J. Appl. Math. & Informatics Vol. **26***(2008), No. 3 - 4, pp. 493 - 500*
Website: http://www.kcam.biz

AN APPROXIMATE GREEDY ALGORITHM FOR TAGSNP SELECTION USING LINKAGE DISEQUILIBRIUM CRITERIA

YING WANG*, ENMIN FENG, RUISHENG WANG

ABSTRACT. In this paper, we first construct a mathematical model for tagSNP selection based on LD measure r^2, then aiming at this kind of model, we develop an efficient algorithm, which is called approximate greedy algorithm. This algorithm is able to make up the disadvantage of the greedy algorithm for tagSNP selection. The key improvement of our approximate algorithm over greedy algorithm lies in that it adds local replacement(or local search) into the greedy search, tagSNP is replaced with the other SNP having greater similarity degree with it, and the local replacement is performed several times for a tagSNP so that it can improve the tagSNP set of the local precinct, thereby improve tagSNP set of whole precinct. The computational results prove that our approximate greedy algorithm can always find more efficient solutions than greedy algorithm, and improve the tagSNP set of whole precinct indeed.

AMS Mathematics Subject Classification : 90C90.
Key words and phrases : TagSNP, SNP marker, Linkage disequilibrium, Local replacement, approximate greedy algorithm.

1. Introduction

The genetic variations in DNA sequences have a major effect on genetic diseases and phenotypic differences. The single nucleotide polymorphism(SNP) is the most frequent form among various genetic variations. SNPs have the significant importance for disease association and drug design. Genome-wide association studies will focus on SNPs because of their high-throughput genotyping[1]. There are more than 10 million verified SNPs in dbSNP(build 124)[2], but current genotyping technology is inadequate and unnecessary for genotyping all SNP markers since the most SNP markers will provide redundant information due to linkage disequilibrium(LD). Hence, only a small subset of representative SNPs is sufficient to capture most information contents of the original set.

Received February 28, 2007. Revised November 10, 2007. *Corresponding author.

Using tagSNPs for association study can greatly reduce the genotyping cost since it needn't genotype all SNPs. Consequently, the selection of tagSNPs has become a very popular topic. Many methods have been proposed to find tagSNPs. These methods are mainly based on three models. The first model assumes that haplotype blocks have been delimited in advance, and some methods are applied to find a minimal set of SNPs which is able to distinguish each pair of haplotypes in a block[3][4][5]. The second model assumes that the number of tagSNPs is first given, and then use various methods to identify tagSNPs which can reconstruct the haplotype of an unknown sample with high accuracy[6][7]. The third model is based on linkage disequilibrium, the methods using the third model are concerned with identifying a minimal set of LD bins such that SNPs within each bin are in high LD. Among above three models, the second model is to find a certain number of SNPs that can predict the haplotype of an unknown sample with high accuracy. However, both LD-based and haplotype block-based methods concentrate on the minimal number of tagSNPs. Although the tagSNPs selection based on haplotype block can distinguish all haplotypes in a block, it is only aiming at haplotype blocks, and the halotype blocks are given in advance. In this paper, we study the third model.

In this manuscript, we first create a simple mathematical model for tagSNPs selection. This model is NP-hard. In order to solve this model, an efficient heuristic algorithm – approximate greedy method with local replacement, is proposed. This method includes two main sections. The first section is concerned with the greedy search[8]. Greedy search is easy to perform, but it can miss more efficient solutions due to selection error of tagSNP for local domain. Hence, in the second section of this method we add local replacement(or local search) into greedy search, tagSNP is replaced by the other SNP having greater similarity degree with it. After each replacement, greedy search is further performed to obtain other tagSNPs. Accordingly, a group of tagSNPs is obtained. Each tagSNP is replaced several times so as to find more efficient solutions. To reduce the computation complexity, all SNP markers from the original set are decomposed into several precincts[9], each of which is made up of markers in strong LD($r^2 \geq 0.5$ or $r^2 \geq 0.8$), our algorithm is preformed in each precinct. The experimental results indicate that our algorithm has obtained much better results than greedy algorithm within reasonable time. This proves that our algorithm has good performance.

2. Model and method

2.1 Model of tagSNP selection

Consider n genotype individuals, each of which contains m bi-allelic SNP markers, denoted by $S = \{a_1, a_2, \ldots, a_m\}$. Almost all the SNP markers have two different alleles, which are common type denoted by 0 and rare type denoted by 1. If two alleles at an SNP site are composed of two common(rare) types, we call this SNP site homozygous site, and the two alleles are shown as 00(11); otherwise call it heterozygous site, and the two alleles are shown as 01 or 10. Hence, a genotype individual may be denoted by a string over $\{0,1,-\}^2$($-$ denoting missing allele or skipped allele). Define a $n \times m$ matrix $M = (m_{ij})$, $m_{ij} \in \{0,1,-\}^2$. Each row of M corresponds to a genotype individual, each column corresponds to a bi-allelic SNP marker, i.e. $a_j(j = 1,2,\ldots,m)$. The elements $m_{ij}(i = 1,2,\ldots,n; j = 1,2,\ldots,m)$ denote a pair of values in the jth site of the ith genotype individual. If m_{ij} corresponds to 01, we assume $m_{ij1} = 0$ and $m_{ij2} = 1$. Further assume that minor allele frequency for each column is above a certain threshold(0.05 was used generally).

The purpose of this paper is to minimize the selected SNP marker for genotyping, simultaneously capturing the most information contents of all SNPs. First, two-SNP haplotype frequencies were estimated[10], and then any pairwise LD measure r^2[11] was calculated using the inferred haplotype frequencies[12]. Consider two SNP markers a_i and a_j corresponding to the ith column and the jth column of matrix M. Let $N_j(x) = |\{m_{ij1}|\ m_{ij1} = x, i = 1,2,\ldots,n\} \bigcup \{m_{ij2}|\ m_{ij2} = x, i = 1,2,\ldots,n\}|$ $(x \in \{0,1\})$ denote the number of allele x in the jth column, and then $p_j(x) = N_j(x)/2n$ indicates its frequency. And let $N_{jk}(x,y)$ and $p_{jk}(x,y) = N_{jk}(x,y)/2n$ denote the number and frequency of 2-SNP haplotypes involving allele x in the jth column and allele y in the kth column, separately. In this paper, we use EM algorithm[13] to obtain 2-SNP haplotype frequencies. Then LD measure for two SNP markers a_j and a_k is shown as $r^2(a_j,a_k) = (p_{jk}(x,y) - p_j(x)p_k(y))^2/(p_j(x)(1-p_j(x))p_k(y)(1-p_k(y)))$. Two markers a_i and a_j are thought to be in strong LD if the LD measure between them is greater than a threshold r_0, namely, $r^2(a_i,a_j) \geq r_0(r_0 = 0.5\ or\ 0.8)$. Then both markers a_i and a_j are thought as tagSNPs for each other. That is, a_i can be used as a surrogate for a_j, or vice versa. Let $C(a_k,S) = \{a_i|\ r^2(a_k,a_i) \geq r_0,\ a_i \in S\ and\ a_i \neq a_k\}$ denote the set that contains all the markers being in strong LD with a_k. At the same time, a_k is regarded as tagSNP of the set $C(a_k,S)$. Assume S_{sub} to denote a subset of S, our purpose is to find a minimal subset of S, such that SNP markers in the subset can represent all other SNP markers. Consequently, we can construct the model as follows:

$$\begin{aligned} \min \quad & |S_{\text{sub}}| \\ s.t. \quad & \bigcup_{a_k \in S_{\text{sub}}} (C(a_k,S) \cup \{a_k\}) = S \end{aligned} \tag{1}$$

where $|\cdot|$ denotes the number of elements in a subset.

2.2 Approximate greedy method with local replacement

First, partition all m SNP markers into several disjoint precincts(referred in [9]), any pairwise markers from different precincts show weak LD, so selecting tagSNPs in the original set can be regarded as selecting tagSNPs in each precinct, and finally combining all tagSNPs together. Precinct partition can reduce computation complexity.

The precinct partition depends on pairwise LD measure r^2. Each precinct is made up of the markers in strong LD. We applied the Breath First Search algorithm[14]. Starting from a new marker(not being partitioned) in a new precinct, this algorithm adds all markers in strong LD with it and all makers being in strong LD with the added markers into the precinct, until there are no markers in strong LD with all the markers in the precinct. This process is restarted from new markers until all markers are assigned to a precinct. The concrete operation steps are shown in Table 1.

TABLE 1. Precinct partition process

Input: A set of m markers $S = \{a_1, a_2, \dots, a_m\}$. Step 0. $Q = S$,$J = 1$,$k = 1$.
Step 1. Let $P_{Jk} = \{a_i\}, a_i \in Q$.
Step 2. Set $P_{J(k+1)} = \Big(\bigcup\limits_{a_j \in P_{Jk}} C(a_j, S)\Big) \bigcup P_{Jk}$ and $Q = Q \backslash P_{J(k+1)}$.
Step 3. If $Q = \phi$, let $P_J = P_{J(k+1)}$, stop.
Step 4. If $P_{J(k+1)} = P_{Jk}$, let $P_J = P_{J(k+1)}$, $J = J + 1$ and $k = 1$.
return to step 1; else let $k = k + 1$, return to step 2.
Output: Precinct number J and marker set of each precinct $P_k, k = 1, 2, \dots, J$.

Greedy search algorithm is very simple, but it is not necessary to obtain the optimal solution or much better solutions by it. Greedy search algorithm can miss more efficient solutions due to selection error of tagSNPs for local precinct. In order to make up this disadvantage, we present an approximate greedy search algorithm, that adds local replacement(or called local search) into greedy search. Since tagSNP selection is performed in each precinct, all operations are aiming at the precinct. Before describing the local replacement, first assume that $C(a_i, P) = \{a_j |\ r^2(a_i, a_j) \geq r_0, a_j \in P \ and \ a_j \neq a_i\}$(P denotes a precinct), then define similarity degree as: For two different markers a_i and a_j from the same precinct, the number of the same markers in $C(a_i, P)$ and $C(a_j, P)$ is called the similarity degree of a_i and a_j in P, denoted as $SD(a_i, a_j) = |C(a_i, P) \bigcap C(a_j, P)|$. The local replacement means that tagSNP of a precinct is replaced with SNP having higher similarity degree with it. Figure 1 shows a simple precinct. It is easy to calculate its tagSNP set – $\{C, D, B\}$ using greedy search algorithm, while the optimal tagSNP set is composed of only two tagSNPs A and B. Our approximate greedy algorithm can find the optimal solution by local replacement. In the precinct, tagSNP C is replaced with SNP A that has higher similarity degree with it, then perform greedy search to obtain the optimal solution $\{A, B\}$.

$|C(a_i, P)|$ represents the number of SNP markers in P having strong LD with a_i. For the SNP marker satisfying $|C(a_i, P)| = 0$, we call them isolated markers of P, there exists no marker having strong LD with it, hence, a_i must be tagged by itself. For each precinct $P_k (k = 1, 2, \dots, J)$, our algorithm is generalized as follows:

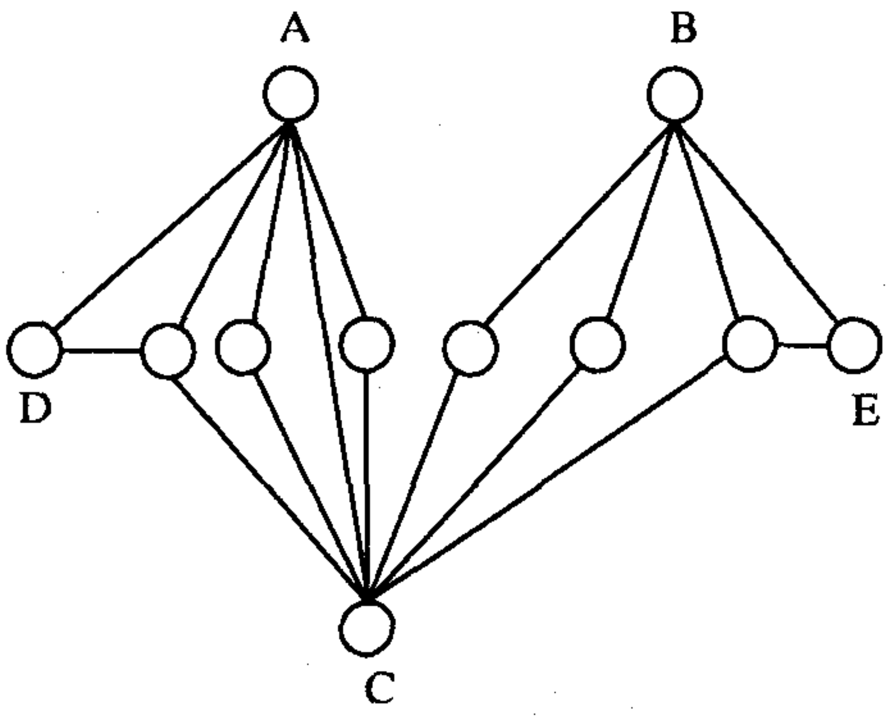

Figure 1.

Approximate greedy algorithm

```
Section I (greedy search)
Step 0. Set a tagSNP set $T_k = \phi$, let $P = P_k$ and $t = 0$.
Step 1. For any $a_i \in P$, calculate $|C(a_i, P)|$.
Step 2. For any $a_i \in P$, if $|C(a_i, P)| = 0$, then let $t = t + 1$, $t_{kt} = a_i$,
        $T_k = T_k \bigcup \{t_{kt}\}$ and $P = P \backslash \{t_{kt}\}$.
Step 3. If $P = \phi$, then stop.
Step 4. Set $a_{max} = argmax\{|C(a_i, P)|,\ a_i \in P\}$, let $t = t + 1$, $t_{kt} = a_{max}$,
        $T_k = T_k \bigcup \{t_{kt}\}$ and $P = P \backslash (\{a_{max}\} \bigcup C(a_{max}, P))$.
Step 5. If $P = \phi$, then stop; else return to step 1.

Section II (approximate greedy search)
Step 0. Set a tagSNP set $T_k^{'} = \phi$, let $P^{'} = P_k$, $t = 0$ and $z = 0$.
Step 1. For any $a_i \in P^{'}$, calculate $|C(a_i, P^{'})|$.
Step 2. If $|T_k| = 1$, then stop.
Step 3. Let $j = t + 1$, $t_{kj}^{''} = t_{kj}$, and calculate $SD(t_{kj}^{''}, a_i)$ for any $a_i \in P^{'}$.
Step 4. Let $t = j$, $a_{maxSD} = argmax\{SD(t_{kj}^{''}, a_i),\ a_i \in P^{'}\}$, and $t_{kj}^{'} =$
        $a_{maxSD}$, $T_k^{'} = T_k^{'} \bigcup \{t_{kj}^{'}\}$, $P^{'} = P^{'} \backslash (\{a_{maxSD}\} \bigcup C(a_{maxSD}, P^{'}))$.
        If $P^{'} = \phi$, then let $|T_k| = |T_k| + 1$ and stop.
Step 5. If $SD(t_{kj}^{''}, a_{maxSD}) = 0$, perform step 12.
Step 6. For any $a_i \in P^{'}$, calculate $|C(a_i, P^{'})|$.
Step 7. Let $a_{max} = argmax\{|C(a_i, P^{'})|, a_i \in P^{'}\}$, $t = t + 1$, $t_{kt}^{'} = a_{max}$,
        $T_k^{'} = T_k^{'} \bigcup \{t_{kt}^{'}\}$ and $P^{'} = P^{'} \backslash (\{a_{max}\} \bigcup C(a_{max}, P^{'}))$.
Step 8. If $P^{'} \neq \phi$, return to step 6.
Step 9. If $|T_k| > |T_k^{'}|$, set $T_k = T_k^{'}$ and $t_{ki} = t_{ki}^{'} (i = 1, 2, \ldots, |T_k^{'}|)$.
Step 10. If $z > z_0 (z_0 = 4,\ 5\ or\ 6)$, then go to Step 12.
Step 11. Let $P^{'} = P_k \backslash \bigcup_{t=1}^{j-1} (\{t_{kt}\} \cup C(t_{kt}, P_k))$, $SD(T_{kj}^{''}, a_{maxSD}) = 0$ and
         $z = z + 1$, return to step 4.
Step 12. If $j = |T_k|$, then stop; else let $P^{'} = P_k \backslash \bigcup_{t=1}^{j} (\{t_{kt}\} \bigcup C(t_{kt}, P_k))$,
```

$j = j + 1$, and $t''_{kj} = t_{kj}$, return to step 4.

3. Results

In order to verify our approximate greedy algorithm, we applied this algorithm and greedy algorithm[8] to four ENCODE regions of all four populations genotyped by the HapMap project(release 16c, June 2005). These four populations are CHB(Chinese), YRI(Yoruban), CEU(European) and JPT(Japanese). The following four tables show the results of four ENCODE regions for all four populations, separately. From Tables 2-5, we can observe that the results of our algorithm are always superior to those of the greedy algorithm. For the region ENr213 in the CHB example, the greedy algorithm identified 47 tagSNPs, while our algorithm identified 42 tagSNPs, the reduction rate is over 10%. When the r^2 threshold is 0.5, the reduction rate of our algorithm relative to greedy algorithm is much bigger than the r^2 threshold 0.8. Under the threshold 0.5, the reduction rate of four populations is between $2.5\% - 11\%$. The average reduction rate for all populations is 5.4%. Compared over all four populations, the number of tagSNPs for YRI population is about twice to three times bigger than that of the other three populations.

TABLE 2. CHB example for ENCODE region

Region	ENr123	ENr213	ENr321	ENr112
No. of SNPs	794	861	610	831
$r^2 \geq 0.5$				
No. of precincts	32	29	42	28
No. of tagSNPs(Greedy)	54	47	68	52
No. of tagSNPs(Our algorithm)	50	42	64	50
$r^2 \geq 0.8$				
No. of precincts	104	79	83	97
No. of tagSNPs(Greedy)	129	89	107	111
No. of tagSNPs(Our algorithm)	124	88	103	108

4. Discussion

In this manuscript, we first present a model for tagSNP selection based on LD measure r^2, then develop an efficient algorithm, which is approximate greedy algorithm with local replacement. This algorithm is able to make up the disadvantage of greedy algorithm. The key advantage of our approximate algorithm over greedy algorithm lies in that it adds local replacement(or local search) into greedy search, so that it has much more chance to find the optimal solution or the better solution. In our algorithm, a tagSNP has been replaced for no more than six times generally. The experimental results prove that our approximate greedy algorithm can always find more efficient solutions than greedy algorithm, and for some examples, the improvement is very obvious. Our algorithm is performed

TABLE 3. JPT example for ENCODE region

Region	ENr112	ENr213	ENr321	ENm014
No. of SNPs	827	498	592	533
$r^2 \geq 0.5$				
No. of precincts	28	26	41	27
No. of tagSNPs(Greedy)	62	45	59	48
No. of tagSNPs(Our algorithm)	57	43	57	45
$r^2 \geq 0.8$				
No. of precincts	107	82	104	89
No. of tagSNPs(Greedy)	120	89	118	106
No. of tagSNPs(Our algorithm)	117	87	116	104

TABLE 4. CEU example for ENCODE region

Region	ENr113	ENr131	ENm010	ENm013
No. of SNPs	1080	989	524	692
$r^2 \geq 0.5$				
No. of precincts	44	81	44	28
No. of tagSNPs(Greedy)	78	114	69	40
No. of tagSNPs(Our algorithm)	74	108	66	39
$r^2 \geq 0.8$				
No. of precincts	137	185	141	72
No. of tagSNPs(Greedy)	154	199	161	90
No. of tagSNPs(Our algorithm)	153	197	159	85

TABLE 5. YRI example for ENCODE region

Region	ENr112	ENr123	ENr321	ENm013
No. of SNPs	912	802	861	736
$r^2 \geq 0.5$				
No. of precincts	86	58	109	40
No. of tagSNPs(Greedy)	163	105	162	90
No. of tagSNPs(Our algorithm)	152	96	156	84
$r^2 \geq 0.8$				
No. of precincts	322	213	313	181
No. of tagSNPs(Greedy)	346	237	320	198
No. of tagSNPs(Our algorithm)	343	231	320	196

on a 1.8 GHz Pentium 3 PC using Microsoft Visual C++. The running time of each test is within seven minutes, and even less than one minute. This proves that our algorithm has good performance, it can find much better solutions within reasonable time, it has made up the disadvantage of greedy algorithm effectively. Our algorithm may also be applied to solve large-scale chromosome sample. Each genotype individual has thousands of SNPs, and these SNPs can be separated into thousands of precincts, we may adopt the parallel computation to speed up computation.

References

1. F.S. Collins, M.S. Guyer, and A. Chakravarti, *Variations on a theme: cataloging human DNA sequence variation*, Science, **278**(1997), 1580-1581.
2. R. Sachidanandam, D. Weissman, S.C. Schmid, et al. *A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms*, International SNP Map Working Group, Nature, **409**(2001), 928-933.
3. N. Patil, A.J. Berno, D.A. Hinds, et al. *Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21*, Science, **294**(2001), 1719-1723.
4. K. Zhang, M. Deng, et al. *A dynamic programming algorithm for haplotype partitioning*, Proc. Natl Acad. Sci. USA, **99**(2002), 7335-7339.
5. C.J. Chang, Y.T. Huang, K.M. Chao, *A greedier approach for finding tag SNPs*, Bioinformatics, **22**(2006), 685-691.
6. E. Halprin, G. Kimmel, R. Shamir, *Tag SNP selection in genotype data for maximizing SNP prediction accuracy*, Bioinformatics, **21**(Suppl. 1)(2005), i195-i203.
7. J.W. He, A. Zelikovsky, *Linear reduction methods for tag SNP selection*, Int. J. Bioinformatics Res. Appl., **1**(2005), 249-260.
8. C.S. Carlson, M.A. Eberle, M.J. Rieder, et al. *Selecting a maximally informative set of single-numcleotide polymorphisms for association analysis using linkage disequilibrium*, Am.J.Hum.Genet., **74**(2004), 106-120.
9. Z.H. Qin, S. Gopalakrishnan, and G.R. Abecasis, *An efficient comprehensive search algorithm for tagSNP selection using linkage disequilibrium criteria*, Bioinformatics, **22** (2006), 220-225.
10. W.G. Hill, *Estimation of linkage disequilibrium in randomly mating populations*, Heredity, **33**(1974), 229-239.
11. B. Devlin and N. Risch, *A comparison of linkage disequilibrium measures for fine-scale mapping*, Genomics, **29**(1995), 311-322.
12. W.G. Hill and A. Robertson, *The effects of inbreeding at loci with heterozygote advantage*, Genetics, **60**(1968), 615-628.
13. S.T. Kalinowski and P.W. Hedrick, *Estimation of linkage disequilibrium for loci with multiple alleles: basic approach and an application using data from lighorn sheep*, Heredity, **87**(2001), 698-708.
14. T.H. Cormen, et al. *Introduction to algorithms*. 2nd edition. MIT Press, Cambridge, 2001.

Ying Wang is now completing her doctors degree in Dalian University of Technology. Her research interests involve DNA sequence analysis , haplotype assembly problem and tag SNP selection.

Department of Applied Mathematics, Dalian University of Technology, Dalian 116024, P.R.China

e-mail: `wwangying2003@@yahoo.com.cn`

Enmin Feng is a professor and Ph.D. advisor in Dalian University of Technology. His research interests center on control theory and optimization.

Department of Applied Mathematics, Dalian University of Technology, Dalian 116024, P.R.China

e-mail: `emfeng@@dult.edu.cn`

Ruisheng Wang is pursuing her work in Renmin University of China. His research interests cover haplotype assembly problem and tag SNP selection.

School of Information, Renmin University of China, Beijing 100872, China

e-mail: `wangrsh@@amss.ac.cn`