Computational Biology and Chemistry xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Research Article

Fast detection of high-order epistatic interactions in genome-wide association studies using information theoretic measure

Sangseob Leem[a], Hyun-hwan Jeong[a], Jungseob Lee[b], Kyubum Wee[a], Kyung-Ah Sohn[a,*]

[a] Department of Information and Computer Engineering, Ajou University, Suwon 443-749, South Korea
[b] Department of Mathematics, Ajou University, Suwon 443-749, South Korea

ARTICLE INFO

Article history:
Accepted 23 December 2013

Keywords:
Genome-wide association studies
High-order epistatic interactions
K-means clustering
Mutual information
WTCCC

ABSTRACT

There are many algorithms for detecting epistatic interactions in GWAS. However, most of these algorithms are applicable only for detecting two-locus interactions. Some algorithms are designed to detect only two-locus interactions from the beginning. Others do not have limits to the order of interactions, but in practice take very long time to detect higher order interactions in real data of GWAS. Even the better ones take days to detect higher order interactions in WTCCC data.

We propose a fast algorithm for detection of high order epistatic interactions in GWAS. It runs *k*-means clustering algorithm on the set of all SNPs. Then candidates are selected from each cluster. These candidates are examined to find the causative SNPs of *k*-locus interactions. We use mutual information from information theory as the measure of association between genotypes and phenotypes.

We tested the power and speed of our method on extensive sets of simulated data. The results show that our method has more or equal power, and runs much faster than previously reported methods. We also applied our algorithm on each of seven diseases in WTCCC data to analyze up to 5-locus interactions. It takes only a few hours to analyze 5-locus interactions in one dataset. From the results we make some interesting and meaningful observations on each disease in WTCCC data.

In this study, a simple yet powerful two-step approach is proposed for fast detection of high order epistatic interaction. Our algorithm makes it possible to detect high order epistatic interactions in GWAS in a matter of hours on a PC.

1. Introduction

There have been many genome-wide association studies (GWAS) to identify associations between phenotypes and whole genome information. Some studies have been successful in finding associations between single SNPs and disease susceptibilities. However, large part of genetic effects to phenotypes remains unresolved. Main reason of this deficiency is that many diseases are associated with effects of gene–gene interactions (Cordell, 2009). These effects are called epistatic interactions (epistasis) (Moore and Williams, 2009). Many studies demonstrated existence of epistatic interactions in such diseases as breast cancer (Ritchie et al., 2001), coronary heart disease (Nelson et al., 2001) and Alzheimer's disease (Zubenko et al., 2001).

Many algorithms for detecting epistatic interactions have been proposed. These algorithms can be largely classified into four categories: exhaustive search methods, stochastic approaches, data mining/machine learning approaches and stepwise approaches.

Exhaustive search strategies test every possible combination of *k*-SNPs. They are guaranteed to find the most susceptible *k*-SNP, but take a long time to execute. CPM (Nelson et al., 2001) enumerates all possible genotype combinations, and its use is limited to candidate SNP association studies due to long execution time. RPM (Culverhouse, 2007) reduces execution time compared with CPM, but its use is still limited. MDR (Ritchie et al., 2001) is a representative of exhaustive search method, which has been applied to many clinical studies. MDR divides all genotypes into one of two categories instead of enumerating all genotypes, which makes it faster than CPM or RPM. However, it is still short of genome-wide association studies. TEAM (Zhang et al., 2010), BOOST (Wan et al., 2010a) and BiForce (Gyenesei et al., 2012) reduce execution time by devising clever data structures. They can be used in GWAS, but are limited to detecting only two-locus interactions.

Stochastic approaches use random sampling and BEAM (Zhang and Liu, 2007) is a representative of them. BEAM maximizes the probability of models by Markov Chain Monte Carlo (MCMC). These algorithms usually have many parameters whose values are set by users, and execution times depend on the values of these

* Corresponding author. Tel.: +82 31 219 2434.
E-mail addresses: leemss@ajou.ac.kr (S. Leem), libe@ajou.ac.kr (H.-h. Jeong), jslee@ajou.ac.kr (J. Lee), kbwee@ajou.ac.kr (K. Wee), kasohn@ajou.ac.kr (K.-A. Sohn).

1476-9271/$ – see front matter © 2014 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compbiolchem.2014.01.005

Table 1
Information theoretic measures.

Measure	Formula
Entropy of a discrete random variable X	$H(X) = -\sum_{x \in X} p(x)\log(p(x))$
Joint entropy of random variables $X_1, X_2, \ldots, X_k$	$H(X_1, \ldots X_k) = -\sum_{x_1} \ldots \sum_{x_k} p(x_1, \ldots, x_k)\log(p(x_1, \ldots, x_k))$
Mutual information of $X_1, X_2, \ldots, X_k$ and Y	$I(X_1, X_2, \ldots, X_k ; Y) = H(X_1, X_2, \ldots, X_k ; Y) + H(Y) - H(X_1, X_2, \ldots, X_k ; Y)$

parameters. In most cases, long execution times are required to obtain good results.

There are many methods of data mining/machine learning approaches. Neural networks (Serretti and Smeraldi, 2004), decision trees (Xie et al., 2005), and random forests (Meng et al., 2009; Winham et al., 2012) have been applied to detect epistatic interactions among candidate SNPs. SNPRuler (Wan et al., 2010b) employs predictive rule learning to make it possible to detect epistasis in GWAS. Various methods of data mining/machine learning approaches have their own pros and cons. In general, they are faster than exhaustive approaches, but may find local optima instead of global ones.

Stepwise approaches consist of two stages: filtering stage and search stage. At the filtering stage, the SNPs that are not meaningful are filtered out; at the search stage, survived SNPs are tested for epistatic interactions. At the search stage, any of exhaustive, stochastic, data mining/machine learning approaches can be used. Feature selection methods Relief (Kenji Kira, 1992) and ReliefF (Kononenko, 1994) are adopted and modified to TuRF (Moore and White, 2007) and SURF (Greene et al., 2009) to be used as filtering methods for GWAS. SNPHarvester (Yang et al., 2009) uses local search algorithm to select the SNPs to be tested at the next stage. Screen and clean (Wu et al., 2010) uses lasso logistic regression to select the significant SNPs. BOOST (Wan et al., 2010a) uses likelihood ratio test to prune insignificant SNPs. EDCF (Xie et al., 2012) recursively detects d-locus interactions from the results of $(d-1)$-locus interactions.

There are many algorithms for detecting epistatic interactions in GWAS. However, most of these algorithms are applicable only for detecting two-locus interactions. Some algorithms are designed to detect only two-locus interactions from the beginning. Such cases are TEAM, BOOST, and BiForce. Others do not have limits to the order of interactions, but in practice take very long time to detect higher order interactions in real data of GWAS. Even the better ones take days to detect higher order interactions in WTCCC data (Burton et al., 2007). SNPRuler reports that it took ~2 days to analyze one WTCCC dataset. SNPHarvester reports that it took about 2 weeks to handle WTCCC data.

We propose a fast algorithm for detection of high order epistatic interactions in GWAS. Our algorithm can analyze 5-locus interactions in a WTCCC dataset in a few hours on a PC. It runs k-means clustering algorithm on the set of all SNPs. Then candidates are selected from each cluster. These candidates are examined to find the causative SNPs of k-locus interactions. We use mutual information from information theory as the measure of distance between two SNPs for clustering. The details of the method, including how candidates are selected and why mutual information is used, will be explained in Section 2.

We tested the power and speed of our method on extensive sets of simulated data. The results show that our method has equal or more power, and runs much faster than previously reported methods including SNPRuler and SNPHarvester. We also applied our algorithm on each of seven diseases in WTCCC data to analyze up to 5-locus interactions. It takes only a few hours to analyze 5-locus interactions in one dataset. From the results we make some interesting and meaningful observations on each disease in WTCCC data. Our algorithm makes it possible to detect high order epistatic interactions in GWAS in a matter of hours on a PC.

2. Methods

2.1. Background on information theoretic measures

Chi-square tests are often used as the association measure in GWAS. However, they are not appropriate to use for the cases where there are cells whose frequencies in contingency tables are less than 5. In the studies of detecting high order interactions, it is highly probable that contingency tables have cells of very low frequencies. Thus we used *mutual information* as the measure of association between genotypes and susceptibilities of diseases.

Mutual Information $I(X;Y)$ represents the amount of the information shared by random variables X and Y. The definition of $I(X;Y)$ and its relation to other information theoretic measures are given in Table 1. In our study X represents the genotype value (minor/minor, major/minor, or major/major) of a particular SNP among samples, and Y represents whether a sample has the disease or not (case or control). Let us call Y *disease state* from this point on. Higher value of $I(X;Y)$ implies the SNP has stronger association with the disease. Note that mutual information can be extended to accommodate multiple SNPs.

In this study, entropy is defined in terms of partitions. The set of samples is partitioned by the values of a SNP, and also partitioned into cases or controls. Let $X = \{A_1, A_2, \ldots, A_n\}$ be a partition of S. In other words, $S = A_1 \cup A_2 \cup \ldots \cup A_n$ and $A_i \cap A_j = \emptyset$ for distinct i and j. The entropy $H(X)$ of X is defined as follows:

$$H(X) = -\sum_{i=1}^{n} \frac{|A_i|}{|S|} \log \frac{|A_i|}{|S|}$$

where $|\cdot|$ denotes the number of the elements of a set. We can extend the definition to any number of partitions. Let $X_j = \{A_1^{(j)}, A_2^{(j)}, \ldots, A_{n_j}^{(j)}\}$, for $j = 1, 2, \ldots, k$, be partitions for a set S. The joint entropy $H(X_1, X_2, \ldots, X_k)$ of the partitions $X_1, X_2, \ldots, X_k$ is defined as

$$-\sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2}\cdots\sum_{i_k=1}^{n_k} \frac{\left|A_{i_1}^{(1)} \cap A_{i_2}^{(2)} \cap \ldots \cap A_{i_k}^{(k)}\right|}{|S|} \log \frac{\left|A_{i_1}^{(1)} \cap A_{i_2}^{(2)} \cap \ldots \cap A_{i_k}^{(k)}\right|}{|S|}$$

Now the mutual information between the joined partition of X_1, $X_2, \ldots X_k$ and a partition Y can be defined as follows:

$$I(X_1, X_2, \ldots, X_k; Y) = H(X_1, X_2, \ldots, X_k) + H(Y) - H(X_1, X_2, \ldots, X_k, Y)$$

Let $X_1, X2, \ldots, X_k$ be partitions for the set of samples induced by the genotypes of the SNPs SNP_1, SNP_2, …, SNP_k, respectively, and Y be the partition by disease state (case or control). Then $I(X_1, X_2, \ldots, X_k ; Y)$ represents the degree of association between genotypes of SNP_1, SNP_2, …, SNP_k and the disease state.

In this study, we try to find the set of k SNPs that maximizes the value $I(X_1, X_2, \ldots, X_k ; Y)$.

Please cite this article in press as: Leem, S., et al., Fast detection of high-order epistatic interactions in genome-wide association studies using information theoretic measure. Comput. Biol. Chem. (2014), http://dx.doi.org/10.1016/j.compbiolchem.2014.01.005