## 2 BN and data-structure learning problem

It is worth highlighting that knowledge representation and the related reasoning, thereof, have given birth to numerous models. The graphic probability models, namely, BN, introduced by Judea Pearl in the 1980s, have been manifested in to practical tools useful for the representation of uncertain knowledge, and reasoning process from incomplete information.

To note, a BN  $B = (G, \Theta)$  is written under the form of:

- $G = (X, E)$  graph managed without circuit summits of whose associates a set of random variables  $X = \{X_1, \dots, X_n\}$
- $\Theta = \{P(X_i | P_a(X_i))\}$ , probability set of every knot  $X_i$  conditional upon the state of its parent relatives  $P_a(X_i)$  in  $G$ .

Hence, the BN graphic representation indicates the dependences (or independences) between variables and provides a visual knowledge representation tool, that turns out to be more easily understood by its users. Furthermore, the use of probability allows to take into account the uncertainty, by quantifying the dependences between variables. These two properties have been at the origin of the first terms allotted, initially, of BN, "probabilistic expert systems", where the graph used to be compared with some set rules pertaining to a classic expert system, and conditional probability presented as a quantification measurement of the uncertainty related to these rules [16].

In this respect, Reference [13] has shown that BN have allowed to represent, in a compact way, the joint probability distribution relevant to all variables:

$$P(X_1, X_2, \dots, X_n) = \prod_{i=1}^n P(X_i | P_a(X_i)) \quad (1)$$

Actually, this decomposition of a global function, undergone by a product of local terms exclusively depending on the considered knot and its relative parents in the graph, is a fundamental property of BN. It is on the basis of the early works pertaining to the development of inference algorithms, that the probability of any model variable could be calculated from observation, even partial, of the other variables. This

problem was proved NP-hard, but ended in various algorithms which can be likened to the information distribution methods in a graph. As can be notes, these methods apply the notion of conditional probability, along with the theorem of Bayes, which allows to calculate the probability  $X_j$  from  $X_i$ , and vice versa, knowing that  $P(X_i | X_j)$  [16].

The number of all BN possible structures has been shown to ascend sharply as a super-exponential on the number of variables. Indeed, Reference [24] derived the following recursive formula for the number of Directed Acyclic Graph (DAG) with  $n$  variables:

$$r(n) = \sum_{i=1}^n (-1)^{i+1} \binom{n}{i} 2^{i(n-i)} r(n-i) = n^{2^{O(n)}} \quad (2)$$

which gives:  $r(1)=1$ ,  $r(2)=3$ ,  $r(3)=25$ ,  $r(5)=29281$ ,  $r(10)=4,2 \cdot 10^{18}$

This means that, it is impossible to perform an exhaustive search of all structures in a reasonable time in cases the number of nodes exceeds seven.

## 3 A New approach for optimizing the number of variables prior to modeling

### 3.1 Background

It is well recognized that the strategy based on single variable analyses has a very limited value in elucidating the mechanisms involved in complex phenomena [8]. In this respect, our proposed approach is fundamentally a four-step operating multivariate analysis. It starts by calculating a statistical score (test value or p-value) for each variable depicting its relevance to a certain phenomenon. It then, clusters variables according to their association to the studied phenomenon as well as their complementarity. In the third step, a global statistical score is calculated for each cluster of variables, which is a function of the correlation between the variables and their scores. Ultimately, the clusters will be ranked in a decreasing order based on their global score (following a logarithmic transformation in order to have a high score if the score statistic value is low), so that a number of them can be selected (Figure 1).

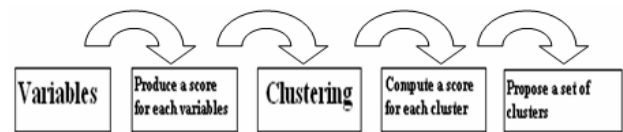


Figure 1. Approach Steps

It is worth mentioning that all the methods and statistical tests that used in this section are available in statistics as well as computer science literature. Noteworthy, however, the novelty lies in their nesting where they have never been built fit before, i.e, in the field of data reduction for learning BN structures.

### 3.2 Single-variable analysis

Chi-square test can validate assumptions raised regarding a certain property contained in the basis of a concrete case base [9]. It is a widely used test applied to measure the association between categorical variables [19]. For cases involving binary variables (two categories), for instance, the disease status and risk factor in epidemiological studies, the Chi-square is easily calculated.

### 3.3 Variables' Clustering

“Clustering is the unsupervised classification of patterns (observations, data items, or feature vectors) into groups (clusters)” [11]. Clustering is an automated process to join related variables into a set in which they are grouped together on the basis of their attributes similar values. There is a variety of algorithms used for clustering, e.g., Generative Models, the Gaussian Mixture Model, C-Means Fuzzy Clustering, Reconstructive Models, K-means...[23]. Clustering can also be achieved based on expert's knowledge in the field [22].

### 3.4 Fusion of separate-cluster scores

In this junction, one might well wonder how to derive a score for each cluster based on the scores of variables within clusters.

Most of the methods used to combine the pertinent scores to computer science literature, and more specifically to knowledge discovery in database, are those that consist in merging such scores of independent variables as: Average and Maximum (MAX) scores [4] and [21], Sum, Minimum (MIN) and Product scores [12].

Yet, the statistical literature provides several score-combining methods by taking into account the correlations among variables. Among these is the Truncated Product Method (TPM) [27] which combines the correlated tests' p-values, whose algorithm is described below:

#### Truncated Product Method (TPM) Algorithm

For each cluster of variables, the following steps are to be undertaken:

- 1: Construct a correlation matrix for variables within the cluster.
  - 2: Calculate the Cholesky matrix  $C$  for each correlation matrix
  - 3: Choose the scores' maximum value  $\pi$  (p-values) to be selected.
  - 4: Calculate  $W_0 = \prod_i^L p_i^{I(p_i \leq \pi)}$
- Where  $L$  designates the number of variables in the cluster
- 5: Put  $A=0$
  - 6: Randomly generate  $L$  independent values from a uniform distribution generating the vector  $R^*: u_1^*, \dots, u_L^* \in [0,1]$
  - 7: Transform the vector  $R^*$  into another vector  $R$  having the values with equation (2):

$$R = 1 - \Phi\{C\Phi^{-1}(1 - R^*)\} \quad (3)$$

$$\Phi(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \quad (4) \quad \Phi^{-1}(x) = \int_x^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \quad (5)$$

- 8: Calculate  $W = \prod_i^L R_i^{I(R_i \leq \pi)}$

9: If  $W \leq W_0$ , then  $A=A+1$

10: repeat steps 6 to 9, B times

11: obtain the combined score (p-value) by means of  $A/B$ .

### 3.5 Ranking

In this way, Clusters of variables can be ranked on the basis of their scores. If a p-value is used as a score, then ranking is based on an increasing-order scoring (smaller p-values are indicative of higher significance). Most often, the score is calculated as the logarithmic transformation  $-\text{Log}_{10}(\text{p-value})$  just as a high score value implies a high degree of significance (association). In this case, the score ranking will be done on decreasing-order.

### 3.6 Selection of a phenomenon closely-related clusters

The purpose of this step is to select the appropriate variables' clusters closely associated to the phenomenon. Actually, there exist numerous methods for selecting most influential variables available in statistical and computer science literature. Nevertheless, to our knowledge, few are those methods that deal with selecting clusters of variables. In this respect, we propose a method inspired from [14], described below.

We consider that among the  $k$  ranked scores obtained for each variables' cluster, the first  $r$  will be selected by applying the following steps.

1: Scores  $S_1, S_2, \dots, S_k$  are used to compute sum statistics as follows:  $T^i = \sum S^j$  where  $i$  varies from 1 to  $k$ .

2: P-values  $(P_T^1, \dots, P_T^k)$  are estimated from the empirical distribution of each  $T^i$  (data are simulated, from a uniform distribution, for a number of times and the p-value is then estimated by the proportion of  $T^i$  values exceeding the observed value  $T_{obs}$ ), thus:

$$P_T = \frac{\text{Cardinality}\{T^B \geq T_{obs}\}}{B} \quad (6)$$

3: The method selects the first  $r$ , where  $r$  corresponds to the first cluster in which a decrease in p-value is initially witnessed  $P_T^{(i+1)} \leq P_T^{(i)}$

that is,  $r = \arg \min(P_T^{(i+1)} \geq P_T^{(i)})$ .

### 3.7 Precepts of genetics and Experimentation

Single Nucleotide Polymorphisms (SNP), indicated in genetics, are variations of a single basic pair (of the same sort) of human genome among individuals. These variations are very frequent (1/1000 pairs of bases in the human genome). The SNP represent 90 % of all the human genetic variations, and SNP with an allelic frequency superior or equal to 1 %, are present in all 100 in 300 basic pairs on average in the human genome, where 2 SNP out of 3 substitute the cytosine with the thymine [28].

Generally speaking, the SNP are bi-allelic (a, A). Every individual will be carrier, at the level of an SNP, of one of the three possible genotypes:

- both homozygous genotypes (aa and AA)
- The heterozygous genotype (aA or Aa imperceptible one of the other). [28]

A gene is a sequence of deoxyribonucleic acid (DNA) which specifies the synthesis of a chain of polypeptide or a ribonucleic acid functional (ARN). We can also define a gene as a unit of genetic information. Therefore, we can say that the DNA is the support of the genetic information. Indeed, it can be considered as a book, an architectural plan of the alive, which directs, dictates the construction of the main constituents and cellular builders which are the proteins. The genotype of an individual (along with that of the animal, plant, bacteria or other) is the sum of the genes which it possesses. As for the phenotype, it corresponds to the sum of the morphological, physiological or behavioural characters which are recognizable from the outside. Consequently, two individuals can have the same genotype but not necessarily the same phenotype, depending on the conditions of expressions of the genes which confer a recognizable, discernible aspect [28].

Genome wide-association studies are which geneticists assess the association of thousands of molecular markers with a disease phenotype. The traditional way of analyzing the data consists in computing the chi-square association tests and the corresponding p-value for each marker. Then, those with the weakest p-values are selected as indicative of interesting genome region, as has been applied on 213 Canadian patients suffering from schizophrenia and 241 Canadian controls, both genotyped for 164 SNPs on chromosome 13.

Eventually, the reached database has been a text file formatted as follows:

```

      213 CASE                241 CONTROLS
-----
Snp 1 0 1 2 1 1 2 0 0 0 0 1 2 2 1 2 0 2 2 2 2 0 0 0 0 1 0 2 2 3 3 2 1 0 0 3 2 1 0 0 0 1 1 2 1 1 1 1 0 3 0 2 2 1 0 2 2 1 0 0 2 1
Snp 2 1 0 2 3 2 0 2 3 2 0 1 0 2 1 2 1 2 1 2 2 3 3 0 0 0 0 0 2 2 1 2 1 0 0 2 0 2 0 2 0 2 1 0 0 3 3 0 2 0 2 1 1 0 1 1 1 1 2 0 0
.....
Snp n 0 2 0 2 2 2 0 1 2 3 3 2 2 0 0 0 1 1 1 2 2 0 2 0 2 1 0 2 1 3 0 2 1 3 0 2 2 1 0 3 0 3 2 1 1 1 0 0 2 1 0 2 1 0 3 2 1 0 1

```

Figure 2. Data base Format.

- Where 0: corresponds to the *aa* genotype,  
 1: corresponds to the *Aa* or *Aa* genotype  
 2: corresponds to the *AA* genotype  
 3: corresponds to missing data

Actually, our objective has been to select genomic regions (clusters of variables) which are most significantly associated to the disease (schizophrenia).

### 3.7.1 Data-Processing Steps

- Calculating a score corresponding to each variable (SNP) which is, in our case, the p-value derived from the chi-square test statistic.

- Variable clustering: we cluster the variable according to the genetic experts' suggesting that a gene might well represent a cluster of SNPs.
- Combine scores from each cluster using the different relevant strategies proposed in Sub-Section 3.4 and compare results.
- Rank the various clusters according to their scores.
- Select clusters involved in the disease (schizophrenia) using the approach described in Sub-section 3.6.

### 3.7.2 Results of different fusion-scoring methods

For the purpose of comparing the p-value combining methods, we have relied on the reference: the region "G72" described by [5], as being the region responsible for Schizophrenia.

TABLE I below depicts the results achieved via both the TPM as well as the MIN Methods. Actually, the discovered genes revealed by these two methods turn out to be very similar and are contained in the "G72" region.

TABLE I. BOTH METHODS' ACHIEVED RESULTS

	Rank	Gene name	Region	Score	Stati Sum	P-value
<b>MIN</b>	1	FOX01	151	1.33	1.33	0.70
	2	NARG1L	140-141	1.20	2.58	0.63
<b>TPM</b>	1	NARG1L	140-141	1.52	1.52	0.15
	2	FOX01	151	1.09	2.62	0.10

In terms of complexity, it clearly appears that the algorithm using the MIN method appears to be the more appropriate choice. Yet, to check the reliability of both methods' results, we turn to study the empirical distribution of the observed minimum p-value ( $P_{\min}^{obs}$ ) via the Monte Carlo simulations, whose principle is the following:

We simulate B times the data by calculating each time relevant minimum p-value ( $P_{\min}^i$ ) and, ultimately, we calculate the overall p-value corresponding to each step through the formula below:

$$P_G = \frac{\text{cardinality}\{P_{\min}^{(i)} \leq P_{\min}^{obs}\}}{B}. \quad (7)$$

Where  $P_{\min}^{obs}$  represent  $P_T^i$  (already explained in the 3.6 section) of the first cluster in which a decrease in p-value.

Thus, the p-value pertinent to the overall process using TPM is equal to 0.09, while it has been 0.41 with respect to the MIN method. One can conclude that the results achieved by means of the TPM turn out to be more significant and that this method is preferably convenient to a subsequent work.

### 3.7.3 Discussion

On applying our innovative approach, we have been able to successfully identify the most significant genes involved in the Schizophrenia disease. Indeed, our attained results have turned out to be conforming to, and to agree with, those published by the specialists in genetics. Above all, we have managed to exclude those genes having no implication, or a very weak relationship, with this illness. As a matter of fact, to our knowledge, the genes eliminated from the information system, subject of study, have not been mentioned by any specialized literature publication pertinent to the genetics field as being involved in the Schizophrenia disease.

Added to this, we have actually been able to reduce the number of initial variables, necessary to this study, from 165 (164 SNP plus one phenotype variable) into four (3 SNP plus a phenotype variable). Thus largely reducing the algorithmic complexity of applying the BN structure, as the number of possible graphs has gone down from  $r(165) \approx 10^{406}$  to  $r(4) = 576$ , without any loss in data information. Yet, we reckon it necessary for our approach to be tested and applied to another data base of similar context, for the sake of validity consideration purposes of the achieved results.

It is worth highlighting, however, that data pertaining to the studies of complex genetic illnesses appear to fit particularly well to our devised approach, seeing the fact that a diverse number of genes would not be involved in certain genetic illnesses. Hence, it could be set, as a proposal for a prospective future research to test this model approach on another domain area data base. Actually, our targeted purpose is to check out the achieved results, above all on those cases where the total numbers of variables appear to be really implicated in the context of certain phenomena.

## 4 Conclusions and suggestions for further research

Our study has defined a novel and appropriately-useful approach for filtering the number of variables in respect of their degree of implications in a given phenomenon. This proposed approach enables to identify clusters of variables that are most frequently involved in a given phenomenon using several steps. This has been illustrated through a simple pertaining to a genetic study on schizophrenia. The comparison of the proposed approach, as a whole, with other similar methods available will be the objective of a prospective publication.

In a future research, we intend to present a new multi-purpose heuristics designed for learning BN structure. Such a process, aimed at reducing the search space for the possible graphs, should be able to combine with the already-existing algorithms and the classic metric-score methods of BN learning. Bound for learning a Bayesian network structure, this heuristic has actually been built on the formalism introduced by [15], called Multi-Entity Bayesian Networks (MEBN). The MEBN formalism unifies the first-order logic jointly with the probability theory. It contains fragments dubbed MFragments, which represent the joint distribution of a

subset of variables. Our principle will be based on the fact that the complexity of learning Bayesian network structure is exponential giving the exponential, increase in the number of variables. Hence, the urgent need for methods allowing to learn the structure with all its contained variables, even when the number of variables is too large. The solution that we reckon to propose would be based on the modulation of learning structure: each cluster has its properly-allotted learning structure, before forming the final single structure encompassing all the variables.

A structure will be devised for a benchmark of databases that depicts the dominating relationships between the selected variables and the phenomenon. Noteworthy, Multi-Entity Bayesian Networks, despite the interest of their use in respect of the complex classical structure-learning algorithms, remain still liable to demonstration.

## 5 References

- [1] V. Nefian, "Learning SNP using embedded Bayesian Networks," IEEE Computational Systems Bioinformatics Conference, 2006.
- [2] L. Bouchaala, A. Masmoudi, F. Gargouri and A. Rebai, "Improving algorithm for structure learning in Bayesian Networks using a new implicit score," Expert Systems with Application, 37, 5470-5475, 2010.
- [3] G. Cooper and E. Hersovits, "A Bayesian method for the induction of probabilistic networks from data," Machine learning, 9, 309-347, 1992.
- [4] M. L. Damian and F. H. Donald, "Combining multiple scoring systems for target tracking using rank-score characteristics," Information Fusion, 10, 124-136, 2009.
- [5] S. Detera-Wadleigh and F. McMahon, "G72/g30 in schizophrenia and bipolar disorder: review and meta-analysis," Biological Psychiatry, 60(2): 106-114, 2006.
- [6] P. Dempster, N. Laird and B. D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," Journal of the Royal Stat Soc B 39: 1-38, 1977.
- [7] O. Francois, and P. Leray, "Evaluation d'algorithmes d'apprentissage de structure pour les réseaux bayésiens," In Proceedings of 14ème Congrès Francophone Reconnaissance des Formes et Intelligence Artificielle, RFIA, pages 1453-1460, Toulouse, France, 2004.
- [8] M. Geudj, J. Wojcik, D. Robelin, M. Hoebeke, M. Lamarine and G. Nuel, "Detecting Local High-Scoring Segments: a First-Stage Approach for Genome-Wide Association Studies," Statistical Applications in Genetics and Molecular Biology, Vol. 5, Iss. 1, Article 22 2006.
- [9] C. Herman and E. L. Lehman, "The use of Maximum Likelihood Estimates in chi-square tests for goodness of fit," The annals of Mathematical Statistics volume 25, Number 3, 579-586, 1954.
- [10] K. Hwang, B. H. Kim and B. T. Zhang, "learning hierarchical Bayesian Networks for large-scale data analysis," In ICONIP: 670-679, 2006.
- [11] K. Jain, M. N. Murty and P. J. Flynn, "Data clustering: A review," ACM Computing Reviews, 264-323, 1999.

- [12] Jain, K. Nandakumar and A. Ross, "Score normalization in multimodal biometric systems," *Pattern Recognition*, volume 38 Issue 12, Pages 2270-2285, Dec 2005.
- [13] P. Judea and V. Tom, "A theory of inferred causation," In James Allen, Richard Fikes and Erik Sandewall, editors, *KR' 91: Principles of knowledge representation and reasoning*, pages 441-452, San Mateo, California, 1991.
- [14] S. Karlin and S. Altshul, "Applications and statistics for multiple high-scoring segments in molecular sequences," *Proceedings of the National Academy of Science USA* 90, 5873-5877, 1993.
- [15] K. B. Laskey, "MEBN: A language for first-order Bayesian knowledge bases," *Artificial Intelligence*, 172, 140-178, 2007.
- [16] P. Leray, "Réseaux Bayésiens: apprentissage et modélisation de systèmes complexes," habilitation à diriger les recherches, Université de Rouen, 2006.
- [17] R. Mourad, C. Sinoquet and P. Leray, "Learning hierarchical Bayesian Networks for genome-wide association studies," In 19<sup>th</sup> International Conference on computational statistics, (COMPSTAT): 549-556, 2010.
- [18] R. Mourad, C. Sinoquet and P. Leray, "A hierarchical Bayesian Network approach for linkage disequilibrium modelling and data dimensionality reduction prior to genome-wide association studies," *BMC Bioinformatics*, ISSN 1471-2105, 2011.
- [19] P. Naim, P. H. WUILLEMIN, P. Leray, O. Ponnet and A. Becker, "Réseaux Bayésiens," Eyrolles, Paris, 3 editions, 2007.
- [20] R. E. Neapolitan, "Learning Bayesian Networks," Newyork, NY, USA: Prentice Hall 2003.
- [21] H. N. Parkash and D. S. Guru, "Offline signature verification: An approach based on score level fusion," *International journal of computer applications*, 0975-8887, Article 10, No.18, 2010.
- [22] A. Peter, G. Patrick and F. Geett, "On the potential of domain literature for clustering and Bayesian Networks learning," *Proceeding of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 405-414, 2002.
- [23] X. Rui, and C. W. Donald, "Clustering," IEEE Press/Wiley, oct 2008.
- [24] R. W. Robinson, "Counting unlabeled acyclic digraphs," *Combinatorial Mathematics*, 622, 28-43, 1977.
- [25] Y. Zhang and L. Ji, "Clustering of SNPs by structural EM algorithm," *International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing*: 147-150, 2009.
- [26] Y. Shulin and K. Chang, "Comparison of score Metrics for Bayesian Networks Learning," *IEEE Transactions on Systems, Man and Cybermetics-part A: Systems and Human*, 32(3), 419-428, 2002.
- [27] D. Zaykin, L. Zhivotovsky, P. Westfall and B. Weir, "Truncated product method for combining P-values," *Genet Epidemiol*, 22(2), 170-85, Feb 2002.
- [28] D. W. Watson, T. A. Baker, S. P. Bell, A. Gann, M. Levine, R. Losick, "Molecular Biology of the gene," Distributed in conjunction with Benjamin cummings, 841 pp, ISBN 978-080539592-1, 2008.