Journal of Theoretical Biology 366 (2015) 84–90

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

A new technique for generating pathogenic barcodes in breast cancer susceptibility analysis

Xiong Li*, Bo Liao, Haowen Chen

Key Laboratory for Embedded and Network Computing of Hunan Province, The College of Information Science and Engineering, Hunan University, Changsha 410082, Hunan, China

HIGHLIGHTS

- A criteria of maximum dissimilarity–minimum entropy is proposed for identifying pathogenic barcodes.
- Low entropy indicates a relatively consistent pattern to cause disease in case samples.
- Large dissimilarity indicates significant distinction between cases and controls.
- Large dissimilarity pathogenic barcodes with consistent pattern in cases are risky.
- From the perspective of statistics, if a shorter barcode contributes to complex diseases, the complex diseases may be more common in population.

ARTICLE INFO

Article history:
Received 22 July 2014
Received in revised form
8 October 2014
Accepted 4 November 2014
Available online 13 November 2014

Keywords:
Breast cancer
Single-nucleotide polymorphism
Epistasis
Entropy
Odds ratio

ABSTRACT

Complex diseases usually involve complex interactions between multiple loci. The artificial intelligent algorithm is a plausible strategy to evade combinatorial explosion. However, the randomness of solution of this algorithm loses decreases the confidence of biological researchers on this algorithm. Meanwhile, the lack of an efficient and effective measure to profile the distribution of cases and controls impedes the discovery of pathogenic epistasis. Here we present an efficient method called maximum dissimilarity–minimum entropy (MDME) to analyze breast cancer single-nucleotide polymorphism (SNP) data. The method searches risky barcodes, which to increase the odds ratio and relative risk of the breast cancer. This method based on the hypothesis that if a specific barcode is associated with a disease, then the barcode permits distinction of cases from controls and more importantly it shows a relative consistent pattern in cases. An analysis based on simulated dataset explains the necessity of minimum entropy. Experimental results show that our method can find the most risky barcode that contributes to breast cancer susceptibility. Our method may also mine several pathogenic barcodes that condition the different subtypes of cancer.

1. Introduction

The epistasis of multiple SNPs is a highly-risky contributor to common complex diseases (e.g. cancer) (Chakravarti et al., 2013; Flintoft, 2013). Genome-wide association study has been shifting its focus from single SNP susceptibility to epistasis effect to reveal the complex relationship between genetic variations and complex diseases. However, the genome-wide epistasis analysis is hampered by huge computational cost, especially for high-order epistasis (Mooney et al., 2012; Yang and Gu, 2013).

The failures of current studies of complex diseases (Yang et al., 2012; Zhu et al., 2012) (e.g. low reproducibility and difficulty of interpretation) may be attributed to several reasons. First, some important rare SNPs are subjectively ignored to simplify the mechanism of complex diseases, which result from high analysis cost. Second, given the moral and economic factors, the number of case samples is not enough to reveal the rule of complex diseases. Recently, lots of rare SNPs are identified and the sample size of association study rapidly grows. Therefore, an increasing demand exists for an efficient bioinformatics method to analyze large-scale data.

Approaches that find a synergic interaction between loci have an important function to predict, diagnose and treat common complex disease. After recognizing the pathogenic epistatic SNP combination, searching a risky genotype barcode (Chang et al., 2009; Daniels et al., 2008) is the key process of diagnosing a specific individual. In this study, a barcode is a particular set of alleles that can be used to amplify a signal of infectious organisms.

* Corresponding author. Tel.: +86 15243647444; fax: +86 731 88821417.
E-mail address: lx_hncs@163.com (X. Li).

http://dx.doi.org/10.1016/j.jtbi.2014.11.005
0022-5193/

Therefore, individuals carrying high-risk barcodes are potential patients.

Many approaches have been designed for generating SNP barcodes using intelligent algorithm such as DBPSO (Chang et al., 2009 and IGA (Yang et al., 2013) or exact algorithm such as branch and bound method IBBFS (Chuang et al., 2013). These methods have some advantages. but still possess several important limitations. For example, IBBFS can generate for an optimal solution, but its computational complexity is relative high. For BPSO and IGA, some important issues are needed to be addressed. First, these methods use odds ratio (OR) or maximum difference ratio to evaluate the susceptibility of a certain barcode. The results may be high false positive. Second, complex diseases are always caused by complex reasons. Therefore, boiling these diseases down to a specific barcode is unreasonable. For example, different types of risky barcodes exist, which produce different tumor subtypes. Third, artificial algorithms (e.g. genetic algorithm and particle swarm optimization) always have a certain degree of randomness, which confuses clinical researchers. This degree of randomness significantly reduces the confidence on bioinformatics methods. Lastly and most importantly, the resolution of IGA is local optima. The barcodes are relative long, which probably violates the mechanism of common diseases. Unfortunately, the DBPSO and IBBFS do not provide executable software or experimental dataset, so that our method is only compared with the latest approach IGA.

This study proposes an exact algorithm to generate SNP barcodes to solve the aforementioned problems. Our method also searches a synergic SNP combination with a maximum difference between the cases and controls. However, we only focus on these SNP combinations which contain a certain pattern in case samples. These combinations are significantly strict and reasonable and will ideally reduce the false positive. Compared with the latest barcode-generating method IGA, MDME is more stable and can generate barcodes with higher odds ratio and relative risk (RR).

2. Methods

As demonstrated in a series of recent publication (Chen et al., 2014; Liu et al., 2014) in response to the suggestion proposed in a comprehensive review (Chou, 2011), to establish a really useful analysis method for a biological or biomedical system, one should make the following procedures very clear: (i) preprocess a valid dataset to test the method or model; (ii) formulate the biological samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted or analyzed; (iii) introduce or develop a powerful algorithm (or engine) to operate the prediction or analysis; (iv) apply statistical measures to objectively evaluate the performance of the method or model; (v) establish a user-friendly web-server that is accessible to the public. Below, let us describe how to deal with these steps one-by-one.

2.1. Breast cancer dataset

In this study, the experiments are based on the breast cancer dataset organized by studies (Pharoah et al., 2007; Yang et al., 2013). A total of 10,000 samples containing 23 SNPs separately belong to 6 genes, namely, COMT, CYP19A1, ESR1, PGR, SHBG and STS. Several studies proved that these genes are associated with the occurrence and progression of breast cancer. For example, COMT encodes catechol-*O*-methyltransferase, which catalyzes inactivation; these metabolites may lead to metabolic redox cycling and production of free radicals, which result in DNA damage (Udler et al., 2009). Gabriel et al. (2013) conducted a candidate association study between PGR and breast cancer. The results show that the progesterone receptor coded by PGR is highly associated with the risk. The latest report on *Nature Genetics* also demonstrates that over two-thirds of breast cancer cases express estrogen receptor-α (encoded by ESR1) (Robinson et al., 2013; Toy et al., 2013).

2.2. Problem formulation

This section formulates the pathogenic barcode (*PB*) problem, which is defined as follows:

Definition (*PB*): Give a set of N SNPs $\{S_1,\ldots,S_N\}$ and a set of M individuals $\{P1,\ldots,PM\}$ containing the cases and controls. Let $Bi=(E,Gm)$ denote the ith barcode consisting of two factors: E is the set of epistatic SNPs and Gm is the genotype of the mth individual in these epistatic SNPs. G is the set of all possible genotype combinations of E. *PB* is a specific genotype that determines whether an individual is a case.

We assume that a pathogenic genotype barcode associated with a certain disease will exhibit a pattern in case samples and show significant dissimilarity between cases and controls. We initially design a maximum dissimilarity–minimum entropy (MDME) measure to evaluate barcodes. Thus, *PB*s should meet the following two conditions.

2.3. Minimum entropy

According to our assumption, the epistatic SNP E of a candidate *PB* must have a relative consistent pattern to cause disease in case samples. However, describing the pattern directly on a limited size of samples is difficult. Thus, we indirectly profile the pattern by measuring the uncertainty of E. Low uncertainty indicates that a consistent pattern functions on the cases. We use entropy theory to measure uncertainty. Information theory has been widely used to measure synergetic effects among multiple loci (Hu et al., 2013). Shannon entropy quantifies the amount of uncertainty of variables. Thus, the first condition is defined by

$$\text{ME}: \quad \text{Min}(HE) \tag{1}$$

with

$$HE = -\sum_{i=1}^{h} pi \log\ pi \tag{2}$$

where h is the total number of different types of barcodes, and pi is the frequency of the ith barcode in cases.

2.4. Maximum dissimilarity

The second condition is maximum dissimilarity (MD) and is defined by the following formulas.

$$\text{MD}: \quad \text{Max}(D_{\text{PB}}) \tag{3}$$

with

$$D_{\text{PB}} = F_{\text{case}} - F_{\text{control}} \tag{4}$$

where F_{case} and F_{control} are the frequencies of *PB* in cases and controls, respectively. The larger value indicates that the barcode has more significant distinction between the cases and controls. If the value is zero, the barcode has no contribution to distinguish the case from the control. If the value is negative, the individual carrying this barcode more likely tends to be normal.

Note that a barcode may satisfy the second condition but violate the first condition. We suppose that this barcode has a high possibility of false positive. Thus, we formulate the *PB* optimization generating problem as

$$\min_{E}\ HE$$
$$\text{s.t.}\ \operatorname*{argmax}_{\text{PB}} D_{PB} \quad (PB \in G) \tag{5}$$