Journal of Computer and Communications, 2014, 2, 1-8
Published Online July 2014 in SciRes. http://www.scirp.org/journal/jcc
http://dx.doi.org/10.4236/jcc.2014.29001

Scientific Research

The Application of Cluster Analysis in Type II Diabetes Genome Association Study

Hankun Hu[1], Weidong Mao[2]

[1]Department of Pharmacy, Zhongnan Hospital, Wuhan University, Wuhan, China
[2]School of Science & Technology, Georgia Gwinnett College, Lawrenceville, USA
Email: hankunhu@hotmail.com, wmao@ggc.edu

Received February 2014

Abstract

Genetic diseases, such as Type II diabetes, are caused by a combination of environmental factors and mutations in multiple genes. Patients who have been diagnosed with such diseases cannot easily be treated. However, many diseases can be avoided if people at high risk change their living style, one example is their diet. Genome association study has been used to identify the risk factor of genetic disease. With the development of DNA microarray technique, it is possible to access the human genetic information related to specific diseases. This paper uses a combinatorial method to analyze the genetic case-control data for Type II diabetes. A distance based cluster method has been applied to publicly available genotype data on Type II diabetes for epidemiological study and achieved a high accurate result.

Keywords

Genetic Disease, Genome Association Study, Cluster Algorithm

1. Introduction

Type 2 diabetes is the most common form of diabetes. In type 2 diabetes, either the body does not produce enough insulin or the cells ignore the insulin. Insulin is necessary for the body to be able to use glucose for energy. When you eat food, the body breaks down all of the sugars and starches into glucose, which is the basic fuel for the cells in the body. Insulin takes the sugar from the blood into the cells. When glucose builds up in the blood instead of going into cells, it can cause two problems: 1) Right away, your cells may be starved for energy. 2) Over time, high blood glucose levels may hurt your eyes, kidneys, nerves or heart. While diabetes occurs in people of all ages and races, some groups have a higher risk for developing type 2 diabetes than others. Research shows that the type 2 diabetes is caused by a complicated interplay of genes, environment, insulin abnormalities, increased glucose production in the liver, increased fat breakdown, and possibly defective hormonal secretions in the intestine. The recent dramatic increase indicates that lifestyle factors (obesity and sedentary lifestyle) may be particularly important in triggering the genetic elements that cause this type of diabetes [1].

Although the Type 2 diabetes cannot easily be treated, it can be avoided if people at high risk change their living style, such as their diet. But how can we tell the susceptibility of people to the disease before symptoms are found and help them make informed decisions about their health? With the development of DNA microarray

How to cite this paper: Hu, H.K. and Mao, W.D. (2014) The Application of Cluster Analysis in Type II Diabetes Genome Association Study. *Journal of Computer and Communications*, **2**, 1-8. http://dx.doi.org/10.4236/jcc.2014.29001

technique, it is possible to access the human genetic information related to specific diseases. Assessing the association between DNA variants and disease has been used widely to identify regions of the genome and candidate genes that contribute to disease [2].

99.9% of one individual's DNA sequences are identical to that of another person. Over 80% of this 0.1% difference will be Single Nucleotide Polymorphisms (SNP) and they promise to significantly advance our ability to understand and treat human disease. A SNP is a single base substitution of one nucleotide with another. Each individual has many single nucleotide polymorphisms that together create a unique DNA pattern for that person. It is important to study SNPs because they represent genetic differences among human beings. Genome-wide association studies require knowledge about common genetic variations and the ability to genotype a sufficiently comprehensive set of variants in a large patient sample [3]. High-throughput SNP genotyping technologies make massive genotype data, with a large number of individuals, publicly available. Accessibility of genetic data makes genome-wide association studies for complex diseases possible.

Success stories when dealing with diseases caused by a single SNP or gene, sometimes called monogenic diseases have been reported [4]. However, most complex diseases, such as psychiatric disorders, are characterized by a non-mendelian, multifactorial genetic contribution with a number of susceptible genes interacting with each other [5]. A fundamental issue in the analysis of SNP data is to define the unit of genetic function that influences disease risk. Is it a single SNP, a regulatory motif, an encoded protein subunit, a combination of SNPs in a combination of genes, an interacting protein complex, a metabolic or a physiological pathway [6]. In general, it may be impossible to associate a single SNP or gene with a disease because a disease may be caused by completely different modifications of alternative pathways, and each gene only makes a small contribution. This makes the identification of genetic factors difficult. Multi-SNP interaction analysis is more reliable but it is computationally infeasible. An exhaustive search among multi-SNP combination is computationally infeasible even for a small number of SNPs. Furthermore, there are no reliable tools applicable to large genome ranges that could rule out or confirm association with a disease.

It's important to search for informative SNPs among a huge number of SNPs. These informative SNPs are assumed to be associated with genetic diseases. Tag SNPs generated by the multiple linear regression based method [7] are good informative SNPs, but they are reconstruction-oriented instead of disease-oriented. Although the combinatorial search method [8] for finding disease-associated multi-SNP combinations has a better result, the exhaustive search is still very slow.

Comparative genomics is the study of the relationship of genome structure and function across different biological species or strains. Comparative genomics is an attempt to take advantage of the information provided by the signatures of selection to understand the function and evolutionary processes that act on genomes. In order to get promising results, we need to compare human chromosomes or genes with those of other species which involves a large amount of information and computation. In epidemiological study, instead of comparing human chromosomes with other species, we just compare the genomic data from one individual to other individuals to get useful information.

The distance-based algorithm and cluster analysis have been used to solve the classification problem. In this algorithm, each item that is mapped to the same class may be thought of as more similar to the other items in that class than it is to the items found in other classes. Therefore, similarity measures may be used to identify the "alikeness" of different items in the database [9]. In our algorithm, the similarity is measured by the distance between the item and some neighbor clusters whose class label are previously known. The algorithm can be applied in our case-control study to predict an individual's susceptibility to type II diabetes by comparing its genetic data with that of other individuals to find the similarity.

In this paper, we first address the disease susceptibility prediction problem [10]-[12]. This problem is to assess accumulated information targeted to predicting genotype susceptibility to complex diseases with significantly high accuracy and statistical power. Next, we introduce the cluster-based distance algorithm and its application in disease susceptibility prediction problem. We will also introduce the case tagging algorithm which is used to reduce the size of data and improve prediction results. The proposed method is applied to a publicly available data for Type II diabetes.

2. Disease Susceptibility Prediction

In this section we first describe the genetic model then we formulate the disease susceptibility prediction prob-

lem.

2.1. Genetic Model

A SNP is a single base substitution of one nucleotide with another. Both substitutions have to be observed in the general population at a frequency greater than 1%. If an individual has a sequence of GAACCT, while another individual has sequence of GAGCCT, the polymorphism or the SNP is a A/G. Each individual has many single nucleotide polymorphisms that together create a unique DNA pattern for that person.

Recent work has suggested that SNP's in human population are not inherited independently; rather, sets of adjacent SNP's are present on alleles in a block pattern, so called *haplotype*. Many haplotype blocks in human have been transmitted through many generations without recombination. This means although a block may contain many SNP's, it takes only a few SNP's to identify or tag each haplotype in the block. A genome-wide haplotype would comprise half of a diploid genome, including one allele from each allelic gene pair. The *genotype* is the descriptor of the genome which is the set of physical DNA molecules inherited from the organism's parents. A pair of haplotype consist a single genotype.

SNP's are bi-allelic and can be referred as 0 if it's a majority and 1, otherwise. If both haplotypes are the same allele, then the corresponding genotype is homogeneous, can be represented as 0 or 1. If the two haplotypes are different, then the genotype is represented as 2.

The case-control sample populations consist of N individuals that are represented in genotype with M SNPs. Each SNP attains one of the three values 0, 1, or 2. The sample G is an (0, 1, 2)-valued $N \times M$ matrix, where each row corresponds to an individual, which is a sequence of 0, 1 and 2, each column corresponds to a SNP.

The data set we use in this paper is coming from the Wellcome Trust Case Control Consortium (WTCCC) [13]. The data has1999 individuals in case and 1999 individuals in control for type II diabetes, respectively. 44,000 SNPs of each chromosome are derived from individuals. Due to the computation difficulty, we take 100 individuals from case and 100 individuals from control. We also reduce the number of SNP to 100 by randomly choose from the 44,000 SNPs.

2.2. Disease Susceptibility Prediction Problem

The disease susceptibility prediction problem can be formulated as follows:

Data sets have n genotypes. The input for a prediction algorithm includes:

(G1) Training genotype set $G = (g_i \mid i = 1 \ldots n)$,

(G2) Disease status $s(g_i) \in \{0,1\}$, indicating if $g_i \mid i = 1 \ldots n$ is in case (1) or in control (0), and

(G3) Testing genotype g_t without any disease status.

We will refer to the parts (G1-G2) of the input as the training set and to the part (G3) as the test set. The output of prediction algorithms is the disease status of the genotype $s(g_t)$.

2.3. Cluster-Based Distance Algorithm

Due to the computational difficulty on finding SNPs associated with diseases, we try to explore another way to find out an individual has the susceptibility to the disease or not. We use the ideal of comparative genomics, but we do not compare human's chromosome with other species, we compare one individual with other patients' or healthy individual's chromosome. In other words, we want to find an individual's disease status from its similarity with other individuals whose disease status is already known.

In the distance-based algorithm, each item that is mapped to the same class may be thought of as more similar to the other items in that class than it is to the items found in other classes. Therefore, similarity measures may be used to identify the "alikeness" of different items in the database. The idea of similarity measures can be abstracted and applied to more general classification problems. The difficulty lies in how the similarity measures are defined and applied to the items.

As we described in above section, genotypes are sequences of 0, 1, and 2, this make it easy to find out the similarity among sequences by computing their hamming distance. The hamming distance between two strings (in our case, two genotypes, each represents an individual) of equal length is the number of positions for which the corresponding symbols are different. For example, the hamming distance between genotype 1 (01021011) and genotype 2 (01021011) is 0, but the hamming distance between genotype 3 (**2**1021**2**1**0**) and genotype 4

(**0**1021**0**11) is 3.

For the training data set, we build graph-based clusters for each class. In other words, we generate *N* clusters in case class and *N* clusters in control class given the threshold *N*. First we generate the graph G_{case} and $G_{control}$ based on the hamming distance among individual genotypes in case class and in control class, respectively. The Kruskal's algorithm is used to find the minimum spanning tree (MST) G_{case} and $G_{control}$. To generate *N* clusters for G_{case} we need to remove the largest *N*-1 edges. As a result, the inter-cluster distance is maximized and the intra-cluster distance is minimized.

For these $2 \times N$ clusters, we find the centroid for each cluster. The centroid is the genotype that has the minimum distance with all other genotype in the same cluster. To begin, an adjacency matrix is created for each individual cluster. Next, the vector is then iterated over adding up each row of the vector to return a total summation of each individuals distance in relation to all other individuals in the vector. The genotype that has the smallest total distance with all other genotype will be selected as the centroid for that cluster. For example in **Table 1**, genotype 2 is selected as the centroid of the cluster with 7 genotypes because the total number of distance with other 6 genotypes is the smallest.

K nearest neighbors (KNN) is used as the classification scheme based on the use of distance measures. The KNN technique assumes that the entire training set includes not only the data in the set but also the desired classification for each item. When a classification is to be made for a new item, its distance to each item in the training set must be determined. Only the *K* closest entries in the training set are considered further. The new item is then placed in the class that contains the most items from this set of *K* closest items. In our case, the hamming distance of the testing genotype to each centroid in the training set (including both case and control set) will be computed, then we find out the *K* closest genotypes which have smaller hamming distance than others. From the set of *K* centroid, if most of them are coming from the case group, then the testing genotype will be classified as case, otherwise, it will be classified as control. Obviously, it will be better if *K* is an odd number. The algorithm is illustrated in **Figure 1** when it is applied in the disease susceptibility prediction problem. **Figure 2** shows how to classify the testing item when *N* is 4 and *K* is 3. In this example, we generate 4 Clusters for case and control, respectively. 8 centroids were found for these 8 clusters. The distance between the testing sample and the 8 centroids is measured and the 3 centroids with the shortest distance are selected for voting. One out of the three centroids is in case group, while the other two are in control group, and we can classify the testing sample as control.

3. Tag SNP

Constructing a complete human haplotype map is helpful when associating complex diseases with their related SNPs. Unfortunately, the number of SNPs is very large and it is costly to sequence many individuals. Therefore, it is desirable to reduce the number of SNPs that should be sequenced to a small number of informative representatives called *tag SNPs*. On the other hand, these important tag SNPs or the subset of genotype/haplotype probably are responsible for diseases. The techniques we described above using all or a random set of SNPs yielded results that were not decisive enough to make any assertion as to the classification of a test sample. As a result, a new method of generating the SNPs to be used was developed in an effort to obtain more accurate results and to improve the quality of the classification result.

Table 1. Centroid generation.

	1	2	3	4	5	6	7	Total Distance
1	0	3	4	16	16	15	15	69
2	3	0	5	4	13	12	12	49
3	4	5	0	12	18	11	13	63
4	16	4	12	0	10	7	7	56
5	16	13	18	10	0	11	8	76
6	15	12	11	7	11	0	4	60
7	15	12	13	7	8	4	0	59

Input: Training genotype set $G = (g_i \mid i = 1...n)$,
Case class C_{case},
Control class $C_{control}$,
Disease status $s(g_i) \in \{0,1\}$,
$g_i = 1$ if it is in class case,
$g_i = 0$ if it is in class control,
Testing genotype g_t
Treshhold *N*, *K*.

Case class $C_{case} = \Phi$
Control class $C_{control} = \Phi$
For i = 1 to n
if $s(g_i) = 1$ then
$C_{case} = C_{case} \cup \{g_i\}$
else
$C_{control} = C_{control} \cup \{g_i\}$
For C_{case} , do
Generate graph G_{case}, based on hamming distance
Find minimum spanning tree MST_{case} for G_{case}
Remove *N* -1 largest edges from to create MST_{case}
to create *N* clusters in case class
Find centroid for each cluster
For $C_{control}$, do
Generate graph $G_{control}$, based on hamming distance
Find minimum spanning tree $MST_{control}$ for $G_{control}$
Remove *N* - 1 largest edges from to create $MST_{control}$
to create *N* clusters in case class
Find centroid for each cluster
Find the distance between g_t and the centroid of each cluster
Find the *K* shortest distance and the corresponding class of the *K* centroid

Output: s (g_t) is classified by majority voting by the class of the *K* centroid.

Figure 1. Cluster-based distance algorithm.

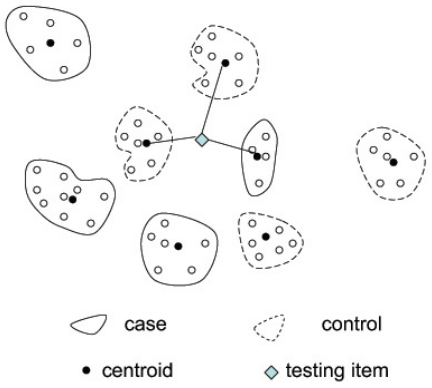

Figure 2. Classify the testing item, $N = 4$, $K = 3$.

"Tagging" is used to generate the SNP subset that will be used throughout the program. This method seeks to strip out data that is common between the Controls and Cases in an effort to reduce data size, and increase the probability that the correct data is being sampled and analyzed. With this method, the algorithm iterates through each position of SNP in both Cases and control. For each group, we find the mode of the position. If the mode at this position from case is different from the mode at the same position from control, then the SNP is added to the set of tag SNPs. If the modes are the same, we compare the frequency of the mode in case and control. If the mode frequency from case is greater than our threshold and the mode frequency from control is smaller than the threshold, or if the mode frequency from case is smaller than our threshold and the mode frequency from control is greater than the threshold, we also added it to the set of tag SNPs. The tagging algorithm is illustrated in **Figure 3**.

Input: Case class C_{case},
Control class $C_{control}$,
Genotype length m,
Treshhold t,
Tag SNP set TAG

$TAG = \Phi$
For SNP_i, $i = 1...m$
Find the mode of C_{case} $CaseM_i$
Find the frequency of the mode $CaseF_i$
Find the mode of $C_{control}$ $ControlM_i$
Find the frequency of the mode $ControlF_i$
if ($CaseM_i \neq ControlM_i$)
$TAG = TAG \cup \{i\}$
else if ($CaseF_i \geq t \&\& ControlF_i \leq t$)
$TAG = TAG \cup \{i\}$
else if ($CaseF_i \leq t \&\& ControlF_i \geq t$)
$TAG = TAG \cup \{i\}$

Output: Tag SNP set TAG

Figure 3. Tagging algorithm.

4. Results & Discussion

In this section we first introduce the measures to evaluate the prediction quality, after that is our experiment results and discussions.

4.1. Measures of Prediction Quality

To measure the quality of prediction methods, we need to measure the deviation between the true disease status and the result of predicted susceptibility, which can be regarded as measurement error. We will present the basic measures used in epidemiology to quantify accuracy of our methods.

The basic measures are:

Sensitivity: the proportion of persons who have the disease who are correctly identified as cases.

Specificity: the proportion of people who do not have the disease who are correctly classified as controls.

The definitions of these two measures of validity are illustrated in **Table 2**.

In this table:

a = True positive, people with the disease who test positive

b = False positive, people without the disease who test positive

c = False negative, people with the disease who test negative

d = True negative, people without the disease who test negative

From the table, we can compute sensitivity (accuracy in classification of case), Specificity (accuracy in classification of controls) and Accuracy:

$$\text{sensitivity} = a/(a+c)$$

$$\text{specificity} = d/(b+d)$$

$$\text{Accuracy} = (a+d)/(a+b+c+d)$$

Sensitivity is the ability to correctly detect a disease. Specificity is the ability to avoid calling normal as disease. Accuracy is the percent of the population that is correctly predicted.

We use K-fold cross validation method to measure the quality of the algorithm. In the K-fold cross validation, the data set is divided into K subsets, and the holdout method is repeated K times. Each time, one of the K subsets is used as the test set and the other K-1 subsets are put together to form a training set. In our experiment, we use 5-fold cross validation.

4.2. Results

Table 3 is the experiment result. In this table, we compare the result when K is 1, 3, 5, 7 and the number of cluster N is 2, 6, 10, 14, 18 and 22. The best result is as high as 100% for sensitivity, 100% for specificity, and

Table 2. Classification contingency table.

		True Status	
		+	-
Classified Status	+	a	b
	-	c	d
Total		a + c	d + b
		Case	Control

Table 3. Experiment results.

K	Measures	Numbers of Clusters (N)					
		2	4	10	14	18	22
1	Sensitivity	100	54	64	69	78	75
	Specificity	100	100	100	99	99	99
	Accuracy	100	77	82	84	88.5	87
3	Sensitivity	0	92	81	82	63	65
	Specificity	100	68	93	97	98	100
	Accuracy	50	80	87	89.5	80.5	83.5
5	Sensitivity		100	95	80	84	84
	Specificity		19	68	71	70	70
	Accuracy		59.5	81.5	75.5	77	77
7	Sensitivity		0	99	88	81	75
	Specificity		100	43	47	69	68
	Accuracy		50	71	67.5	75	71.5

91.3% for accuracy.

Please note, when N is 1, we have only 4 clusters/centroids, two from case and two from control. So we don't have results for K is 5 and for K is 7.

5. Conclusion

In this paper, we discuss the potential of applying a cluster-based distance algorithm on epidemiological studies. The proposed classification method based on cluster and distance is shown to have a high prediction rate without finding SNPs associated with the disease which may reduce the running time. Also using tag SNPs reduces the size of the problem and provides a better result. Current disease status prediction methods depend on the association studies. The genetic factors associated with the disease have to be identified first. Our methods can predict the individual's susceptibility without searching the genetic factors, thus save lots of money and efforts on that. In our future work we are going to continue validation of the proposed method.

References

[1] Type 2 Diabetes. http://ezinearticles.com/Type-2-Diabetes

[2] Cardon, L.R. and Bell, J.I. (2001) Association Study Designs for Complex Diseases. *Nature Reviews*: *Genetics*, **2**, 91-98. http://dx.doi.org/10.1038/35052543

[3] Hirschhorn, J.N. and Daly, M.J. (2005) Genome-Wide Association Studies for Common Diseases and Complex Diseases. *Nature Reviews*: *Genetics*, **6**, 95-108. http://dx.doi.org/10.1038/nrg1521

[4] Merikangas, K.R. and Risch, N. (2003) Will the Genomics Revolution Revolutionize Psychiatry. *The American Journal of Psychiatry*, **160**, 625-635. http://dx.doi.org/10.1176/appi.ajp.160.4.625

[5] Botstein, D. and Risch, N. (2003) Discovering Genotypes Underlying Human Phenotypes: Past Successes for Mendelian Disease, Future Approaches for Complex Disease. *Nature Genetics*, **33**, 228-237. http://dx.doi.org/10.1038/ng1090

[6] Clark, A.G., Boerwinkle, E., Hixson, J. and Sing, C.F. (2005) Determinants of the Success of Whole-Genome Association Testing. *Genome Research*, **15**, 1463-1467. http://dx.doi.org/10.1101/gr.4244005

[7] He, J. and Zelikovsky, A. (2006) Tag SNP Selection Based on Multivariate Linear Regression. *Proceedings of International Conference on Computational Science*, LNCS 3992, 750-757.

[8] Brinza, D., He, J. and Zelikovsky, A. (2006) Combinatorial Search Methods for Multi-SNP Disease Association. *Proceedings of International Conference of the IEEE Engineering in Medicine and Biology*, **1**, 5802-5805.

[9] Margaret, H.D. Data Mining—Intrdocution and Advanced Topics. Prentice Hall, Upper Saddle River.

[10] Mao, W., Brinza, D., Hundewale, N., Gremalschi, S. and Zelikovsky, A. (2006) Genotype Susceptibility and Integrated Risk Factors for Complex Diseases. *Proceedings of IEEE International Conference on Granular Computing*, 2006, 754-757.

[11] Kimmel, G. and Shamir, R. (2005) A Block-Free Hidden Markov Model for Genotypes and Its Application to Disease Association. *Journal of Computational Biology*, **12**, 1243-1260. http://dx.doi.org/10.1089/cmb.2005.12.1243

[12] Listgarten, J., Damaraju, S., Poulin, B., Cook, L., Dufour, J., Driga, A., Mackey, J., Wishart, D., Greiner, R. and Zanke, B. (2004) Predictive Models for Breast Cancer Susceptibility from Multiple Single Nucleotide Polymorphisms. *Clinical Cancer Research*, **10**, 2725-2737. http://dx.doi.org/10.1158/1078-0432.CCR-1115-03

[13] Wellcome Trust Case Control Consortium (WTCCC). http://www.wtccc.org.uk/