An informative SNP selection method based on artificial neural network

Zejun Li[1,2*], Lijun Cai[1], Min Chen[2], Lijun Zeng[2]

1 College of Information Science and Engineering, Hunan University, Changsha, Hunan, 410082, China
2 School of Computer and Information Science, Hunan Institute of Technology, Hengyang , Hunan,412002,China

Abstract
Because of low prediction accuracy and high time complexity of information of the current SNP selection methods, we design an information SNP selection method based on artificial neural network, which uses a two-stage design framework to optimize information SNP selection process.The first stage aims to eliminate redundant role sites, so we construct a heuristic function by a new multiple loci linkage disequilibrium measure to improve the ant colony algorithm performance. In the second stage, the neural network is applied to reconstruct alleles of non-informative SNP. In order to achieve the purpose of improving the prediction accuracy of the reconstruction process, the greedy algorithm is proposed to search the best combination of informative SNPs which has lower noise and redundancy. The experimental results show that this method has some improvements in improving prediction accuracy and running time.
Keywords: ANT COLONY ALGORITHM; LINKAGE DISEQUILIBRIUM; ARTIFICIAL NEURAL NETWORK; INOFORMATIVE SNP

1. Introduction

In the genome, polymorphism caused by a single nucleotide mutation is called Single Nucleotide Polymorphism, SNP. The study found that among different samples, a small amount of SNP loci may represent the majority of bits of information points. The small number of SNP loci are called Information SNP (Informative SNP). This is due to the widespread presence of linkage disequilibrium in genome. If the two sites are completely linked, then it is known genotype a site, you can use bioinformatics methods to restore the other genotypes, without the need to use chemical and biological means to genotyping experiments. Although the biochemical experiments can draw more accurate genotyping results, but when for the huge number of sites, the reliability of the process continues to reduce and costs are significantly. Therefore, the use of data mining method selection information SNP, has become one of the key contents of the current bioinformatics research [1] [2] [3].

So far, there have been a lot of intelligent optimization algorithms (such as genetic algorithms, particle swarm algorithm [4] [5] [6]), data mining (support vector machine [6], clustering [7]) is used in SNP information selection for the study. Each method has been improved for a specific problem.

In order to effectively handle two allele and more alleles, Ting [5] proposed a modified Bayesian network (Bayesian network) model. The biggest advantage of this method is that the input genotype data can be processed directly without Haplotyping process. Chuang [6] combined with particle swarm algorithm

for large space optimization capabilities directly to the prediction accuracy as an optimization goal, proposed an improved algorithm.

2012 Zeng Jin Ping [8] proposed information SNP selection method based on intelligent algorithm. The method includes two stages of filtering and selection. Filtration stage uses two linkage between the degrees of value judgment as a redundant site basis. The selection phase uses genetic algorithms to optimize the accuracy of SNP candidate information collection.

Li [9] 2013 presents a message SNP selection method based on the maximum correlation minimum redundancy. This method regards information SNP as a SNP subset of non-information SNP with the highest correlation, between the interior and with minimum redundancy. Since this method avoids the wound sample reconstruction, so they can select information on a large scale SNP datasets.

Although the above methods have certain information SNP advantages, they still need to further enhance performance. For example, the above methods have higher spending less time complexity. Such as improved genetic algorithms proposed by Jin Ping, each individual will need to use the leave-one-out cross validation, conduct individual performance evaluation, so that the time complexity is too high.

Further analysis showed that the time complexity is too high is due to that the above methods repeatedly reconstruct the accuracy wound evaluate of information on each candidate SNP solution process, thus they greatly wasted running time.

2. The Method This Paper Proposed

In order to overcome the shortcomings that the time complexity is too high, and at the same time guarantee the accuracy of information on the reconstruction of the SNP, The proposed method still uses a two-stage technology framework. Two phases are filtered phase and selection phase. The main difference for the previous stage of filtration process is that in this stage, The proposed method does not use the two-point chain metrics as a redundant measure, but to build a multi-site correlation measure by information entropy theory. This measure not only directly measure the degree of redundancy at multiple sites, and the computer is relatively simple, and thus has high efficiency than a wound selecting method. In order to guarantee the degree of redundant information SNP as small as possible, in the second stage, we use a greedy algorithm to further optimize the results of the first phase. The stage regards reconstruct accurate as the optimization target. The following methods are described in detail herein.

2.1. Noise SNP Filtering Based on Ant Colony Algorithm

In practical applications, the information SNP required to directly assess correlation between multiple sites, and the use of the two-point evaluation is difficult to directly measure. Therefore, Liu et al. [10] used information entropy theory to construct multi-site chain metrics ER method. While this measure has certain partiality, which with the increase of the number of SNP subsets of information, the ER value is difficult to accurately reflect changes in the degree of redundancy. Therefore, this paper proposes a new measurement method. Assumed that the sample data set contains m different haplotypes, and each haplotype sample contains n sites, then this article multisite metrics as shown in Equation 1.

$$R = \frac{\sum_{i=1}^{m} p(\mathbf{x}i) \log p(\mathbf{x}i)}{\sum_{j=1}^{n} p(xij)} \quad (1)$$

Wherein the vector $\mathbf{x}i$ represents a haplotype, and $p(\mathbf{x}i)$ indicates the frequency of the haplotype present in the sample, $p(xij)$ indicates the frequency of allele on the j-th site of the i-th haplotype.

(1) Path Selection function

In epistasis analysis of SNP data, the biggest challenge facing is explosive of the SNP combination space. And for all the SNP combinations exhaustive analysis, it is NP-hard problem. To reduce the complexity of epistasis analysis time, an effective strategy is pretreat to all the initial SNP, through some simple and effective and accurate filter, thus reducing the computational complexity of the follow-up sample reconstruction phase. The main purpose of this article in the framework of the second phase of the first stage is in this. Ant colony algorithm is an excellent combination and optimization algorithms. Thus, the stage uses ant colony algorithm to search of candidate information SNP subset with low redundancy. As one important function of the ant colony algorithm, the path selection function as shown in Equation 2.

$$p_i^k(t) = \begin{cases} \dfrac{[\tau i]^{\alpha} \cdot [\eta i]^{\beta}}{\sum_{i \in R} [\tau i]^{\alpha} \cdot [\eta i]^{\beta}} & i \in R \\ 0 & \text{else} \end{cases} \quad (2)$$

In Equation 2, α is the weight of pheromone residual, and β is the weight of inspire information, R

represents the current iteration unselected SNP set. The main difference with the traditional traveling salesman problem is that the traditional TSP problem is a path as short as possible, and in order to avoid important information SNP were excluded, the candidate subset of the number of non-redundant site is the bigger the better for this article ant colony algorithm optimization goals. On the basis of ensuring the degree of redundancy among all important SNP is less, then these SNP candidate information is as input data, passing to the next stage.

(2) Heuristic and pheromone update function

Linkage disequilibrium metrics between two loci although cannot directly measure for the chain between multiple sites, but because the measurement method is easy to calculate, and it is intuitively reflect the interdependence between the two sites. In fact, this metric can better represent relationships between the sites. The higher the dependence between the sites, so a site is able to more accurately represent another site.To some extent, it shows superior predictability between the two. Therefore, this article will use the metric r^2 as heuristic information to improve the convergence rate.

$$\eta i = \frac{\sum_{j=1}^{n} r_{ij}^2}{n} \tag{3}$$

Where r_{ij}^2 represents the degree of linkage between sites i and j sites, the larger the value, the greater the correlation between the two sites. n indicates the number of all the SNP. In Equation 3, molecules represent a SNP loci i and all the other SNP loci LD value. A higher number indicates the degree of association between them is stronger, then the show is more conducive to construct a subset of SNP candidate information, and then dividing by n to reduce the molecule. In this article, its value is directly as heuristic information. The key is that the ant colony as a whole choose SNP concentration of pheromone on each SNP. So after each iteration, according to a number of linkage metrics to evaluate the performance of the current candidates for the subset. Then depending on the metric change the pheromone of the sites, as shown in Equation 4.

$$\Delta\tau^k i(t) = \begin{cases} \dfrac{L^k(t)}{\varepsilon} & i \in T^k(t) \\ 0 \quad \text{otherwise} \end{cases} \tag{4}$$

Where ε represents a subset of the candidate chain metric. As can be seen, the greater its value indicates greater redundancy therein, so that the candidate subset is worse. A represents the size of the candidate subset. As this article hoping to avoid significant sites are removed, therefore, hope that the largest possible of the candidate subset.

2.2. Greedy Strategy to Constructe SNP Subset of Candidate Information

Suppose the current candidate information SNP subset is Tm. The subset of features is with m features. Then each iteration of m features removed from a worst candidate information SNP, to maximize satisfy Equation 5.

$$\max[P_{m-1} - P_m] \tag{5}$$

Wherein the sample Pm-1 indicates the reconstruction accuracy of corresponding subset.

2.3. information SNP Selection based on Artificial Neural Networks

The following uses pseudo-code to describes the artificial neural networks as a learning model and gradually optimized feature selection until to select the optimal t information SNP subset. In the back-propagation neural network learning algorithm, initially, the network on each edge is set a right value of a smaller non-zero value. With the deepening of the learning process, each subsequent iteration, using supervised learning model, constantly calculates the difference between the target and the predicted value, and then continue to modify the weights between nodes in the network based on the degree of difference until the accuracy of this learning model to reach a satisfactory level. Adjust the weights is a gradual optimization process, usually using a gradient method to modify the weights. End conditions of neural network algorithm can be set a fixed number of iterations or prediction accuracy. This article sets a maximum number of iterations 50 as the end condition.

Information SNP selection algorithm based on Greedy algorithm:

Input: the first stage reserved t SNP subset of candidate information $T = \{T1, T2, ..., T t\}$

Output: Information SNP

Begin:

For i=1:t // Select a candidate subset

For j=1:| | // do exploratory culling for all SNP of the subset

For k=1: max_iteration // Neural network algorithm maximum number of iterations

According to the SNP subset number to initialize the network;

From front to back calculate the nerve cell layer by layer, obtaine the output layer values and compared with the actual value;

Correction weights based on the error value;

End For

According to the formula 5 Select an optimal Tm-1;

End For

Choose a minimal subset of Tm-1 from all over the optimal solution;

End For

End

3. Data Collection and Evaluation Index

For a more objective evaluation of the proposed method, the paper do compared experiment for a plurality of real data sets were with other methods in a number of indicators on (time complexity and reconstruction accuracy) were compared. Real data set used in this paper is haplotype data from the international human genome. In order to more fully evaluate the improved performance of the method, a plurality of haplotypes were used to test data sets, as shown in Table 1.

Table 1. Dataset properties

Data name	SNP number
TRPM8	101
ENm013	360
ENr112	411
5q31	55

In order to avoid peak phenomenon occurred during the reconstruction, we use cross method to reconstruction method accuracy, the formula is as follows:

$$\text{Acci}=1-\frac{|hi-hi'|}{N} \quad (6)$$

$$AvgAcc=\frac{\sum_{i}^{|O|}Acci}{|O|} \quad (7)$$

Where N represents the total number of samples genotype, Acci represents an average accuracy rate of a non-informative SNP i-th point is reconstructed information SNP. In Equation 6, hi represents the actual value of the haplotype genotypes and hi' represents neural network algorithm predicted gene value. Because it is two alleles, so that only two kinds of value 0,1. So if two prediction values and the actual values are equal, then 0, 1 otherwise.

4. Simulation Experiment and Results Analysis

In order to obtain a more accurate evaluation of the performance of the proposed method, the paper will be compared with experimental MCMR, GA [8] and MLR methods.

In 2007, He proposed MLR (multiple linear regression), the multiple linear regression model is as haplotype data learning model to reconstruct the non-information SNP.

(1) Reconstruction accuracy

Reconstruction is an important measure of the accuracy of the information SNP evaluation. The higher the accuracy of the information indicates that the SNP can represent the better, the better the efficiency of the subsequent association studies it is able to maintain. Figure 1-4 are comparative experiments on four datasets.

As can be seen from the above diagram, the performance advantages of this method are obvious. The GA method has better performance than other methods. GA combination of this method and are used in the optimization phase intelligent algorithm. But the proposed method is better than the GA method performance. This measure not only the amount of information can better save information SNP, while selection stage can accurately identify the real information SNP. While the accuracy of this method advantages over GA only about 1 percent higher, but can be seen from the subsequent computational complexity, GA consumes more running time.

(2) Running time

The time complexity of the evaluation information is as SNP selection method performance a secondary index. When the reconstruction accuracy of the different methods of the same, the running time of less method better performance. Figure 5 is running time in different ways on an ordinary PC (2.80GHz and 2GB memory). From the figure it can be found in the larger data set, the processing time for this dataset is more. Running time of this method is the lowest of four methods, indicating that the proposed method first stage of filtering policy is designed to be reasonable and effective. Running time of this method was significantly less than MCMR and MLR methods.

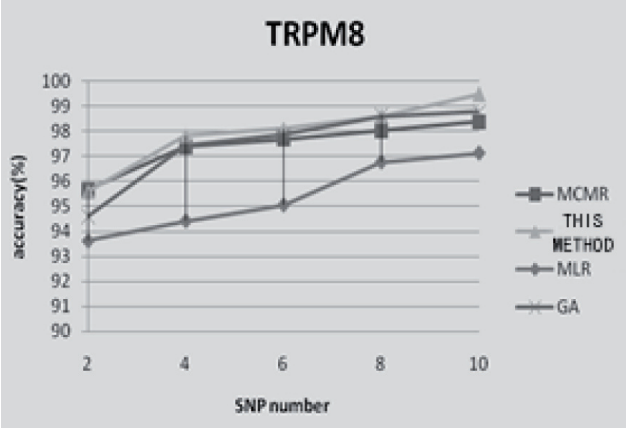

Figure 1. TRPM8 Comparison of the accuracy of the reconstruction

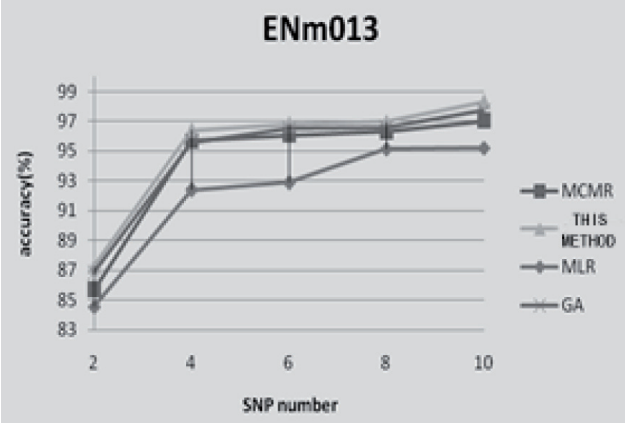

Figure 2. ENm013 Comparison of the accuracy of the reconstruction

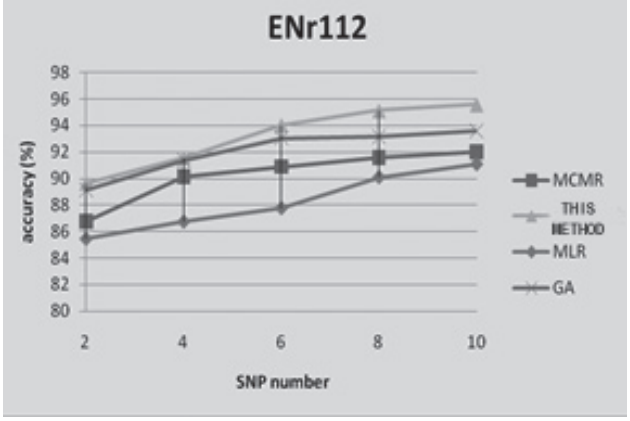

Figure 3. ENr112 Comparison of the accuracy of the reconstruction

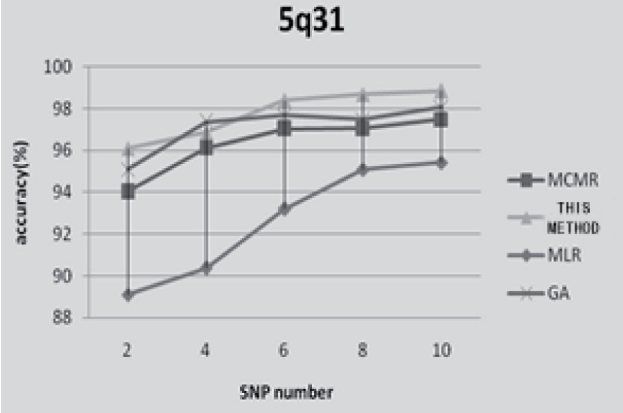

Figure 4. 5q31 Comparison of the accuracy of the reconstruction

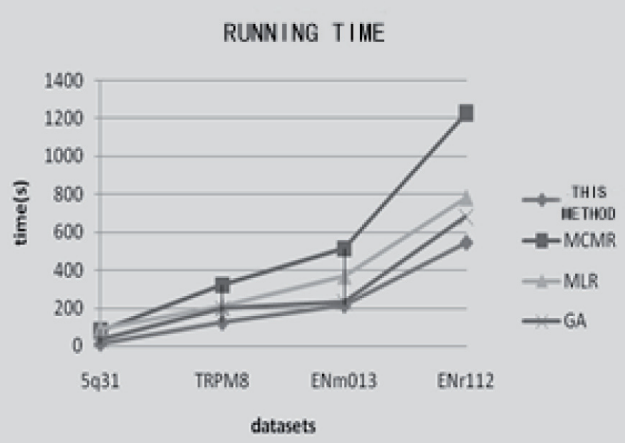

Figure 5. Run time comparison

Conclusions

For lack of available information on the presence of SNP selection methods, this paper presents a framework for selecting information SNP based on artificial neural network learning model. The framework has two phases, the main purpose of the first phase is to filter. First, this phase designs a number of points associated with metrics information entropy theory, and then use ant colony algorithm to search the optimal subset of candidate. While the second stage is to use neural networks on a candidate subset of non-informative SNP genotype reconstruction sites, to identify the real information SNP site. Finally, on multiple datasets we demonstrate the effectiveness of this method.

Acknowledgements

This work is supported by the National Natural Science Fund Project (No. 61472127) and the National Natural Science Fund of Hunan, China (No. 13JJ9026).

References

1. Liao B, Li X, Zhu W, et al. A Novel Method to Select Informative SNPs and Their Application in Genetic Association Studies. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2012, 5(9): 1529-1534.
2. Kimmel G, Shamir R. GERBIL. Genotype resolution and block identification using likelihood. PNAS, 2004, 102(1): 158-162
3. Flintoft L. Complex disease: A SNP for disease prognosis[J]. London: Nature Reviews Genetics, 2013, 14(11): 746-746.
4. Bo Liao, Xiong Li, Wen Zhu et al. Multiple Ant Colony Algorithm Method for Selecting tag SNPs[J]. The Netherlands: Journal of Biomedical Informatics,2012,45(3):931–937.
5. Ting C K., Lin W T, Huang Y T. Multi-objective tag SNPs selection using evolutionary algorithms[J]. London: Bioinformatics, 2010, 26(5): 1446-1452.

6. Chuang L Y, Yang C S, Ho C H.et al. Tag SNP Selection Using Particle Swarm Optimization[J]. San Francisco: Biotechnology Progress, 2010, 26(2): 580-588.
7. Bo Liao, Xiong Li, Wen Zhu, et al. A Novel Method to Select Informative SNPs and Their Application in Genetic Association Studies[J]. New York: IEEE/ACM Transactions on Computational Biology and Bioinformatics: 2012, 5(9):1529-1534.
8. Chuang L Y, Hou Y J, Yang C H. Fuzzy guided BPSO method for haplotype tag SNP selection. FUZZ-IEEE, 2009, 20-24.
9. Li X, Liao B, Cai L, et al. Informative SNPs selection based on two-locus and multilocus linkage disequilibrium: Criteria of Max-Correlation and Min-Redundancy[J]. 2013, 10(3):688-95.
10. Liu Z. and Lin S. Multilocus LD measure and tagging SNP selection with generalized mutual information[J]. Genet Epidemiol, 2005, 29(4): 353-364.
11. Hudson R R. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics, 2012, 18: 337-338.
12. J.W. He, and A. Zelikovsky. Informative SNP Selection Methods Based on SNP Prediction[J]. New York:IEEE TRANSACTIONS ON NANOBIOSCIENCE, 2007, 6(1):60-67.
13. Halldorsson B V. Optimal haplotype block-free selection of tagging SNPs for genome-wide association studies. Genome Res, 2004, 14(3): 1633-1640
14. Ting C K, Lin W T, Huang Y T. Multi-objective tag SNPs selection using evolutionary algorithms. Bioinformatics, 2010, 26(5): 1446-1452.
15. Datta S, Satten G A. A Signed-Rank Test for Clustered Data. Biometrics, 2008, 64(2): 501-507
16. Mahdevar G, Zahiri J, Sadeghi M, et al. Tag SNP selection via a genetic algorithm. Journal of Biomedical informatics, 2010, 43(5): 800-804.
17. Chuang L Y, Yang C S, Ho C H, et al. Tag SNP Selection Using Particle Swarm Optimization. Biotechnology Progress, 2010, 26(2): 580-588.