Improving algorithms for knowledge discovery in genetics databases

Heni Bouhamed
University of Rouen
LITIS EA 4108
76821 Mont-Saint-Aignan Cedex, France
Heni_bouhamed@yahoo.fr

Thierry Lecroq
University of Rouen
LITIS EA 4108
76821 Mont-Saint-Aignan Cedex, France
Tierry.Lecroq@univ-rouen.fr

Ahmed Rebai
University of Sfax
Centre of Biotechnologie of Sfax
Sfax, Tunisia
Ahmed.Rebai@cbs.rnu.tn

Maher Jaoua
University of Sfax
MIRACL
Sfax, Tunisia
Maher.Jaoua@fsegs.rnu.tn

***Abstract*—Extracting Knowledge from genetics pertinent data-base has still remained one of the most exciting challenges in the data mining area. To note, most of the association studies have exclusively been devoted to highlight the basic determinant lying behind to a certain specific complex genetic disease. In this respect, the present study's major objective consists in devising a novel knowledge-discovery approach, whereby a genetics data base has been developed. Hence, this work has been primarily designed to propose some improvements to the predominantly applied algorithms, widely-applied in this field. In the second places, we intend to demonstrate that our newly-devised algorithm, dubbed NCA, has by far achieved highly accurate and effective results in respect of the prevalent algorithms. As a matter of fact, we have, willingly, applied and compared our approach, along with the existing approaches, to some biological ideas relevant to some acute hereditary complex illnesses, in which the concerned biological literature has identified the pertinent responsible variables. As for the last-section part, it depicts our concluding and proposed suggestions for further research.**

Keywords-component; association studies; clustering; knowledge discovery; genetic database; local score.

I. INTRODUCTION

It is a well-known fact that the major aim of genetic epidemiology is to identify the relevant biological mechanisms responsible for certain human diseases. Actually, genome-wide association studies have been made possible, thanks to some recently-undertaken improvements introduced in the field of genotyping technologies [4]. Actually, genome-wide associations studies seek to detect a potentiality possible association between a genetic variant and a disease, not only at a familial level but also at a whole population level. Naturally, it falls into case-control framework: the association between genetic variant and case-control status can easily be established by means of classical epidemiological methods relevant to associations' studies [18].

In this framework, Reference [4] has proposed a local-score based algorithm concept, called Local High-scoring Segments for Association (LHiSA). It has showed better results than the single marker analyses as corrected by Bonferonni [17]. The use of the association local scores studies turns out to be so simple and quick that it can easily detect a disease associated genomic regions. Noteworthy, a local score is based on the accumulation of statistical-association high values, including one or more neighbour markers [4]. In fact, on applying the LHiSA algorithm on a case-control database (whose cases are suffering from Schizophrenia disease), some areas, confirmed by the specialized literature relevant to the Schizophrenia disease, have been identified. However, the overall significance of the approach is far from being convincing to the disease associated regions' set. In this respect, we reckon to propose, within the framework of the present paper, a prior-clustering markers' based approach, whereby markers are set accurate to their affiliation to the genes, and each clusters' statistical signals are combined with respect to the "Truncated Product Method" [11]. It is also worth noting that our newly designed approach has been compared to that applying LHiSA algorithm on the same database.

Regarding the remaining constituent sections of the present research work, they are organized as follows. The next section is allotted to the introductory genetics precepts. In the following section, the LHiSA algorithm is going to be presented. As for the fourth section, our newly-conceived algorithm is going to be presented and compared to the LHiSA algorithm, which are both going to be applied and tested on a specific biological database framework. As regards the last section, it depicts our conclusion along with some perspectives for further future researches.

II. GENETICS PRECEPTS

It is worth noting that the Single Nucleotide Polymorphisms (SNP), widely-applied in genetics, are variations of a single basic pair (of the same sort) of human genome highly perceptible among individuals. These variations are very frequent (1/1000 pairs of bases in the human genome). Actually, the SNPs represent 90 % of all the human genetic variations, and with an allelic frequency superior or equal to 1 %, they are present in every 100 out of 300 basic pairs, on average, in the human genome, where 2 SNPs out of 3 substitute the cytosine with the thymine [19].

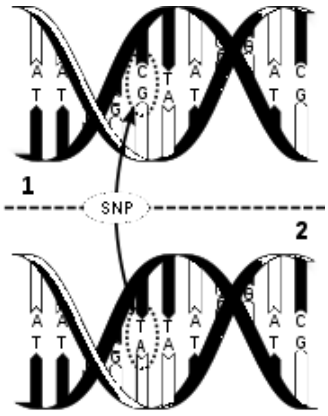

Figure 1. Single Nucleotide polymorphisms

Generally speaking, the SNPs are bi-allelic in nature (a, A). As a matter of fact, every individual will be a carrier, at the level of an SNP, of one of the three possible genotypes:

- both homozygous genotypes (aa and AA)
- The heterozygous genotypes (aA or Aa, imperceptible one of the other). [19]

Noteworthy, a gene is a sequence of a deoxyribonucleic acid (DNA) which specifies the synthesis of a chain of polypeptide or a functional ribonucleic acid (ARN). Besides, a gene can also be defined as a unit of genetic information. One can, therefore, say that the DNA is the basic support or original background of the genetic information. Indeed, it can be considered as a book, an architectural plan of the alive, which directs and dictates the construction of his main constituents and cellular builders, i.e., the proteins. The genotype of an individual (along with that of the animal, plant, bacteria or others) is the sum of the genes which it possesses. As for the phenotype, it corresponds to the sum of the apparent 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 genes' expression conditions which confer a special recognizable, discernible aspect [19].

III. THE LHISA ALGORITHM

A. Local-score definition

Let $X=(X_i)_{i=1,\dots,n}$ be a sequence of real random variables:

$$H = max_{1\le i\le j\le n}\left(\sum_{i}^{j} X_k\right) \quad (1)$$

Such definition of the local score restrains the search for the highest scoring sequence. Noteworthy, however, data may not contain, unique or one single region of interest. Yet, despite the fact that the next highest scoring sub-sequences are as potentially interesting, it turns out to be more convenient to look for disjoint segments. Hence, we consider $H_1 \ge \dots \ge H_k$ as being the scores of the k first and distinct highest-scoring segments [4].

B. The First Step

In the first place, we assign an "individual score" X_i to each marker, where a high score designates a high chance of association to the disease. This score is calculated by means of a chi square test based on the allelic contingency table. Resultant p-value will be transformed into $X_i = -log_{10}(p_i)$. X_i will be transformed into $X_i' = X_i - \delta$, where $\delta(\delta=0,1 or\ 0,05)$ will improve the cumulative score of a given segment, while the threshold will penalize it.

C. The Second Step

The aim is, then, to identify the best high-score segment, with a local score concept (refer to local score definition). In this framework, the algorithm of [20] is used for the purpose of implementing and executing the highest local-score set.

D. The Third Step

On a third step, an appropriately selective method of the most interesting segments has been devised. Reference [6] has proposed the below-described method.

Given $H_1 \ge \dots \ge H_k$, the successive local-scores (explained in sub-section *A*) have been combined into new sum statistics $T^1,\dots,T^k$, with $T^i = H^1 + \dots + H^i$. As for the Corresponding p-values $P_{T1},\dots,P_{TK}$, they will be computed via the Monte Carlo simulations (permuting case and control labels). Finally, we resort to select the most interesting segments as being the r first ones, with $r = argmin_i(P_{Ti+1} > P_{Ti})$ and an associated statistic $P_{\min}^{obs} = P_{Tr}$ suitably adopted for this selection.

E. The Fourth Step

As a last step, it consists in assessing the global significance, P_G, of the process via Monte Carlo simulations. We reiterate N times steps 1 to 3, per mutating each time case and control labels, and computing P_{min}. At last, the whole procedure's p-value turns out to be:

$$P_G = \frac{card\left\{P_{\min} < P_{\min}^{obs}\right\}}{N} \quad (2)$$

Where $P_{\min}^{obs}$ is the P_{min} initially calculated, in the first time, with the real data.

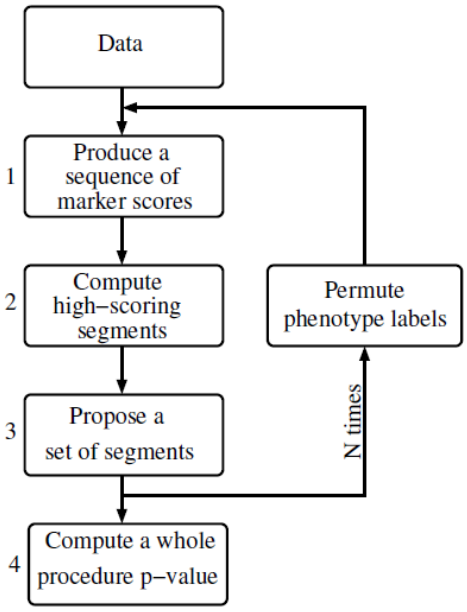

Figure 2. LHiSA Algorithm.

IV. New Algorithm with a Priori Clustering (NCA)

Regarding our conceived approach, we will primarily rely on the notion that a gene is the unit of genetic information. Each gene is formed by a set of markers that interact together [19]. Accordingly, one might well believe that the clustering of markers according to their genes belonging seems to be more appropriate than applying the local score (the local score ignores this very important concept pertinent to the knowledge discovery field in genomic data). Actually, our novel contribution to the field will highly prevail throughout the second step of the LHiSA algorithm, while the same concepts will be preserved and abided by throughout the other steps (Fig. 3). Thus, during the second step, we first resort to cluster genetic markers according to their gene membership. Secondly, marker-gene's scores are going to be combined into a single score that will represent this gene. Ultimately, these genes will be classified according to their scores in descending order.

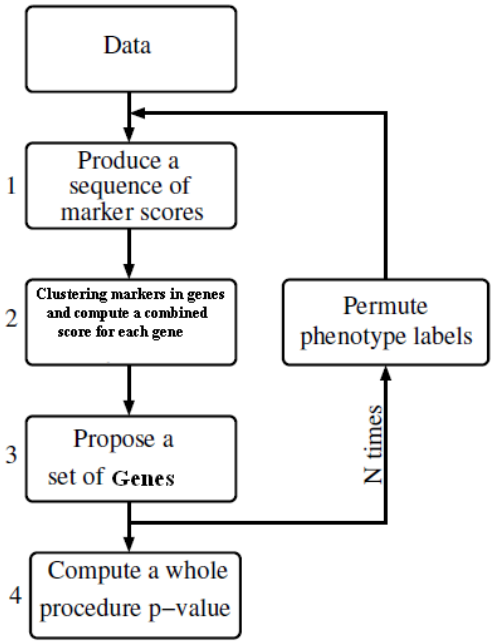

Figure 3. NCA Algorithm.

To note, the algorithm we intend to use so as to combine each single-gene's statistical scores is dubbed as the "Truncated Product Method" (TPM) [11] described below.

Truncated Product Method (TPM) Algorithm

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

1: Constructing a correlation matrix for variables within the cluster.

2: Calculating the Cholesky matrix C for each correlation matrix

3: Choosing the scores' maximum value π (p-values) to be selected.

4: Calculating $W_0 = \prod_i^L p_i^{I(p_i \le \pi)}$

Where L designates the number of variables in the cluster

5: Putting $A=0$

6: Randomly generating L independent values from a uniform distribution generating the vector $R^*: u_1^*, \ldots, u_L^* \in [0,1]$

7: Transforming the vector R^* into another vector R having the values with equation:

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

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

8: Calculating $W = \prod_i^L R_i^{I(R_i \le \pi)}$

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

10: repeating steps 6 to 9, B times

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

V. Applying and Comparing of Both Algorithms' Results

It is worth highlighting, that both the LHiSA Algorithm and our newly-designed algorithm have been applied on 213 Canadian patients suffering from schizophrenia as well as on 241 Canadian controls, both genotyped for 164 SNPs on chromosome 13.

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

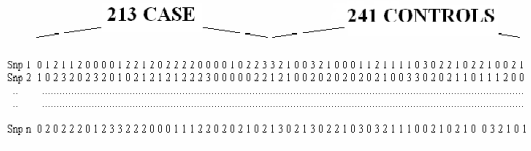

Figure 4. Data base Format.

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 with the disease (schizophrenia).

TABLE I below depicts the results achieved via both the LHiSA as well as the NCA Methods. In fact, the discovered regions revealed by these two algorithms turn out to be very similar and are contained in the "G72" region [2].

TABLE I. BOTH ALGORITHMS' RESULTS

	Rank	Gene name	Region	Score H	SUM T	P_T	P_G
LHiSA	1	---	140-144	2.54	2.54	0.24	0.22
	2	---	150-152	1,97	4.52	0.17	
NCA	1	NARG1L	140-141	1,52	1.52	0.15	0.09
	2	FOX01	151	1.09	2.62	0.10	

As can be noticed, both algorithms identify two areas belonging to the "G72" region, cited in the specialized literature as being responsible for the Schizophrenia disease [2]. Yet, the global p-value of our algorithm application (equal to 0.09) appears to be more meaningful and convincing than the LHiSA algorithm global p-value (equal to 0.22). Hence, one can well conclude that our algorithm has provide a noticeable improvement in

results. Actually, this might well be due to the importance of the gene concept in dealing with the genome-sequence studies.

CONCLUSION

Our algorithm has enabled to identify genes that are most frequently involved in a given disease by following certain relevant steps. This has been illustrated through a sample pertaining to a genetic study on Schizophrenia. Still, the proposed algorithm can be used to filter the number of variables in respect of their degree of implications in a given phenomenon prior to learning a Bayesian Network structure. The aim has been to reduce the computational complexity of such a process.

In this framework, it is worth mentioning that we intend to present a new multi-purpose heuristics, designed for learning BN structure, as a future research subject. Such a process, aimed at reducing the search space for the possible potentially graphs, should be able to combine with the already-existing algorithms and the classic metric-score methods of BN learning. Bound to help learn a Bayesian network structure, this heuristic method has actually been built upon the formalism introduced by [7], labelled Multi-Entity Bayesian Networks (MEBN). Noteworthy, the MEBN formalism has been designed to unify the first-order logic jointly with the probability theory. It contains fragments dubbed MFrags, which represent the joint distribution of a subset of variables. Our principle will be based on the fact that the complexity associated with of learning a 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 relevant 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 further demonstration.

REFERENCES

[1] M. L. Damian and F. H. Donald, "Combining multiple scoring systems for target tracking using rank-score characteristics," Information Fusion, 10, 124-136, 2009.

[2] S. Detera-Wadleigh and F. McMahon, "G72/g30 in schizophrenia and bipolar disorder: review and meta-analysis," Biological Psychiatry, 60(2): 106-114, 2006.

[3] 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.

[4] 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.

[5] A. Jain, K. Nandakumar and A. Ross, "Score normalization in multimodal biometric systems," Pattern Recognition, volume 38 Issue 12, Pages 2270-2285, Dec 2005.

[6] 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.

[7] K. B. Laskey, "MEBN: A language for first-order Bayesian knowledge bases," Artificial Intelligence, 172, 140-178, 2007.

[8] 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.

[9] 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.

[10] R. W. Robinson, "Counting unlabeled acyclic digraphs," Combinatorial Mathematics, 622, 28–43, 1977.

[11] D. Zaykin, L. Zhivotovsky, P. Westfall and B. Weir, "Truncated product method for combining P-values," Genet Epidemiol, 22(2), 170-85, Feb 2002.

[12] O. François, "De l'identification de structure de réseaux bayésiens à la reconnaissance de formes à partir d'informations completes où incompletes, "Thèse de doctorat. Institut National des Science Appliquées de Rouen, 2006.

[13] P. Leray, "Réseaux Bayésiens: apprentissage et modélisation de systèmes complexes," habilitation à diriger les recherches, Université de Rouen, 2006.

[14] B. Efron, "The lenght heuristic for simultaneous hypothesis tests," Biometrica, 84, 143-157, 1997.

[15] 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.

[16] X. Rui, and C. W. Donald, "Clustering," IEEE Press/Wiley, oct 2008.

[17] Y. Benjamini, and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing." Journal of the Royal Statistical Society B 57, 289–300, 1995.

[18] N. Breslow, and N. Day, "Statistical methods in cancer resarch." Schlesselman, 1982.

[19] 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.

[20] W. Ruzzo, and M. Tompa, "A linear time algorithm for finding all maximal scoring sub-sequences." In 7th Int. Conf. Intelligent Systems for Molecular Biology , pages 234_241, 1999.