Identification of Minimum Redundancy Tagging SNPs via Gibbs Sampling

Gaolin Zheng
Department of Math and Computer Science, North Carolina Central University, Durham, NC 27707, USA

Abstract—*Single nucleotide polymorphisms (SNPs) are genetic changes that can occur within a DNA sequence. Due to the high frequency of SNPs in the human genome, it is desirable to select a small set of SNPs (tagging SNPs) that can be used to represent the majority of SNPs. We propose a Gibbs sampling approach to find a small set of SNPs with minimum redundancy for tagging purposes. Pre-clustering is added in the basic Gibbs sampling procedure to avoid the disturbance caused by local optima. We also propose two general purpose correlation measures that are able to accommodate SNPs with three or more alleles. Our experimental results show that Gibbs sampling process converges faster and finds better optimum if pre-clustering is conducted before the sampling process. While our tagging process is not guided by any prediction algorithm, we are able to obtain comparable results as the SNP prediction guided algorithm SVM/STSA [1] while requiring much less time.*

Keywords: minimum redundancy, Chi-squared statistic, mutual information, single nucleotide polymorphism, Gibbs sampling

1 Introduction

Single nucleotide polymorphism (SNPs) are the most frequent variations in the human genome [2], and many SNPs show correlated genotypes because of their shared evolutionary history [3]. Many known polymorphic sites need not be genotyped when testing for genotype-phenotype associations because of this redundancy. There is considerable interest in finding an informative and minimal set of common polymorphisms (tagging SNPs) to detect genetic associations while controlling cost [1, 4-7]. Halldorsson et al. gave an in-depth review of these approaches [8].

Popular tagging SNP selection algorithms are typically based on block-based heuristics such as LD-Select [9], MultiPop-TagSelect [10]. The main drawback of block-based approaches is that the definition of blocks is not always straightforward and there is no consensus on how blocks must be formed [11]. Several researchers have focused on looking for tagging SNPs using block-free methods [1, 8, 11-13]. Most of these methods are based on some greedy deterministic searching procedures that are susceptible to local optimum. Furthermore, most of these methods are using the r^2 similarity/correlation measure between two SNPs. This measure is not able to handle three or more alleles. SNPs with three or more alleles are usually ignored for processing conveniences. To accommodate more alleles, we propose two correlation measures that are more general purpose for handling nominal data. The first one is mutual information and the second one is the Chi-squared statistic.

Finding a set of k tagging SNPs out of a total set of n SNPs requires evaluating $\binom{n}{k}$ different combinations. It is computationally infeasible to exhaustively search the optimal solution when n is usually large. In this study, we describe a global search heuristic based on a randomized procedure (Gibbs sampling) that aims to find a set of tagging SNPs with minimum redundancy. Although the stochastic nature of Gibbs sampling is presumed to prevent it from becoming completely trapped in local optima, it still requires a better initial value due to strong disturbance from the local optima. We propose a pre-clustering approach to obtain a better initial SNP set. The effect of pre-clustering will be investigated.

The paper is organized as follows. In Section 2, we explain our Gibbs sampling approach to obtain the minimum redundancy SNP set. The experiments and results will be discussed in Sections 3 & 4. We conclude our paper in Section 5.

2 Methods

2.1 Redundancy Measures

Consider two biallelic loci, locus 1 with alleles a and A, locus 2 with alleles b and B. Suppose the frequencies for alleles A and a are P_A and $1 - P_A$, the frequencies for alleles B and b are P_B and $1 - P_B$, and the the frequency of genotypes having allele A at locus 1 and allele B at locus 2 is P_{AB}. The commonly used *linkage disequilibrium measure* r^2 [14] is defined as

$$r^2 = \frac{(P_{AB} - P_A P_B)^2}{P_A(1 - P_A)P_B(1 - P_B)} \qquad (1)$$

The mutual dependency of two random variables can also be used as a redundancy measure. Here redundancy and

correlation are used interchangeably. The *mutual information* between SNP *X* and SNP *Y* is defined as

$$MI(X,Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) log \frac{p(x,y)}{p(x)p(y)} \qquad (2)$$

where both *X* and *Y* are discrete variables, *p(x,y)* is the joint probability and *p(x)* and *p(y)* are marginal probabilities.

Chi-squared test of independence is adopted here to measure the correlation between two SNPs. For SNP *X* and SNP *Y*, we first obtain a contingency table between the two SNPs. The *Chi-squared statistic* is defined as

$$\chi_s^2(X,Y) = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} \qquad (3)$$

where *r* is the number of alleles for SNP *X*, and *c* is the number of alleles for SNP *Y*, O_{ij} is the observed joint frequency for i^{th} allele of SNP *X* and j^{th} allele of SNP *Y*, and E_{ij} is the expected frequency which is given by

$$E_{ij} = \frac{\sum_{k=1}^{c} O_{ik} \sum_{k=1}^{r} O_{kj}}{N} \qquad (4)$$

where N is the total number of samples. A higher value of χ_s^2 indicates a stronger association between the two SNPs.

For a set *S* consisting of *k* SNPs, the total pair-wise mutual information is defined as

$$MISUM(S) = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} MI(SNP_i, SNP_j) \qquad (5)$$

The total pair-wise Chi-squared statistics is defined as

$$CHISUM(S) = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \chi_s^2(SNP_i, SNP_j) \qquad (6)$$

The total pair-wise r^2 measure is give by

$$R2SUM(S) = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} r^2(SNP_i, SNP_j) \qquad (7)$$

2.2 Clustering of SNP Data

Due to the nominal nature of SNP data, the commonly used K-means clustering and its many variants are not suitable. In this study, we first obtain a similarity matrix using a similarity measure that is applicable for nominal data such as Chi-square statistic or mutual information. The distance matrix is then obtained by subtracting each entry from the maximum of all the values. We then apply the agglomerative clustering procedure with complete linkage to obtain the desired number of clusters.

2.3 Gibbs Sampling

Gibbs sampling is a special case of the Metropolis–Hastings algorithm. It is a stochastic global search heuristic for optimization problems. However, it still requires a better starting set to avoid being trapped in local optima. A pre-clustering is proposed to avoid the disturbance from local optima. To achieve this, we first cluster the SNPs into K groups, and randomly select an SNP from each group to form the initial SNP set. We then follow a Gibbs sampling procedure to find a set of K SNPs that minimize a goal function. The goal function can be one of the functions defined in equation 5-7. The detail of our approach is summarized in Figure 1.

```
Input: S is the total set of N SNPs, ε is a predefined threshold value
Output: C is the set of K chosen SNPs
minRedundancySNP(S, C, ε)
        Cluster the set S into K groups via customized hierarchical clustering
        Form set G of K members where G_i is the i^th cluster
        Form initial set C by randomly pick one SNP from each of the K clusters
        while( a predefined maximum iteration is not reached)
                Randomly pick a number n from 1 to K
                Find a SNP x in G_n that minimizes MISUM/CHISUM/R2SUM
                Replace C_n with x
                Return C if the improvement of MISUM/CHISUM/R2SUM is less than ε
        end
        Return C
```

Figure 1. The pseudocode for finding the minimum redundancy SNP set via Gibbs sampling with pre-clustering.

2.4 Prediction of Non-tagging SNPs with Tagging SNPs

Once the tagging SNP set is found, they can be used to predict the genotype values of the non-tagging SNPs. Many machine learning and statistical models can be used for this goal, including logistic regression [15], neural networks, support vector machines (SVM) [16], and random forest [17]. In this study, we conduct our experiments using logistic regression and SVM. We choose a *K*-fold cross validation to evaluate the effectiveness of our method. Our *K*-fold cross validation procedure is similar to the leave-one-out cross validation procedure for SNP prediction described in [1] where *K* is equal to the number of observations in the original sample.

3 Experimental Data

The following datasets are used to validated our method.

IBD 5q31: This data set is from an inflammatory bowel disease study of father-mother-child trios [18]. The original data set contained 103 SNPs in 387 subjects. Using the PHASE 2.0.2 software to derive haplotypes resulted in 103 non-singletons from 774 phased chromosomes.

TRPM8: The phased haplotype data was downloaded from Hapmap Data release 24. It contains 101 SNPs from 119 phased chromosomes.

4 Results and Discussion

4.1 Effect of Pre-clustering on the Convergence of the Gibbs Sampling Process

In order to test how fast the Gibbs sampling process converges, we obtained the convergence curve using all three measures introduced in Section 2.1 (i.e., the linkage disequilibrium measure, mutual information, and Chi-squared statistic). Figure 2 shows the convergence process while attempting to find 10 tagging SNPs.

In each case, the Gibbs sampling process converged within 100 iterations regardless of whether or not pre-clustering was applied. However, the resulting set of SNPs had smaller redundancy measures when pre-clustering was used. Without pre-clustering, there is still some disturbance from local optima that affect the global minimum search process through Gibbs sampling.

4.2 Tagging Results

We conducted our experiments on the three distance measures to find tagging SNPs using the randomized algorithm mentioned above. The tagging results for IBD data set using r^2, Chi-squared statistic and mutual information are shown in Table I, II, III respectively. The tagging results for TRPM8 data set using r^2, Chi-squared statistic and mutual information are shown in Table IV, V, VI respectively.

For IBD data, pre-clustering is able to improve prediction performance. With pre-clustering, our 10-fold cross validation results are comparable with published leave-one-out cross validation results obtained by He et al. [1] and better than the results obtained by FSFS [11] (Table II, III). SVM and logistic regression show similar performance. Although our method does not present significant advantages over He's method [1], our method is simpler and does not rely on specific machine learning model to guide the selection process which is susceptible to over-fitting. In addition, the prediction based selection method SVM/STSA [1] requires calling SVM model during each stepwise selection process. This can be expensive due to the overhead of the prediction algorithm.

Among the three distance measures, both Chi-squared statistic and mutual information performed better than r^2 measure. This proves both of them can be used to study SNP association, and they are particularly useful for genotype data that sometimes involve more than two alleles.

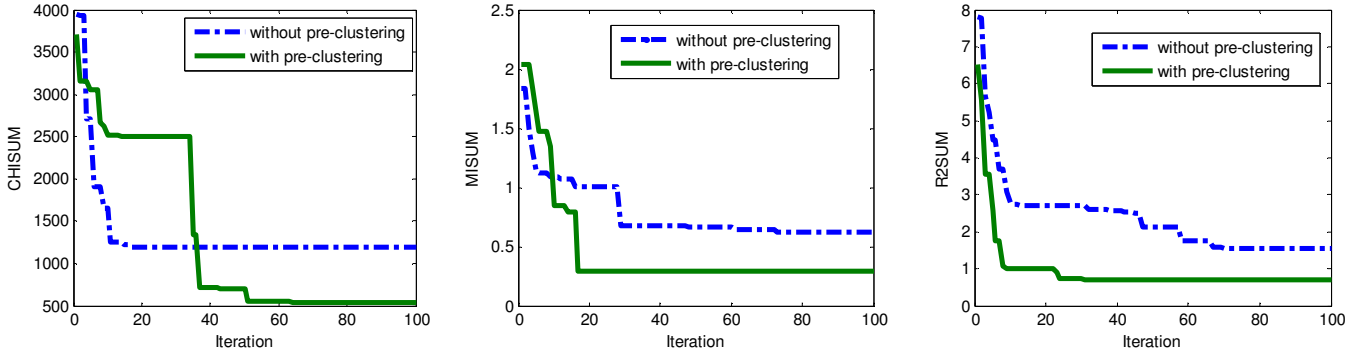

Figure 2. Convergence curve for the Gibbs sampling process based on three redundancy measures. (a) minimization process of CHISUM (b) minimization process of MISUM (c) minimization process of R2SUM.

For TRPM8, our ten-fold cross validation prediction performance is better than SVM/STSA when mutual information is used as correlation measure (Table V). Similar performances are observed between r^2 measure and Chi-squared statistic. The pre-clustering does not improve the prediction performance significantly. This indicates that the disturbance from local optima in this data set is not as strong as in IBD data set.

TABLE I. TEN-FOLD CROSS VALIDATION EXPERIMENT RESULTS ON IBD DATA SET (CORRELATION IS MEASURED WITH r^2 MEASURE, K IS THE NUMBER OF TAGGING SNPS).

K	With Pre-clustering		Without Pre-clustering	
	SVM	Logistic Regression	SVM	Logistic Regression
3	81.1%	80.7%	76.6%	76.6%
5	81.5%	81.2%	78.8%	77.4%
10	93.5%	91.7%	77.4%	77.4%
20	98.2%	97.8%	85.7%	86.0%
30	98.5%	98.3%	93.5%	93.5%

TABLE II. TEN-FOLD CROSS VALIDATION EXPERIMENT RESULTS ON IBD DATA SET (CORRELATION IS MEASURED WITH CHI-SQUARED STATISTIC, K IS THE NUMBER OF TAGGING SNPS).

K	With Pre-clustering		Without Pre-clustering	
	SVM	Logistic Regression	SVM	Logistic Regression
3	85.6%	85.5%	79.3%	79.3%
5	86.0%	85.1%	81.5%	81.1%
10	95.0%	93.3%	94.4%	93.5%
20	98.2%	97.9%	98.1%	97.5%
30	98.5%	98.5%	97.8%	97.4%

TABLE III. TEN-FOLD CROSS VALIDATION EXPERIMENT RESULTS ON IBD DATA SET (CORRELATION IS MEASURED WITH MUTUAL INFORMATION, K IS THE NUMBER OF TAGGING SNPS).

K	With Pre-clustering		Without Pre-clustering	
	SVM	Logistic Regression	SVM	Logistic Regression
3	86.6%	86.5%	80.0%	79.9%
5	87.3%	86.1%	79.8%	79.8%
10	96.0%	95.0%	77.4%	77.3%
20	98.2%	97.8%	89.8%	88.4%
30	98.5%	98.3%	97.3%	96.4%

TABLE IV. TEN-FOLD CROSS VALIDATION EXPERIMENT RESULTS ON TRPM8 DATA SET (CORRELATION IS MEASURED WITH r^2 MEASURE, K IS THE NUMBER OF TAGGING SNPS).

K	With Pre-clustering		Without Pre-clustering	
	SVM	Logistic Regression	SVM	Logistic Regression
3	89.1%	87.9%	87.6%	81.3%
5	88.9%	86.2%	87.5%	92.3%
10	91.7%	91.3%	91.3%	92.4%
20	99.5%	99.7%	99.2%	99.2%
30	99.7%	99.7%	99.3%	99.7%

TABLE V. TEN-FOLD CROSS VALIDATION EXPERIMENT RESULTS ON TRPM8 DATA SET (CORRELATION IS MEASURED WITH CHI-SQUARED STATISTIC, K IS THE NUMBER OF TAGGING SNPS).

K	With Pre-clustering		Without Pre-clustering	
	SVM	Logistic Regression	SVM	Logistic Regression
3	89.1%	87.9%	87.6%	81.3%
5	88.9%	86.2%	87.5%	92.3%
10	91.7%	91.3%	91.3%	92.4%
20	99.5%	99.6%	99.2%	98.6%
30	99.7%	99.7%	99.3%	99.7%

TABLE VI. TEN-FOLD CROSS VALIDATION EXPERIMENT RESULTS ON TRPM8 DATA SET (CORRELATION IS MEASURED WITH MUTUAL INFORMATION, K IS THE NUMBER OF TAGGING SNPS).

K	With Pre-clustering		Without Pre-clustering	
	SVM	Logistic Regression	SVM	Logistic Regression
3	96.7%	90.1%	89.3%	81.8%
5	97.3%	92.1%	96.3%	92.3%
10	97.3%	92.3%	96.0%	92.2%
20	99.7%	99.7%	97.4%	98.5%
30	99.7%	99.7%	98.1%	99.7%

4.3 Running Time Results

The running time required to select different number of tagging SNPs using our Gibbs sampling procedure is shown in Table VII. Our Gibbs sampling code is implemented with R statistical programming language.

The running time increases as the number of tagging SNPs increases. The running time results are similar between the Chi-squared statistic and mutual information. The program often ran a little faster with r^2 as correlation measure. Comparing with prediction guided SNP selection SVM/STSA [1] which takes up to 1 day to find 10 tagging SNPs for IBD data set, and 23 hours to find 10 tagging SNPs for TRPM8 data. It even took several hours to find 1 tagging SNPs[1], our Gibbs sampling procedure runs much faster and usually completes within 5 minutes for up to 30 tagging SNPs.

TABLE VII. RUNNING TIME REQUIRED (SECONDS) TO SELECT TAGGING SNPS USING DIFFERENT CORRELATION MEASURES (K IS THE NUMBER OF TAGGING SNPS, ALL EXPERIMENTS ARE PERFORMED ON A COMPUTER WITH AMD ATHLON II X4 620, 2.61 GHZ PROCESSOR AND 2 GB OF RAM)

Data set	IBD			TRPM8		
K	Correlation measure			Correlation measure		
	r^2	χ^2	MI	r^2	χ^2	MI
3	11.57	11.75	11.28	9.11	10.52	9.25
5	20.68	19.83	23.11	16.62	10.86	21.32
10	53.44	63.20	53.00	25.56	27.30	34.15
20	94.22	136.23	120.62	80.38	107.73	91.69
30	196.94	195.83	191.46	121.27	157.03	157.81

5 Conclusions

We investigated a block-free stochastic global search heuristic to find a set of minimum redundancy tagging SNPs. It is a randomized search technique based on Gibbs sampling. We modified the basic Gibbs sampling procedure by adding a pre-clustering step to find a better starting set. In order to properly cluster the SNP data, we applied hierarchical clustering with a distance measure that is applicable for nominal data. The Gibbs sampling process typically converges faster and reaches lower minimum if a pre-clustering is used. Pre-clustering improves the tagging prediction accuracy if there is a disturbance from local optima. If there is little disturbance from local optima, pre-clustering at least does no harm.

Although our tagging process is driven by a simple objective function that aims to minimize redundancy among a set of SNPs instead of being driven by a prediction method such as SVM, we are able to obtain comparable prediction results while running much faster than prediction based SNP selection method [1].

We also proposed two correlation measures to study SNP association. They proved to be as effective as the commonly used r^2 measure. These two measures can be useful for genetic features (e.g. genotypes) that could have more than two alleles.

6 Acknowledgments

This work was supported by the National Institutes of Health [5T36GM008789-08].

7 References

[1] J. He and A. Zelikovsky, "Informative SNP Selection Methods Based on SNP Prediction," *Nanobioscience,* vol. 6, pp. 60-67, July 18, 2006 2006.

[2] L. Kruglyak and D. Nickerson, "Variation is the spice of life.," *Nat Genet,* vol. 27, pp. 234-236, 2001.

[3] D. Reich*, et al.*, "Linkage disequilibrium in the human genome.," *Nature,* vol. 411, pp. 199-204, 2001.

[4] H. Ackerman*, et al.*, "Haplotypic analysis of the TNF locus by association efficiency and entropy.," *Genome Biology,* vol. 4, p. R24, 2003.

[5] X. Ke and L. R. Cardon, "Efficient selective screening of haplotype tag SNPs," *Bioinformatics,* vol. 19, pp. 287-288, January 22, 2003 2003.

[6] Z. Meng*, et al.*, "Selection of genetic markers for association analyses, using linkage disequilibrium and haplotypes.," *The American Society of Human Genetics,* vol. 73, pp. 115-130, June 5 2003.

[7] K. Zhang*, et al.*, "A dynamic programming algorithm for haplotype block partitioning.," *Proc Natl Acad Sci,* vol. 99, pp. 7335-7339, 2002.

[8] B. Halldorsson*, et al.*, "Optimal selection of SNP markers for disease association studies.," *Hum Hered,* vol. 58, pp. 190-202, 2004.

[9] C. Carlson*, et al.*, "Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium.," *Am J Hum Genet. ,* vol. 74, pp. 106-120, 2004.

[10] B. Howie*, et al.*, "Efficient selection of tagging single-nucleotide polymorphisms in multiple populations," *Human Genetics,* vol. 120, pp. 58-68, 2006.

[11] T. M. Phuong*, et al.*, "Choosing SNPs Using Feature Selection," *Proceedings IEEE Computational Systems Bioinformatics Conference,* pp. 301-309, 2005.

[12] P. Sebastian*, et al.*, "Minimal Haplotype tagging," *Proceedings of the National Academy of Sciences,* vol. 100, pp. 9900-9905, 2003.

[13] J. He and A. Zelikovsky, "MLR-tagging: informative SNP selection for unphased genotypes based on multiple linear regression," *Bioinformatics,* vol. 22, pp. 2558-2561, October 15, 2006 2006.

[14] W. G. Hill and A. Robertson, "Linkage disequilibrium in finite populations," *TAG Theoretical and Applied Genetics,* vol. 38, pp. 226-231, 1968.

[15] A. Agresti, *Categorical Data Analysis.* New York: Wiley-Interscience, 2002.

[16] C. Cortes and V. Vapnik, "Support Vector Network," *Machine Learning,* vol. 20, pp. 273-297, 1995.

[17] L. Breiman, "Random Forests.," *Machine Learning,* vol. 45, pp. 5-32, 2001.

[18] M. J. Daly*, et al.*, "High-resolution haplotype structure in the human genome," *Nature Genetics,* vol. 29, pp. 229-232, 2001.