A SNP Prioritization Method Using Linkage Disequilibrium Network for Disease Association Study

Erkhmbayar Jadamba and Miyoung Shin
Bio-Intelligence & Data Mining Lab.,
Graduate School of EECS,
Kyungpook National University,
Daegu, South Korea
e-mail: erkhembayar@knu.ac.kr, shinmy@knu.ac.kr

***Abstract*—The problem of identifying and prioritizing various types of genetic markers including *single nucleotide polymorphisms* (SNPs), which are involved in human diseases such as cancer, is a one of primary challenge in current disease association studies. In this work, we propose a prioritization method, SNPRank that employs linkage disequilibrium (LD) network to improve the prioritization of candidate SNPs in disease association study. For the construction of LD network structure, we defined mutual links between SNPs based on $r^2 >$ 0.6, and prioritized such SNPs that are linked to other highly ranked SNPs. For experiments, we applied our method to identify SNP markers associated with prostate cancers. The results showed that the proposed method can improve upon existing approaches by newly finding disease related SNPs which could not be identified by existing approaches.**

Keywords-SNP marker; disease association study; linkage disequilibrium network; SNP ranking.

I. INTRODUCTION

After completion of Human Genome Project in 2003 [1], most of researchers were interested in specific areas which are varied between individuals to individuals. Out of all the genetic variations, a *single nucleotide polymorphism* (SNP, pronounced snip) is known to contribute to 90% of them with being almost uniformly distributed across the genome. The SNP is a DNA sequence variation occurring when a single nucleotide –A, T, G, or C- in genome (or other shared sequence) differs between members of a biological species [2]. In recent disease association study, the presence of certain SNPs is often used as a significant clue to identify gene markers which predispose individuals to specific diseases. That is, some SNPs can be involved in increasing the risk of human disease, although most SNPs are not responsible for causing a particular disease phenotype. Thus, the problem of identifying such SNPs that are associated with disease in humans is a major task of disease association studies.

In this paper, we have overviewed current existing methods such as *single SNP analysis* methods and introduced our new approach in order to solve existing approach problems. In last section, we have showed that by allowing the usage of LD based network construction, SNPRank improves the performance over the state-of-the-art ranking method such as GWAS approach [3].

II. RELATED METHODS

. Most of existing methods use *single SNP analysis*, which include a chi-square test, Fisher`s test and Cochrane-Armitage trend test [3]. In these approaches, candidate SNPs are ranked based on the statistical significance of the test and top few SNPs are chosen to be highly associated with the phenotype.

A. Cochraen- Armitage Trend Test

Cochrane-Artimage test for trend, named for William Cochran and Peter Artimage, is used in categorical data analysis when the aim is to assess for the presence of an association between a variable with two categories and variable with k categories [4].

$$T = \sum_{i=1}^{k} t_i (S_{1i} R_2 - S_{2i} R_1) \qquad (1)$$

Trend test statistic can be shown as in (1). In genetic application, the weight t_i can be different according to genetic models described in [3]. In order to test allele is dominant A over allele B, the choice is: t = (1,1,0); if we assume Allele A is recessive to allele B, the choice is: t = (0,1,1).To test whether alleles A and B are codominant, the choice is: t = (0,1,2) [4]. In disease association study, the additive (or codominant) version of the test is mainly used.

However, when number of SNPs are in millions, statistical significance of each SNP would be too small to rely on; this leads to the difficulty in finding significant SNPs in top ranked results. To solve such problems, in this work, we propose a new SNP ranking method, called *SNPRank*.

III. PROPOSED METHOD ‘SNPRANK’

The newly proposed method SNPRank is taking some ideas from Google`s popular PageRank [5] algorithm. Adapting this concept in bioinformatics field was firstly attempted on gene expression data analysis with GeneRank [6] algorithm by Morrison et al. in 2005. Here, our method employs *linkage disequilibrium* [7][8] based network structure along with ordinary GWAS test result to produce an efficient prioritization of the SNPs in a disease association study. In particular, SNPRank method attempts to improve ranking results in such a way that relative ranking of a SNP makes it higher if it is linked to other highly connected SNPs.

ISBN: 978-1-61208-224-0

A. LD Network Construction

Network construction can be summarized into following steps.

- Order candidate SNPs according to chromosome position value
- Calculate LD values (r^2) [7][8] between two SNPs
- Define each SNPs as nodes on network structure
- If r^2 between two SNP is greater than the threshold add the edge between the SNPs to the network
- Build adjacent matrix for SNPRank

Our aim here is to construct a network structure by using correlation between SNPs. The correlation between two SNPs can be estimated by using ***r square* measurements** [7][8], which can be obtained by using (2), between them.

$$r^2 = \frac{D^2}{P_A \times P_a \times P_B \times P_b} \tag{2}$$

where P_A , P_B , P_a , P_b are frequency of each allele and *D is LD measurement* defined by [6]. When the two alleles are not independent, we consider them to be in a state of linkage disequilibrium (LD). When the dependence between SNP is high, the two SNPs are considered to be in a state of high LD. After estimating the LD measurements we constructed network structure and considered each SNPs as nodes in the graph structure. We assumed there that there is an edge between SNPs if r^2 between two SNP is greater than ≥ threshold. We have tried different threshold values in range of (0.2 to 0.9), see Table I. SNPs are presented as a node in network structure. From the network structure, we have built the adjacent matrix(*4*) structure which is used as an input in in our SNPRank.

B. SNPRank

Letting $r_j^{[n]}$ denote the ranking of SNP *j* after the n^{th} iteration, it is defined by

$$r_j^{[n]} = (1-d) \times tr_j + d \times \sum_{i=1}^{N} \frac{w_{ij} \times r_i^{[n-1]}}{\deg_i} \tag{3}$$

Here, tr_j denotes ordinary GWAS test statistic of i^{th} SNP and w_{ij} denotes an element of the adjacent matrix *W* representing LD network on candidate SNPs. In particular, $w_{ij} = w_{ji} = 1$ if *i* and *j* are adjacent and $w_{ij} = w_{ji} = 0$ otherwise. Also, $d \in (0,1)$ is a control parameter which is to define the weight of network structure reflected to calculate ranking statistic.

The value *d* = 0.80 is appears to be used by Google. From previous studies, *d* = 0.6 gave the best result in GeneRank algorithm in case of gene expression data [5].

$$\deg_i = \sum_{j=1}^{N} w_{ij} \tag{4}$$

Formula (4) indicates the degree of i^{th} SNP. The SNPRank method proceeds iteratively, updating the ranking for *j* th page from $r_j^{[n-1]}$ to $r_j^{[n]}$ according to the formula (3).

IV. Experiments and Results

A. Dataset

For experiments, we have used dataset from GSE [8], which include genotype called data profiles of 20 prostate cancer tumors paired with normal samples for 500568 SNPs. For evaluation, we counted how many *truly disease related SNPs* are in top n-ranked result by using prostate cancer related gene list [9] as gold standard. That is, SNPs are considered biologically meaningful if its associated genes match with any one of *gold standard genes* [10].

B. Results

To obtain better result, we implemented matching process in different ranges of parameter *d* and r^2. The best improvement of performance was when $r^2 \geq 0.6, d = 0.5$ when comparing current approach. We have implemented SNPRank, when $\boldsymbol{r^2} \geq$ in range of [0.4 to 0.9] and ***d*** is in range [0 to 1]; if *d = 0*, the ranking returned is based on solely on the absolute value of Cochrane-Armitage test results for that SNP. For *d = 1*, we return the ranking based on Linkage Disequilibrium Network connectivity. By setting *d* in the range [0 to 1], we interpolate between two extremes.

TABLE I. PERFORMANCE SENSIVITY TO R^2, WHEN D=0.5

d = 0.5	**50**	**100**	**150**	**200**	**250**	**300**	**400**	**500**
Cochrane Rank	**4**	**6**	**10**	**11**	**13**	**16**	**19**	**21**
SNPRank								
r^2>0.5	3	7	8	11	12	12	18	21
r^2>0.6	4	7	10	13	16	18	20	23
r^2>0.7	4	9	10	11	12	15	17	20
r^2>0.8	4	7	11	11	13	16	19	22
r^2>0.9	4	7	9	11	12	14	18	21

Since the choice of *d =0.5* was suggested in original GeneRank algorithm, we have checked performance sensitivity to the choice of $\boldsymbol{r^2}$. In Table I, column heads represent top rank SNPs in range of 100 to 500. We compared how many '*gold standard'* genes are matched in top SNPs in two prioritization method classical Cochrane Rank and new SNPRank. We noted the best performance was when $r^2 \geq 0.6$.To evaluate the performance for novel SNP identification we compared the SNP ids and its associated genes for SNPRank with GWAS Cochrane Test ranking. Comparison was performed for top 50 SNPs to 500 SNPs when $r^2 \geq 0.6, d = 0.5$ in Table II.

TABLE II. COMPARISON OF THE EXISTENCE OF PROSTATE CANCER GOLD STANDARD SNPS AND GENES IN SNPRANK AND GWAS RESULTS: O - EXIST , X- NOT EXIST , RED – SNPS NOT IN GWAS RESULT, GREEN – GENES NOT IN GWAS RESULT

		Top 500 SNPs	
SNPs(rs ID)	**Gene Name**	**SNPRank**	**GWAS rank**

rs41488045	NR5A2	O	O
rs41330844	CDH9	O	O
rs17162712	NR5A2	O	O
rs41401450	RNASEL	O	O
rs4261554	CDH8	O	O
rs41498345	HK2	O	O
rs6801782	FHIT	O	O
rs16966932	CDH8	O	O
rs1448988	FGF16	O	O
rs8047093	CDH8	O	O
rs4287583	**CDH8**	O	X
rs231150	**TRPS1**	O	X
rs1019731	IGF1	O	O
rs34011899	**CDKN2A**	O	X
rs41517846	MYC	O	O
rs7194529	CDH1	O	O
rs395920	CDH13	O	O
rs41348046	TRPS1	O	O
rs17098265	PRKCH	O	O
rs10079737	**CDH9**	O	X
rs9936929	**CDH13**	O	X
rs5749939	**MAPK1**	O	X
rs6560010	**DAPK1**	O	X

V. Conclusion

In this work, we have addressed the problem of ranking and prioritizing biomarkers called SNPs which are the most common form of genetic variations on the human genome, and they have been widely used as genetic markers for studying common and complex human diseases. The tremendous number of SNPs, which is estimated at more than eleven million, poses new challenges for discovering and ranking procedures associated with such studies. Our purpose is to support effective disease association studies by providing operational prioritization methods for SNP markers based on both their allele frequency information and Linkage disequilibrium measurement. To achieve this purpose , we have proposed a novel integrative approach, SNPRank method, which allows us to combine linkage disequilibrium based SNP connectivities and conventional rank statistics to produce more robust SNP markers in disease association study, compared with traditional methods only based SNP genotype frequency. In particular, with $d = 0.5$ when $r^2 \geq 0.6$ is used, we observed no deterioration and overall improvement over original Cochrane-Armitage test results. Also, our new method SNPRank incorporated with LD network structure was shown to improve GWAS performance by newly identifying some of *truly disease related SNPs*, which include rs4287583, r231150, rs34011899, rs10079737, rs9936929, rs5749939, and rs6560010. In addition, our SNPRank identified new genes (e.g., TRPS1, CDKN2A, CDH9, CDH13, MAPK1, DAPK1) in top ranks, which could not be identified by conventional approach.

VI. Future Work

The work described in this paper comprises one step toward the goal of identifying disease variants, SNP, which underlying human diseases. For extending the work, we are interested in conducting simulation studies to examine the performance of the proposed method under various genomic experimental conditions, e.g., using the Next Generation Sequencing data. Finally, we mention the main lines of research of prioritizing genetic variation for certain disease will be still remain open for us after finishing this paper. In future, our particular would be using Next Generation Sequencing methods for identifying and prioritizing bio-markers in common and complex human disease.

Acknowledgment

This work was supported by the Korean Research Foundation of Korea (KRF) grant funded by the Korea Government (MEST) (No. 2009-0067724).

References

[1] Collins, F. S., M. Morgan, and A. Patrinos, "The Human Genome Project: Lessons from Large-Scale Biology." *Science* 300, no. 5617 (Apr 11 2003): 286-90.

[2] Wikipedia, The Free Encyclopedia. Wikimedia Foundation, Inc.. 22 Jan 2004. Web. 04 Dec 2011. <http://en.wikipedia.org/wiki/Single-nucleotide_polymorphism>

[3] Lewis, C. M, "Genetic Association Studies: Design, Analysis and Interpretation." *Brief Bioinform* 3, no. 2 (Jun 2002): 146-53.

[4] Wikipedia, The Free Encyclopedia. Wikimedia Foundation, Inc.. 22 Jan 2004. Web. 04 Dec 2011. <http://en.wikipedia.org/wiki/Cochran-Armitage_test_for_trend>

[5] Page, Larry, "PageRank: Bringing Order to the Web", Stanford Digital Library Project, talk. August 18, 1997 (archived 2002)

[6] Morrison, Julie, Rainer Breitling, Desmond Higham, and David Gilbert, "Generank: Using Search Engine Technology for the Analysis of Microarray Experiments." *BMC Bioinformatics* 6, no. 1 (2005): 233

[7] Hill, W. G, "Estimation of Linkage Disequilibrium in Randomly Mating Populations." *Heredity (Edinb)* 33, no. 2 (Oct 1974): 229-39.

[8] Barrett, J. C., B. Fry, J. Maller, and M. J. Daly, "Haploview: Analysis and Visualization of Ld and Haplotype Maps." *Bioinformatics* 21, no. 2 (Jan 15 2005): 263-5.

[9] Danford, T., A. Rolfe, and D. Gifford, "GSE: A Comprehensive Database System for the Representation, Retrieval, and Analysis of Microarray Data." *Pac Symp Biocomput* (2008): 539-50.

[10] Castro, P., C. J. Creighton, M. Ozen, D. Berel, M. P. Mims, and M. Ittmann, "Genomic Profiling of Prostate Cancers from African American Men." *Neoplasia* 11, no. 3 (Mar 2009): 305-12.