Gene Clustering using Independent Component Analysis

Michel Sarkis[1], Zaher Dawy[2], Joachim Hagenauer[3], and Jakob C. Mueller[4]
[1]Munich University of Technology, Institute for Data Processing (LDV), Munich, Germany
[2]American University of Beirut, Department of Electrical and Computer Engineering, Beirut, Lebanon
[3]Munich University of Technology, Institute for Communications Engineering (LNT), Munich, Germany
[4]Munich University of Technology, Institute for Medical Statistics and Epidemiology, Munich, Germany
Email: michel@tum.de, zaher.dawy@aub.edu.lb, hagenauer@tum.de, jakob.mueller@imse.med.tu-muenchen.de

***Abstract*— Linkage disequilibrium has gained a lot of attention recently since it can be effectively utilized in various problems in the field of statistical genetics, for example gene mapping and evolutionary inference. In this work, we propose and analyze a new algorithm for linkage disequilibrium based on independent component analysis (ICA). The results comply with results obtained using other published methods. However, the proposed algorithm is able in some cases to discover new patterns due to the inherent properties of ICA and is more robust compared to other techniques since it estimates the missing values.**

I. Introduction

Two genetic markers are said to be in linkage disequilibrium (LD) if their alleles are dependent. LD increases the fine mapping ability of complex diseases since the single nucleotide polymorphisms (SNPs) that were skipped in genotyping could be located by dependency. However, the ability to detect these associations depends on several factors such as the properties and the locations of the SNPs [1]. Various methods have been developed to measure LD in genomic data. Some are limited to single marker analysis [2], [3] while others like [4], [5] investigate the dependency among groups of polymorphisms. The latter class of techniques is capable of better capturing the genetical variations among individuals.

In this direction, we propose a novel algorithm that is capable of detecting the intragenic dependencies based on independent component analysis (ICA). ICA is a well-known signal processing technique that tries to extract the components that are nearly independent of each other from a mixture of signals, e.g. see [6], [7]. Hence, this property will be employed to capture the dependencies among different groups of SNPs. In other words, ICA will be used to identify the dependent genes by considering them as one component.

Section II presents how the data was modeled to use ICA. Section III shows the different steps of the proposed algorithm. Section IV illustrates the results obtained by testing the algorithm. Finally, Section V draws some conclusions.

II. Problem Model

Given a study comprising a population of N individuals where for each individual a bi-allelic SNP sequence of length M is provided. We assume that the SNPs have been transformed by some unknown means to form some independent SNP expressions, e.g. see Fig. 1.

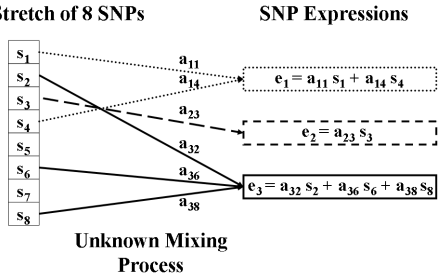

Fig. 1. An example of SNPs transforming to SNP expressions.

Assuming that the expressions are independent sources and the transformation process is a mixing environment, the problem can be translated into an ICA based problem where a set of SNP expressions that are almost independent of each other have to be estimated along with some mixing environment. Consequently, the SNPs that contribute to one expression will be dependent while the others are independent. The expressions that have to be estimated by ICA must be non-Gaussian distributed for the algorithm to work. In addition, the model is assumed to be linear. This is motivated by recent results in microarray data analysis [8], [9] and in gene mapping [10].

III. The Gene Clustering Algorithm

The proposed clustering algorithm is demonstrated in Fig. 2. Besides ICA, it is composed of three additional steps. First, the missing values due to genotyping errors should be estimated. This is performed by using Bayesian principal component analysis (BPCA) since it gives lower error rates when compared to other techniques and it does not require any model assumption [11]. Next, the number of components (clusters) to

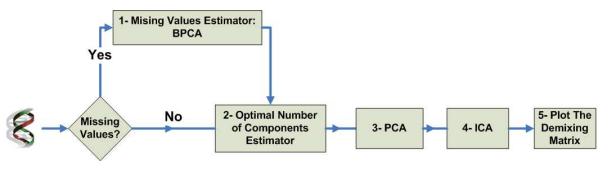

Fig. 2. Blind gene clustering algorithm.

be determined by ICA should be computed. The methodology used is similar to the one implemented by [12] in factor analysis. However, this method has been adapted to be compatible with PCA [10]. What has to be done is to compute the SNP covariance matrix, perform its singular value decomposition (SVD), and calculate its approximate covariance matrix by removing one or more singular values. The optimal dimension is found by taking the minimal dimension where the standard deviation of the error difference between the two matrices is less than that of the standard deviation of a distribution with zero correlation, i.e. $\sigma_{r=0} = N^{-1/2}$. The dimension found is then passed to the third step which is PCA for whitening and dimension reduction. Finally, the new transformed data is processed with ICA. The plot of the obtained demixing matrix will reflect the clusters of SNPs.

IV. Results and Analysis

We have tested the algorithm on various data sets. We present results for two clinical data sets. In the ICA block, we used the FastICA algorithm [7]. The first study was performed on the data set of the T-lymphocyte regulatory genes used in [13]. It contains a sequence of 108 SNPs genotyped from 1036 individuals. Among all the genotype values, 2.87% are missing which we estimate using BPCA. The optimal number of clusters is found to be 9 and the total variance is 74% of the total one. Due to space limitations, Fig. 3 illustrates the plot of only one of the obtained clusters where every SNP location in kilo-base pairs (kbp) is plotted versus its contribution (SNP factor). Compared to Fig. 1 in [13], it is shown that the proposed algorithm detects successfully the LD block. The presence of other clusters in the final result suggests that there are other clusters in the data. This outcome could not be achieved by the other algorithm.

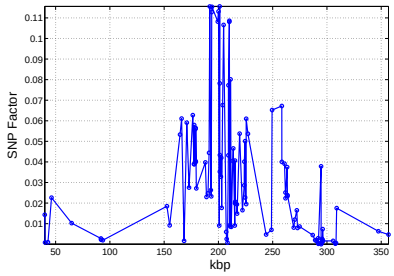

Fig. 3. Outcome of the proposed algorithm: Plot of the component relative to Fig. 1 in [13].

The second tested data set is described in [14]. It comprises a sequence of 103 bi-allelic SNPs collected from 387 individuals where 10% of the total genotypes are missing. The authors categorized the data into 11 major blocks. The proposed clustering algorithm was capable of detecting a subset of these blocks. However, some of the obtained clusters captured more than one block. This is due to the fact that ICA tries to find similarities among any combination of SNPs, a job that is not performed by the other techniques. Fig. 4 shows the plots of the third block (SNPs 16 to 24) and seventh block (SNPs 46 to 76) where each SNP position is plotted versus its contribution.

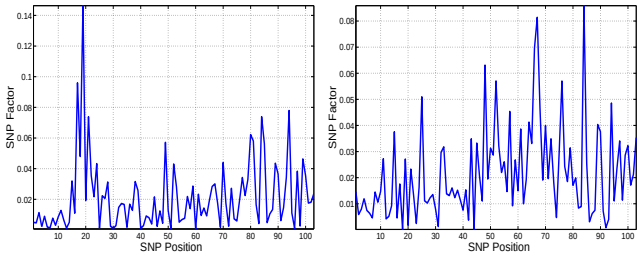

Fig. 4. Outcome of the proposed algorithm for the data set from [14]. Left: Plot of component relative to block 3 in [14]. Right: Plot of the component relative to block 7 [14].

V. Conclusions

This paper exploited the use of ICA in gene clustering. ICA was selected due to its capability of finding components that are as much as possible independent while PCA alone finds the components that are only uncorrelated but not necessarily independent [6]. Nevertheless, PCA was necessary to reduce the dimension of the data. The proposed algorithm estimates the missing values eliminating the need to neglect samples and determines the number of clusters automatically taking into account the structure of the data. Results obtained comply with other techniques with the capability of detecting more LD combinations in some cases.

References

[1] D. J. Balding, M. Bishop, and C. Cannings, *Handbook of statistical genetics*, John Wiley and Sons, Chichester, 2001.

[2] S. W. Guo, "Linkage disequilibrium measures for fine-scale mapping: a comparison," *Human Heredity*, vol. 47, no. 6, pp. 301–314, November 1997.

[3] L. B. Jorde, "Linkage disequilibrium and the search for complex disease genes," *Genome Research*, vol. 10, no. 10, pp. 1435–1444, October 2000.

[4] M. J. Daly, J. D. Rioux, S. F. Schaffner, T.J. hudson, and E. S. Lander, "High-resolution haplotype structure in the human genome," *Nature Genetics*, vol. 29, no. 2, pp. 229–232, October 2001.

[5] B. D. Horne and N. J. Camp, "Principal component analysis for selection of optimal SNP-sets that capture intragenic genetic variation," *Genetic Epidemiology*, vol. 26, no. 1, pp. 11–21, January 2004.

[6] A. Hyvaerinen, J. Karhunen, and E. Oja, *Independent component analysis*, Wiley, New York, USA, 2001.

[7] A. Hyvaerinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Trans. on Neural Networks*, vol. 10, no. 3, pp. 626–634, May 1999.

[8] W. Liebermeister, "Linear modes of gene expression determined by independent component analysis," *BIOINFORMATICS*, vol. 18, no. 1, pp. 51–60, February 2002.

[9] S. Lee and S. Batzoglou, "Application of independent component analysis to microarrays," *Genome Biology*, vol. 4, no. 76, October 2003.

[10] Z. Dawy, M. Sarkis, J. Hagenauer, and J. C. Mueller, "A novel gene mapping algorithm based on independent component analysis," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005)*, March 2005.

[11] S. Oba, M. Sato, I. Takemasa, M. Monden, K. Matsubara, and S. Ishii, "A bayesian missing value estimation method for gene expression profile data," *BIOINFORMATICS*, vol. 19, no. 16, pp. 2088–2096, November 2003.

[12] A. Machado, J. Gee, and M. Campos, "Visual data mining for modeling prior distributions in morphometry," *IEEE Signal Processing Mag.*, vol. 21, no. 3, pp. 20–27, May 2004.

[13] H. Ueda et al, "Association of the T-cell regulatory CTLA4 with susceptibility to autoimmune disease," *Nature*, vol. 423, pp. 506–511, May 2003.

[14] M. J. Daly, J. D. Rioux, S. F. Schaffner, T.J. hudson, and E. S. Lander, "High-resolution haplotype structure in the human genome," *Nature Genetics*, vol. 29, no. 2, pp. 229–232, October 2001.