

Genome-wide Tagging SNPs with Entropy Based Monte Carlo Method

Zhenqiu Liu¹, Shili Lin^{2,*}, Ming Tan¹

¹Division of Biostatistics, University of Maryland Greenebaum Cancer Center, 22 South Greene Street, Baltimore, MD 21201, USA.

²Department of Statistics, The Ohio State University, 1958 Neil Avenue, Columbus, OH 43210, USA

*Address for correspondence:

Shili Lin, PhD

Department of Statistics

The Ohio State University

1958 Neil Avenue

Columbus, OH 43210-1247

Tel: (614) 292-7404

Fax: (614) 292-2096

Email: shili@stat.ohio-state.edu

Abstract

The number of common single nucleotide polymorphisms (SNPs) in the human genome is estimated to be around 3-6 millions. It is highly anticipated that the study of SNPs will help provide a means for elucidating the genetic component of complex diseases and variable drug responses. High throughput technologies such as oligonucleotide arrays have produced enormous amount of SNP data, which creates great challenges in genome-wide disease linkage and association studies. In this paper, we present an adaptation of the cross entropy (CE) method and propose an iterative CE Monte Carlo (CEMC) algorithm for tagging SNP selection. This differs from most of SNP selection algorithms in the literature in that our method is independent of the notion of haplotype block. Thus, the method is applicable to whole genome SNP selection without prior knowledge of block boundaries. We applied this block-free algorithm to three large datasets (two simulated and one real) that are in the order of thousands of SNPs. The successful applications to these large scale datasets demonstrate that CEMC is computationally feasible for whole genome SNP selection. Furthermore, the results show that CEMC is significantly better than random selection, and it also outperformed another block-free selection algorithm for the dataset considered.

1 Introduction

Single-nucleotide polymorphism (SNP) tagging is widely used as a way of saving genotyping costs in association studies. It is even more important with the availability of HapMap. In an association study of a candidate region, it is usually too costly to genotype all SNPs within the region, as the number of SNPs can be very large given their abundant nature. On the other hand, genotyping only an appropriately selected subset may not lead to much, if any, reduction in information. This is due to limited diversity among densely spaced SNP markers. Thus, the challenge is to select a minimal subset that retains most of the information provided by the full set.

Tagging SNPs is usually associated with the concept of haplotype block, despite significant theoretical and empirical evidence which shows that conserved substructure may be lost when the data is being fitted to a block structure (Wall and Pritchard 2003). Nonetheless, one strong argument for assuming a block structure has been algorithmic convenience. Haplotype block is defined as discrete regions of low diversity whose boundaries are conserved across distinct haplotypes. Most algorithms in the literature are haplotype block tagging, that is, grouping SNPs into segments of low haplotype diversity and then tagging a subset of the SNPs within that block (Avi-Itzhak et al. 2003; Byng et al. 2003; Carlson et al. 2004; Ke and Cardon 2003; Sebastiani et al. 2003; Zhang et al. 2004). The main concern of these papers is the problem of defining a quality measure, that is, how well a set of tagging SNPs captures the variance observed. It generally assumes that the region and the number of SNPs dealt with are not very large, which imply tagging within each block. Classical algorithms such as exhaustive search, forward and backward selections have been applied extensively for SNP selections in the literature.

There are only a handful of block free methods available currently. Halldorsson et al. (2004) proposed a block free tagging SNP method based on weak correlations that occur

across neighboring blocks. This method makes use of the neighborhoods of potentially predictive SNPs, which is not totally block free but more flexible than the rigid notation of haplotype blocks. Rinaldo et al. (2005) developed a method that select tagging SNPs with hierarchy clustering instead of blocks. This method chooses SNPs that are the closest to the center of each cluster. It does not aim to optimize any criterion explicitly.

In this paper, we propose an entropy based block free method for SNP tagging that is applicable to large data sets. The proposed method makes use of a global optimization technique, cross entropy Monte Carlo (CEMC), and searches the tagging SNPs from the full set that optimizes a criterion. Although the entropy based multilocus LD measure ER (Liu and Lin 2005) was used as the criterion to be optimized in the applications in this paper, the proposed method is general and is amenable to any other optimization criterion. We evaluated our algorithm with both simulated and real data sets and compared our method with that of Halldorsson et al. (2004). Our successful applications demonstrate that CEMC is computationally feasible for whole genome SNP selection. Furthermore, the results show that CEMC is significantly better than random selection, and it also outperformed the other block-free selection algorithm in our comparison for the real dataset.

2 Methods

Information theory provides a natural way to quantify relevant information. Liu and Lin (2005) proposed a multilocus linkage equilibrium measure based on Kullback-Leibler (K-L) distance for two probability distributions. Assuming that there are n SNPs with m observed haplotypes. Let \mathbf{x}_i be the i th haplotype, with x_{ij} being the allele at locus j , then the LD measure is defined as

$$E = \sum_{i=1}^m p(\mathbf{x}_i) \log_2 \frac{p(\mathbf{x}_i)}{\prod_{j=1}^n p_j(x_{ij})},$$

where the p 's denote the appropriate observed frequencies of the haplotypes or the alleles. Because of the properties of the K-L distance, this LD measure is nonnegative and is zero if and only if the variables are independent. This measure is bounded. The bound can be found in terms of entropies of component variables (SNPs) \mathbf{x}_j :

$$E \leq \sum_{j=1}^n H(\mathbf{x}_j) - \max_j H(\mathbf{x}_j) = E_{\max}.$$

Consequently, Liu and Lin (2005) defined the normalized LD measure with entropy ratio (ER):

$$ER = \frac{E}{E_{\max}} = \frac{\sum_{i=1}^m p(\mathbf{x}_i) \log_2 \frac{p(\mathbf{x}_i)}{\prod_{j=1}^n p_j(x_{ij})}}{\sum_{j=1}^n H(\mathbf{x}_j) - \max_j H(\mathbf{x}_j)}. \quad (1)$$

They showed that ER can deal with a large number of SNPs and genotype data very efficiently.

Based on ER, they proposed a weighted criterion

$$\omega(S) = (1 - \lambda) \frac{H(X_S)}{H(X)} + \lambda(1 - ER(X_S)) \quad (2)$$

for selecting a SNP set, where S is the index set of the selected SNPs, i.e., $S \subset \{1, \dots, n\}$, $H(X_S)$ and $H(X)$ are the joint entropy of the selected and the total SNPs, respectively, and $\lambda \in [0, 1]$ is a free parameter to be determined. The criterion ω allows us to choose a set of tagging SNPs with large joint information yet small LD simultaneously. Note that λ can affect the number of SNPs chosen: a smaller λ usually leads to a bigger tagging SNPs set than a larger λ , since it puts less penalty on choosing more SNPs coupled with the fact that $H(X_S)$ is a monotone increasing function. We suggest choosing a λ in the range of 0.3-0.7, following Liu and Lin (2005).

Cross Entropy Monte Carlo Method

The objective of SNP tagging is to choose a smallest subset of SNPs that maximizes the ω criterion. Mathematically the problem can be defined as finding S^* such that

$$S^* = \arg \max_S \{\omega(S), S \subset \{1, \dots, n\}\}. \quad (3)$$

This problem is combinatorial in nature and an exhaustive search requires searching through all subsets of indexes of the SNPs. This is not tractable even for a moderate number of SNPs. To select the tagging SNPs from a large full set through optimizing ω , we propose to adopt the cross entropy (CE) Monte Carlo (MC) approach. The original CE algorithm was proposed by Rubinstein (1997) for estimating probabilities of rare events that involves variance minimization, which has now been expanded to solving difficult combinatorial problems (Rubinstein and Kroese 2004) as well. Thus, it is suitable for our purpose.

Suppose there are n SNPs in the full set. Let $\mathcal{Z} = \{(z_1, z_2, \dots, z_n) | z_i \in \{0, 1\}\}$ be the collection of all possible binary vectors of length n . Each $\mathbf{z} = (z_1, z_2, \dots, z_n) \in \mathcal{Z}$ gives rise to the index set of a subset of the n SNPs through a 1-1 mapping function $S = S(\mathbf{z}) = \{j | z_j = 1, j = 1, \dots, n\}$. Thus the goal of finding S^* that leads to the maximum value y^* of the ω criterion in (2) is equivalent to finding \mathbf{z}^* that maximizes Φ as follows:

$$y^* = \omega(S^*) = \omega(S(\mathbf{z}^*)) = \Phi(\mathbf{z}^*) = \max_{\mathbf{z} \in \mathcal{Z}} \Phi(\mathbf{z}).$$

Instead of finding \mathbf{z}^* by solving the combinatorial problem directly, we used the CE method that addresses the problem through an iterative procedure. Suppose that $\{f_{\mathbf{v}} = f(\mathbf{z}; \mathbf{v})\}$ is a known parametric family of probability mass functions (pmf) defined on \mathcal{Z} that is indexed by an unknown parameter vector \mathbf{v} that will be updated iteratively. Iteration by iteration, $f(\mathbf{z}; \mathbf{v})$ will place more and more of its probability mass on the \mathbf{z} 's that are in the “neighborhood” of \mathbf{z}^* . By a “neighborhood” of \mathbf{z}^* , we mean those \mathbf{z} 's whose corresponding

y ($= \Phi(\mathbf{z})$) values are close to the maximum y^* . Each iteration can be considered as composed of two stages: (stage 1) generation of a random sample of \mathbf{z} 's from $f(\mathbf{z}; \mathbf{v})$ based on the current parameter vector \mathbf{v} , and (stage 2) updating of the parameter vector \mathbf{v} based on the \mathbf{z} 's generated in order to produce a sample in the next iteration that will be mostly from even a “tighter” neighborhood.

The remaining question is how to update \mathbf{v} to meet the above described objectives. This is accomplished by considering the estimation of the following quantity

$$P_{\mathbf{v}}(y) = P_{\mathbf{v}}(\Phi(\mathbf{z}) \geq y) = \sum_{\mathbf{z}} I(\Phi(\mathbf{z}) \geq y) f(\mathbf{z}; \mathbf{v}) = E_{\mathbf{v}} I(\Phi(\mathbf{z}) \geq y), \quad (4)$$

where I is an indication function that is equal to 1 if $\Phi(\mathbf{z}) \geq y$ and 0 otherwise. A Monte Carlo solution is possible by obtaining a sample $\mathbf{z}_1, \dots, \mathbf{z}_N$ from $f(\mathbf{z}; \mathbf{v})$ and estimate it by

$$\hat{P}_{\mathbf{z}}(y) = \frac{1}{N} \sum_{i=1}^N I(\Phi(\mathbf{z}_i) \geq y).$$

But this estimate may not be efficient (i.e., with a large Monte Carlo error) unless $f(\mathbf{z}; \mathbf{v})$ puts most of its mass on the \mathbf{z} 's that satisfy $\Phi(\mathbf{z}) \geq y$. One way to reduce Monte Carlo variance is by importance sampling, in which a sample is drawn from a distribution $g(\mathbf{z})$ different from $f(\mathbf{z}; \mathbf{v})$ but defined on the same sampling space \mathcal{Z} . The “ideal” (but unavailable due to the unknown normalizing constant) sampling distribution would be

$$g^*(\mathbf{z}) = \frac{I(\Phi(\mathbf{z}) \geq y) f(\mathbf{z}; \mathbf{v})}{P_{\mathbf{v}}(y)},$$

which would lead to an estimator with zero variance. This gives us an idea for specifying $f(\mathbf{z}; \mathbf{v})$ (i.e., updating the parameter vector \mathbf{v}) such that it is as close to g^* as possible. Specifically, let \mathbf{v} be the current estimate of the parameter vector. We choose the next parameter update \mathbf{v}' to minimize the Kullback-Leibler distance (also known as cross entropy) between the distributions $f_{\mathbf{v}'} = f(\mathbf{z}; \mathbf{v}')$ and g^* as given in the following:

$$CE(g^*, f_{\mathbf{v}'}) = \sum_{\mathbf{z}} g^*(\mathbf{z}) \log g^*(\mathbf{z}) - \frac{1}{P_{\mathbf{v}}(y)} \sum_{\mathbf{z}} [I(\Phi(\mathbf{z}) \geq y) \log f(\mathbf{z}; \mathbf{v}')] f(\mathbf{z}; \mathbf{v}).$$

The first term, as well as $P_{\mathbf{v}}(y)$, in the above formula only involve \mathbf{v} but not \mathbf{v}' , therefore, minimizing $CE(g^*, f_{\mathbf{v}'})$ is equivalent to maximizing

$$\sum_{\mathbf{z}} [I(\Phi(\mathbf{z}) \geq y) \log f(\mathbf{z}; \mathbf{v}')] f(\mathbf{z}; \mathbf{v}) = E_{\mathbf{v}} [I(\Phi(\mathbf{z}) \geq y) \log f(\mathbf{z}; \mathbf{v}')],$$

which can be approximated by a Monte Carlo estimator using a sample, $\mathbf{z}_1, \dots, \mathbf{z}_N$, drawn from $f(\mathbf{z}; \mathbf{v})$ with the current estimate of the parameter vector \mathbf{v} . Thus

$$\mathbf{v}_{\text{new}} = \arg \max_{\mathbf{v}'} \frac{1}{N} \sum_{i=1}^N [I(\Phi(\mathbf{z}_i) \geq y) \log f(\mathbf{z}_i; \mathbf{v}')] \quad (5)$$

will be used as the next estimate of the parameter vector \mathbf{v} . In practice, the threshold y will also be updated iteratively, which will be configured into the updating scheme of \mathbf{v} . This exercise will lead to the construction of a sequence, y_0, y_1, \dots , that will converge to a value (y_{∞}) close to y^* . Furthermore, $\mathbf{v}_0, \mathbf{v}_1, \dots$ will converge to \mathbf{v}_{∞} , with the corresponding $f(\mathbf{z}, \mathbf{v}_{\infty})$ placing most of its probability mass on the \mathbf{z} 's that satisfy $\Phi(\mathbf{z}) \geq y_{\infty}$.

The CEMC Algorithm for SNP Selection

First, we need to specify the pmf $f(\mathbf{z}; \mathbf{v})$ up to the parameter vector \mathbf{v} . We assume that each z_j is independently distributed as Bernoulli (p_j), and thus the parameter vector is $\mathbf{v} = \mathbf{p} = (p_1, \dots, p_n)$. We define the pmf to be

$$f(\mathbf{z}; \mathbf{p}) = \prod_{j=1}^n p_j^{z_j} (1 - p_j)^{1 - z_j}, \quad z \in \mathcal{Z}.$$

To derive the specific update procedure for \mathbf{p} , we need to solve (5) by setting the first derivatives with respect to p'_j , $j = 1, \dots, n$, equal to zero:

$$\frac{\partial}{\partial p'_j} \left\{ \frac{1}{N} \sum_{i=1}^N I(\Phi(\mathbf{z}_i) \geq y) \log f(\mathbf{z}_i; \mathbf{p}') \right\} = \frac{1}{(1 - p'_j)p'_j} \sum_{i=1}^N I(\Phi(\mathbf{z}_i) \geq y) (z_{ij} - p'_j) = 0,$$

where $\mathbf{z}_i = (z_{i1}, \dots, z_{in})$, $i = 1, \dots, N$, is a sample from $f(\mathbf{z}, \mathbf{p})$ with the current estimate of \mathbf{p} . Solving the above equations gives the following specific update schemes:

$$p'_j = \frac{\sum_{i=1}^N I(\Phi(\mathbf{z}_i) \geq y) z_{ij}}{\sum_{i=1}^N I(\Phi(\mathbf{z}_i) \geq y)}, \quad j = 1, \dots, n. \quad (6)$$

The Algorithm:

1. Set \mathbf{p}^0 with each $p_j^0 \in (0, 1)$. For instance, $p_j^0 = 0.5$ indicates that each SNP can be selected with 50% chances. Set $t = 0$.
2. Draw a sample $\mathbf{z}_i = (z_{i1}, \dots, z_{in})$, $i = 1, \dots, N$, of Bernoulli vectors with success probability \mathbf{p}^t . Find the tagging index set $S_i = \{j | z_{ij} = 1, j = 1, \dots, n\}$ and calculate $\Phi(\mathbf{z}_i) = \omega(S_i)$ for all i 's and sort them in ascending order: $\Phi_{(1)} \leq \dots \leq \Phi_{(N)}$. Let $[(1 - \rho)N]$ be the integer part of $(1 - \rho)N$, then we have the sample $(1 - \rho)$ -quantile of the performances: $y^t = \Phi_{([(1 - \rho)N])}$, where $\rho < 1$ is a free parameter needed to be specified.
3. Use the same sample \mathbf{z}_i , $i = 1, \dots, N$ to update the parameter vector $\mathbf{p}^{t+1} = (p_1^{t+1}, \dots, p_n^{t+1})$ via

$$p_j^{t+1} = \frac{\sum_{i=1}^N I(\Phi(\mathbf{z}_i) \geq y^t) z_{ij}}{\sum_{i=1}^N I(\Phi(\mathbf{z}_i) \geq y^t)}, \quad j = 1, \dots, n.$$

4. if $\|\mathbf{p}^{t+1} - \mathbf{p}^t\| < \varepsilon_1$ and $|y^{t+1} - y^t| < \varepsilon_2$, then go to Step 5; otherwise set $t = t + 1$ and go back to step 2. Note that $\|\cdot\|$ denotes a norm such as the sum of squared component distances.
5. Output $y = \Psi_{(N)}$ and the corresponding selected SNPs set S , which will be taken as the estimate of our tagging SNP set S^* .

One can prove that the y 's and the \mathbf{p} 's sequences obtained from our algorithm converge following Margolin (2005). The choices for the Monte Carlo sample size N and the parameter ρ depends on the number of SNPs. These two parameters control the efficiency of the algorithm as well as the ability of the algorithm to escape from a local maximum, and thus should be chosen carefully. Based on the experience gained from several analyses, we would suggest setting $N = cn$, where n is the number of SNPs and c is the constant of at least 5.

For ρ , setting it equal to 0 will trap the algorithm in its local maximum, but setting it too large will lead to computational inefficiency. Our experience indicates that ρ in the range of 0.05-0.1 works adequately.

3 Applications and Results

We describe applications of the CEMC algorithm to four sets of data to assess its performance and to compare with a few other existing methods for SNP selection. The first (51-SNP dataset) was a small dataset used to mainly assess the quality of the algorithm, as near-optimal results can be obtained from an existing method. The second real data set has 4120 SNPs on chromosome 22. This, together with two large sets of simulated data, were for the purpose of demonstrating the utility of the CEMC algorithm for whole genome SNP selection. In all applications and algorithms, the performances were assessed using the haplotype r^2 (RSQ) criterion discussed in Weale et al. (2003).

A 51-SNP Dataset

To assess the performance of the proposed tagging SNP algorithm, we applied the CEMC algorithm to Clayton's 51-SNP genotype data (<http://www-gene.cimr.cam.ac.uk/clayton>). The haplotype frequencies were estimated with the PHASE software (Stephens and Donnelly 2003). There were a total of 49 haplotypes, with four having frequencies above 5% (0.295, 0.193, 0.07, and 0.05). In addition to CEMC, we also analyzed this dataset using two forward selection algorithms (FSA), one based on the ω criterion in (2) and the other based on the RSQ criterion. Although FSA with RSQ (FSA(RSQ)) is not an exhaustive search (which is infeasible for datasets with more than 30 SNPs using PCs at their current computational power (Liu and Lin 2005)), the resulting tagSNP set should be close to the optimum evaluated

using the RSQ criterion. Therefore, the results from FSA(RSQ) was treated as the “gold standard” for comparisons. For CEMC and FSA(ω), we used $\lambda = 0.4$. Furthermore, for CEMC, we set $p_j^0 = 0.5, j = 1, \dots, n, N = 1000, \rho = 0.1$, and $\varepsilon_1 = \varepsilon_2 = 10^{-6}$.

We first performed the CEMC algorithm on a PC with a 2.8 GHz processor, which took 16 iterations to converge and selected 14 SNPs. We then performed both FSA(RSQ) and FSA(ω) for selecting 14 SNPs. The computational times for all three algorithms were short and comparable (0.2 sec for CEMC versus 0.15 sec for the FSAs). The three tagged SNP sets selected from the three algorithms were all evaluated using the RSQ criterion, and the results are shown in Table 1. We see that CEMC is very close to the gold standard set by FSA(RSQ), while FSA(ω) is slightly lagging behind, according to the RSQ criterion. These results demonstrate that CEMC is indeed a viable alternative for SNP selection, as it is capable of selecting a tagging set that is composed of only 27% of the full set but has retained 95% of the haplotype diversity. In terms of the tagged SNPs, FSA(ω) and FSA(RSQ) have 12 out of 14 in common, while CEMC have 11 in common with FSA(RSQ) and 11 with FSA(ω).

Two Large Simulated Datasets

The intention of our proposed algorithm is to deal with a large number of SNPs. To demonstrate this utility of the algorithm, two large SNP sets were generated using the MS program of Hudson (2002) for analysis. Both datasets were generated based on the same coalescent model, which has a cross-over rate of 20, with the number of base pairs in the locus being 1000 and the gene conversion parameter of 6. Detailed descriptions of these parameters can be found in Hudson (2002). The first dataset, the smaller of the two, was set to have 250 haplotypes and 1000 SNPs. The larger dataset doubles these benchmarks, with 500 haplotypes and 2000 SNPs. Since both simulated datasets resulted in some SNPs having very

small minor allele frequencies, we only used 781 and 1390 SNPs, respectively for the first and the second dataset, that have minor allele frequencies > 0.01 , for our analysis. The CEMC algorithm with various λ values in the range of 0.3-0.7 was applied to each of the datasets. We used Monte Carlo sample size of $N = 5000$ and 8000 for the 781- and the 1390-SNP set, respectively. For both datasets, we set the quantile parameter and the parameters for the stopping rule to be $\rho = 0.1$ and $\varepsilon_1 = \varepsilon_2 = 10^{-6}$. Since there is no publicly available software for k-MIS, we could not compare our results with those that might have been obtained from it. Thus, in addition to CEMC, we also performed random selections (RAND) of SNPs with the same subset sizes as those of the corresponding CEMC to gauge the performance of CEMC.

The results are given in Table 2 For the 781-SNP dataset, the different λ parameters lead to different sizes of the tagged SNP sets, with the size tending to be smaller for a larger λ , as we discussed in Methods. For CEMC, considering the fact that the tagging sets consist of only 4.6 - 6.1 % of the SNPs in the full set, its performance of achieving 81 - 92% of haplotype diversity (based on the RSQ criterion) is very encouraging. As expected, larger tagging sets tend to have higher RSQ values. On the other hand, with random selection, the average RSQ values over 100 such selections merely reached 27% at the most, with a standard deviation of about 8-9%. These results clearly show the value of CEMC, as it did perform much better than random selection, as one would expect of any good tag SNP selection method that can truly capture the haplotype variation in the full set. Similar results are observed for the 1390-SNP set. Despite the performance of random selection being improved considerably, the CEMC algorithm still showed at least 68% greater corresponding RSQ values. Using the same PC as described earlier, it took 78 and 355 minutes for the first and second simulated dataset, respectively.

A Large Real SNP Dataset

This data set consists of 4120 SNPs distributed along chromosome 22 with a median spacing of 4kb, genotyped by the 5' nuclease assay (de la Vega et al. 2002) on 45 DNA samples of Caucasian individuals obtained from the NIGMS Human Variation Panel (Coriell Institute of Medical Research, Camden, NJ). It is particularly interesting to analyze this dataset since its density and sample size are similar to those in the International HapMap Project. This data set was also analyzed by Halldorsson et al. (2004) using their proposed k-MIS algorithm, which is also a block-free SNP selection method. Therefore, in addition to evaluate CEMC relative to random selection, we were also interested in comparing the result of CEMC with that from k-MIS.

We first ran CEMC for a wide range of the λ parameter from 0.1-0.7 (which is expanded from our recommended range of 0.3-0.7) to obtain a good range of the sizes of the tagged SNP sets for comparison purpose. As discussed and seen earlier, the smaller the λ value, the more SNPs were selected. The RSQ values of the selected SNP sets were computed, and they are plotted as the solid curve in Figure 1. For the range of the numbers of SNPs selected by CEMC, we carried out analyses using the k-MIS algorithm and based on random selection. The results in Figure 1 reconfirm our observations with the two simulated datasets, that is, CEMC performed significantly better over random selection. Compared to k-MIS based on the RSQ criterion (dot-dashed curve; results obtained from Halldorsson et al. 2004), CEMC also outperformed with tagging sets of any size considered (\sim 720-1780 SNPs, or 17-43% of full set). The RSQ value is consistently higher, about 4% higher on the average, across the entire range of tagging sets. Although this dataset is seemingly larger than the 1390-SNP simulated dataset, it took only 192 minutes to compute due to the smaller number of observed haplotypes. Regardless of the datasets, the CEMC algorithm converged rather quickly in all of them, usually taking only 12-20 iterations.

4 Discussion

In this paper, we present an adaptation of the cross entropy method of Rubinstein (Rubinstein 1997; Rubinstein and Kroese 2004) and propose an CEMC algorithm for tagging SNP selection. This differs from most of SNP selection algorithms in the literature in that our method is independent of the notion of haplotype block. Thus, the method is applicable to whole genome SNP selection without prior block boundary detection. We have applied this block-free algorithm to three large datasets (two simulated and one real) that are in the order of thousands of SNPs. The successful applications to these large scale datasets demonstrate that CEMC is computationally feasible for whole genome SNP selection. Furthermore, the results show that CEMC is significantly better than random selection, and it also outperformed another block-free selection algorithm for the dataset considered.

Reduction of genotyping cost was the major objective that fanned the research in devising algorithms for tagging SNP selection. With cost-effective mass SNP typing using microarray technology becoming a routine matter, the problem of tagging SNP selection is being pushed into a new stage. Here, the concern may not be genotyping cost, but the reduction of dependency of densely situated SNPs. For whole genome linkage or association studies, such reduction will lead to less burden on multiplicity adjustment as well as more valid assumptions (such as linkage equilibrium among the SNPs) without sacrificing essential information. Thus, our proposed whole genome SNP selection algorithm, and others, will likely to play an important role in the traditional gene mapping arena in the post-genome era.

For demonstration of the CEMC algorithm, we have chosen to use a SNP selection criterion (ω) that is based on both achieving high haplotype diversity as well as low multilocus linkage disequilibrium among the SNPs. For the datasets analyzed in this article, this selection criterion led to satisfactory results evaluated by an objective criterion RSQ. However,

CEMC is applicable to any other SNP selection criterion. For example, for the purpose of reducing the degree of linkage disequilibrium among the SNPs for genomewide linkage analysis, one may be interested in using a criterion that is dependent on the measure of informativeness widely accepted in linkage analysis, such as the entropy based information content measure in Kruglyak et al. (1996).

CEMC is basically a computational approach that turns the combinatorial optimization problem (NP-hard) into a solvable problem (in terms of computational feasibility) and devises an iterative algorithm for it based on the notion of Monte Carlo importance sampling, coupled with the Kullback-Leibler distance for two probability distributions. Thus, although it can be proved that the sequence of tagged SNP sets from the iterative algorithm will converge, following the argument of Margolin (2005), the index set to which the sequence converges may not achieve the global maximum. Whether the global maximum can be achieved, and how fast the algorithm will converge, depend on the starting values of the iterative algorithm, its internal control parameters, as well as the size of the Monte Carlo samples. Our suggested values for these parameters have been tested in our applications, and they appear to work satisfactorily. However, we are by no means claiming optimality for any of these choices. It would of course be of great interest to investigate the settings of these parameters, especially the starting points, that would lead to the global maximum, but this is out of the scope of the present study.

Acknowledgments

This work was supported in part by NSF grant DMS-0306800 and NIH grant 5R01HG002657-03.

- Avi-Itzhak, H.I., Su, X., and De La Vega, F.M. 2003. Selection of minimum subsets of single nucleotide polymorphisms to capture haplotype block diversity. *Pac Symp Biocomput.* 2003, 466-477.
- Byng, M.C., Whittaker, J.C., Cuthbert, A.P., Mathew, C.G., and Lewis, C.M. 2003. SNP subset selection for genetic association studies. *Ann Hum Genet.* 67, 543-556.
- Carlson, C.S., Eberle, M.A., Rieder, M.J., Yi, Q., Kruglyak, L., and Nickerson D.A. 2004. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium, *Am J Hum Genet.* 74, 106-120.
- De La Vega, F.M., Dailey, D., Ziegler, J., Williams, J., Madden, D., and Gilbert, D.A. 2002. New generation pharmacogenomic tools: a SNP linkage disequilibrium Map, validated SNP assay resource, and high-throughput instrumentation system for large-scale genetic studies. *Biotechniques. Suppl,* 48-54.
- Halldrsson, B.V., Bafna, V., Lippert, R., Schwartz, R., De La Vega, F.M., Clark, A.G., and Istraili, S. 2004 Optimal Haplotype Block-Free Selection of Tagging SNPs for Genome-Wide Association Studies *Genome Res.* 14,1633-1640.
- Hudson, R.R. 2002 Generating samples under a Wright-Fisher neutral model. *Bioinformatics* 18, 337-338.
- Ke, X., Cardon, L.R. 2003. Efficient selective screening of haplotypes tag SNPs. *Bioinformatics* 19, 287 -288.
- Kruglyak, L., Daly, M.J., Reeve-Daly, M.P., Lander, E.S. (1996) Parametric and nonparametric linkage analysis A unified multipoint approach. *Am J Hum Genet.* 58, 1347-1363.
- Liu, Z., and Lin, S. 2005 Multilocus LD measure and tagging SNP selection with generalized mutual information. *Genet Epidemiol.* 29, 353-364.

- Margolin, L. 2005. On the convergence of the cross-entropy method. *Annals of Operations Research*. 134, 201-214.
- Rinaldo, A., Bacanu, S., Devlin, B., Sonpar, V., Wasserman, L., and Roeder, K. 2005. Characterization of Multilocus Linkage Disequilibrium. *Genet Epidemiol*. 28, 193-206.
- Rubinstein, R.Y. 1997. Optimization of computer simulation models with rare events. *European Journal of Operations Research*. 99, 89-112.
- Rubinstein, R.Y., Kroese, D.P. 2004 *The Cross-Entropy Method. A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning*. Springer, UK.
- Sebastiani, P., Lazarus, R., Weiss, S.T, Kunkel, K.M., Kohane, I.S., and Ramoni, M.F. 2003 Minimal haplotype tagging. *Proc Natl Acad Sci USA*. 100, 9900-9905.
- Stephens, M., and Donnelly, P. 2003. A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet*. 73, 1162-1169.
- Wall, J.D., and Pritchard, J.K. 2003. Assessing the performance of the haplotype block model of linkage disequilibrium. *Am J Hum Genet*. 73, 502-515.
- Weale, M.E., Depondt, C., Macdonald, S.J., Smith, A., Lai, P.S., Shorvon, S.D., Wood, N.W., and Goldstein, D.B. 2003. Selection and evaluation of tagging SNPs in the neuronal-sodium-channel gene SCN1A: implications for linkage-disequilibrium gene mapping. *Am J Hum Genet* 73, 551-565.
- Zhang, K., Qin, Z., Liu, J., Chen, T., Waterman, MS, and Sun, F. 2004. Haplotype Block Partitioning and Tag SNP Selection Using Genotype Data and Their Applications to Association Studies. *Genome Res*. 14, 908-16.

Table 1: Comparisons of tagging SNP sets and their performances evaluated using the RSQ criterion for three methods on a small dataset.

Algorithm	SNP Index														RSQ
CEMC	1	2	5	7	13	16	17	19	22	25	30	32	41	46	0.953
FSA(ω)	1	5	7	10	13	16	17	22	25	32	39	41	43	46	0.940
FSA(RSQ)	1	5	7	10	13	16	17	22	25	30	31	41	43	46	0.954

Table 2: Sizes of tagging SNP sets and their RSQ values with various λ settings.

Data	Algorithm		λ				
			0.3	0.4	0.5	0.6	0.7
781-SNP Set	CEMC/RAND	No. SNPs	47	48	47	42	36
	CEMC	RSQ	0.921	0.907	0.903	0.872	0.809
	RAND ^a	RSQ	0.265	0.271	0.265	0.212	0.164
		(SD)	0.081	0.097	0.102	0.087	0.092
1390-SNP set	CEMC/RAND	No. SNPs	69	65	64	60	63
	CEMC	RSQ	0.909	0.854	0.819	0.772	0.749
	RAND ^a	RSQ	0.496	0.463	0.457	0.401	0.445
		(SD)	0.032	0.043	0.048	0.039	0.026

^aFor random selection (RAND), the process was repeated 100 times. The RSQ and SD are the average and standard deviation over the 100 selected SNP sets.

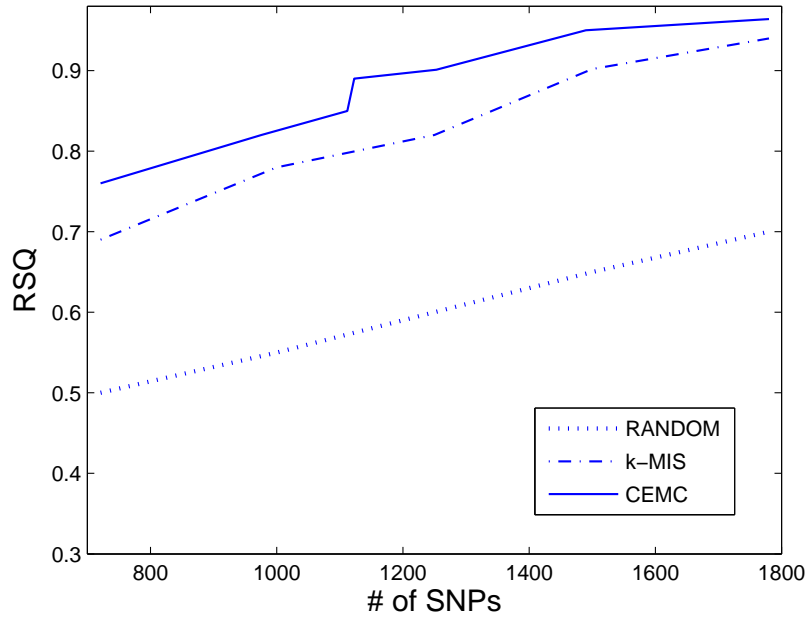


Figure 1: The performance of different algorithms on the Chromosome 22 Caucasian genotype data set. The x-axis represents the number of SNPs selected (size of tagging set), and the y-axis shows the fraction of haplotype RSQ captured. The solid line shows the haplotype RSQ for SNPs chosen with CEMC, while the dashed line are for tagging sets selected with k-MIS. The dotted line shows the RSQ's for SNPs chosen randomly.