Haplotype Block Partitioning and TagSNP Selection using Mutual Information[☆]

Katanforoush A. [1,2], Sadeghi M. [1,2], Pezeshk H. [3]

Abstract

Most experimental and computational studies on high-throughput SNP data are applicable only on a limited region of genome or limited subset of SNPs. Using mutual information on SNP pairs, we define the regions of high information consistency as haplotype blocks. Applying the dynamic programming approach with the proposed score on HapMap samples results in wider chromosomal blocks while adopting the recombination hotspots.

Keywords: Haplotype Blocks, HapMap, TagSNP selection

Introduction

Current high-throughput techniques in molecular genetics can determine millions of single nucleotide polymorphism on human genome. These data provide informative materials for disease association studies and analyzing models of population genetic. Most practical and computational studies on these genome-wide data are applicable only on a limited region of genome or limited subset of SNPs. Using the mutual information of SNP pairs, we define the chromosomal regions of high information consistency as haplotype blocks. We have incorporated dynamic programming to find the block partitioning maximizing the sum of block scores. Also using the linear programming approaches for the problem of dominating set, we develop an efficient algorithm for tagSNP selection. The power of the case-control association test obtained by this tagging method also compared with some other common methods. It is shown that the proposed tagging SNP's improves levels of performance while holding the desired accuracy.

Method

Essentially the mutual information reflects the genetic association between two SNPs. By doing some algebra, it would be shown that mutual information asymptotically tends to the logarithm of the Fisher exact test value.

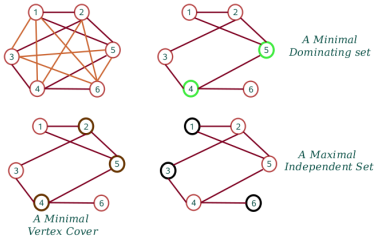
$$I(i,j)=\sum_{\substack{a\in\{0,1\}\\ b\in\{0,1\}}} P(i=a\wedge j=b).\log_2\left(\frac{P(i=a\wedge j=b)}{P(i=a).P(j=b)}\right)$$

$$F_{ss}=\frac{\binom{n_{1X}}{n_{11}}\binom{n_{0X}}{n_{01}}}{\binom{n}{n_{X1}}}$$

The association by itself may be affected by several kinds of evolutionary events such as selections, migrations and genetic drift. Such messy bunch of factors behaves like a noisy channel communicates words from one SNP to the other. Assuming that there is no other operation affecting SNPs then the capacity of such a channel is determined by the mutual information. Therefore high mutual information shows a strong associated pair of SNPs. The maximum value for the mutual information defined for bi-allelic SNPs is one bit per message and its minimum is zero.

TagSNP selection

The general approach to define tagging SNPs aims at finding smallest subset of SNPs that any difference between every two distinct haplotypes can be captured by at least one tagSNP. These methods rely on some parsimonious assumption on haplotype diversity. Instead, we devise a method based on minimal dominating set from graph theory. By this approach every SNP either is a tagSNP or is in strong association with at least one tagging SNP.

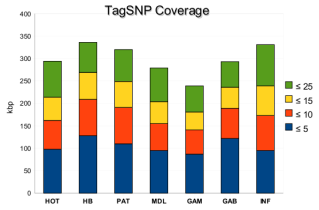


Method comparison

Among methods for haplotypes block partitioning, we have selected some of the most cited (Table 1). Also, recombination hotspots can realize boundaries of haplotype blocks on genome. In hotspot regions the recombination rate is high and the probability of observing less diverse haplotype is low. Hotspots for all individuals are not necessarily the same. However, there are some regions on genome which could be recognized as the consensus.

Result

It is observed that the blocks produced by LD-based methods like Gabriel, Infres also the blocks defined by hotspots are broader than what determined by the greedy algorithm and 4-gamete test.



In general, the methods producing broader blocks entail less number of tagSNPs. It reveals that there is not any essential relation between the minimal number of tagSNPs and grouping high associated SNPs within blocks. The higher numbers corresponded to the wider blocks indicate that a usual tagSNP is capable of capturing much longer haplotypes than those broken by diversity criteria. The conventional methods based on block diversity usually generate more redundant tagSNPs over all genome.

The sum of all mutual information of SNP pairs in a block is introduced as score after subtracting the sum by a constant factor which bounds the type I error.

$$S(i;d)=\sum_{i-d\le s<t\le i} f(s,t)-\theta$$

For each interval of length d ending at SNP i, $S(i;d)$ denotes the score of the block and shows how strong a certain set of SNPs tends to be integrated into a block.

The free parameter, $\theta\in[0,1]$ incorporated in the formula allows to alter average length of blocks by the following optimization step. For θ's close to zero, the algorithm tends to produce wider blocks while for θ's close to one, it produces smaller blocks same as the 4-gamete test approach.

The optimum solution is obtained using a dynamic programming approach. The number of fundamental operations to calculate blocks scores and to find optimal partitioning is $O(nwL)$ where n is the number of haplotype samples, w is the window size, and L is the number of SNPs. In general, we refer to the method described above by "Infres".☼

Linear Programming

0/1 variable x_j for every SNP

$x_j = 1$ if SNP j is selected as a tag SNP

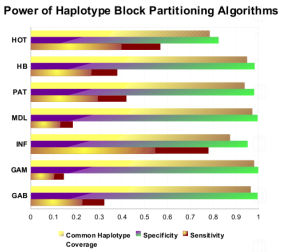
$x_j = 0$ otherwise

$$Min \sum_{j=1}^{m} x_j$$

$$s.t. \sum_{j=1}^{m} f_{ij} x_j \ge k,\ i=1,\ldots,m$$

Method	Block criteria	Implementation	Literature
Patil's	diversity / greedy algorithm	HapBlock v.3	Patil et al. 2001
HapBlock	diversity / dynamic programming	HapBlock v.3	K. Zhang and L. Jin 2003
Gabriel's	pairwise LD	Haploview v.4	Gabriel et al. 2002
4 gamete test	perfect phylogeny	Haploview v.4	Wang et al. 2002
MDBlock	Minimum Description Length	MDBlock v.1	Anderson and Novembre 2003
Hotspots	Fine-scale LD map		Myres et al. 2005

Table 1: Haplotype Block partitioning methods



Conclusion

Finding minimum number of tagSNPs can pose a two fold object. The first is the genome partitioning in a way that reported blocks could be captured by as limited number of tagSNPs as possible. The second object is to get fewer numbers of tagSNPs to be associated with all SNPs. Results obtained by available methods show that in practice there is a gap between two extremes. The algorithms aimed to recognize evolutionary conserved haplotype blocks such as greedy diversity based algorithms, four-gamete-test, and Gabriel's LD based method obtain the blocks which are tagged by just few number of tagSNPs, but they assign too many tagSNPs for the whole chromosome. On the other hand, the blocks being introduced by hotspots neglect the diversity considerations while giving less number of tagSNPs. The information theoretic method, Infres, seems to cover this gap by satisfying both the objects on average.

References

Anderson,E.C. and Novembre,J. (2003) Finding Haplotype Block Boundaries by Using the Minimum-Description-Length Principle. *The American Journal of Human Genetics*, 73, 336-354.

Gabriel,S. et al. (2002) The structure of haplotype blocks in the human genome. *Science*, 296, 2225-2229.

Myers,S. et al. (2005) Genetics: A fine-scale map of recombination rates and hotspots across the human genome. *Science*, 310, 321-324.

Patil,N. et al. (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science*, 294, 1719-1723.

Wang,N. et al. (2002) Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. *Am. J. Hum. Genet.*, 71, 1227-34.

Zhang,K. and Jin,L. (2003) HaploBlockFinder: Haplotype block analyses. *Bioinformatics*, 19, 1300-1301.

(1) Department of Bioinformatics, Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran,
(2) Institute for Studies in Theoretical Physics and Mathematics (IPM), Niavaran Square, Tehran, Iran,
(3) School of Mathematics, Statistics and Computer Science, University College of Science, University of Tehran, Tehran, Iran

☆ This research was partially supported by grants from IPM. (No. CS 1385-1-02)

○ World Wide Web Availability
http://bioinf.cs.ipm.ac.ir/softwares/infres