Statistical estimation of diagnosis with genetic markers based on decision tree analysis of complex disease

Liu Hui*, Gai Liping
College of Medical Laboratory, Dalian Medical University, Dalian 116027, China

A R T I C L E   I N F O
Article history:
Received 25 November 2006
Accepted 27 July 2009
Keywords:
Genetic diagnosis
Decision trees
Multivariate analysis

A B S T R A C T
To explore combinations of genetic markers and to estimate their joint action, decision trees are built on the basis of marker frequencies in both disease and control groups. Youden's index (0.1–0.9 for a single marker) is calculated for genetic markers with different diagnostic capacities. When 23 single genetic markers with diagnostic power 0.10 are combined, the resulting diagnostic power is 0.5. Medium diagnostic power (Youden's index 0.7) can be obtained by combining four low effect diagnostic items. High diagnostic power (Youden's index 0.9) can be obtained by combining either eight low power items or four medium power ones. This implies that selection of about 100 genetic markers, differing in capacity to distinguish between the disease and control groups by (say) 10%, will meet the requirement for clinical diagnosis. Thus, diagnosis of complex diseases by genetic markers is possible through the discovery and characterization of markers throughout the human genome and the development of genotyping technology.

© 2009 Elsevier Ltd. All rights reserved.

1. Introduction
Analysis of genetic linkage has been highly successful in mapping the genes responsible for Mendelian diseases. Many disorders with an underlying single-gene mode of inheritance have been identified by positional cloning. In the past decade, attempts have been made to extend this approach to multifactorial disorders and other health-related traits. It has proved difficult, however, to find strong and replicable linkages since a significant feature of a complex disease is the modest contribution of each susceptibility gene to its onset [1,2]. Therefore, diagnosis of complex diseases using single genes is limited in efficacy.

Recent progress in genotyping techniques enables us to use short tandem repeats (STR) or single nucleotide polymorphisms (SNP) as allelic markers of complex diseases [3,4]. It is possible by combining several genetic markers to improve the diagnostic power to meet the requirements for clinical diagnosis. Consequently, it is necessary to analyze the frequency indices and to evaluate the joint action of multiple genetic markers.

Multivariate logistic regression is the first choice for multivariate analysis [5]. However, it requires original data and thus is not suitable for theoretical illustration. In this paper we have chosen the decision tree analytical method [6,7], which solves these problems satisfactorily and lays a theoretical foundation for research and the clinical application of genetic markers.

2. Models and method

2.1. Setting of decision tree
We assume three genetic markers with those expected under the hypothesis of panmixia (Hardy–Weinberg equation), X1, X2 and X3, the frequencies of which are PX1, PX2 and PX3, respectively, and establish the decision tree shown in Fig. 1.

Calculations are performed using the VB Program as shown in Fig. 2 and VB Program Code is listed in Appendix A.

We can suppose three observation indices X1, X2 and X3, whose positive rates are Ptx1, Ptx2 and Ptx3, respectively, in a test group and Pcx1, Pcx2 and Pcx3, respectively, in the control group, thus decision tree of the test group can be set with Ptx1, Ptx2 and Ptx3 and decision tree of control group can be set with Pcx1, Pcx2 and Pcx3. We can work out Pt(i) and Pc(i) (i = 1, 2, 3, . . . ), respectively, according to probability of multiplication principle. The Pt(i) represents the probability calculated by the test group decision tree and Pc(i) represents the probability calculated by control group decision tree, equivalent to P1, P2 and P3 as shown in Fig. 1.

The algorithm is Pt(1) = Pt(1) × X1 × X2 × X3, Pt(2) = Pt(1) × X1 × (1 − X2) × X3, Pt(3) = Pt(1) × (1 − X1) × X2 × X3, Pt(4) = Pt(1) × (1 − X1) × (1 − X2) × X3, Pt(5) = Pt(1) × (1 − X1) × (1 − X2) × (1 − X3), Pt(6) = (1 − Pt(1)) × X2 × (1 − X3), Pt(7) = (1 − Pt(1)) × (1 − X2) × X3,

© 2009 Elsevier Ltd. All rights reserved.

AVI reprints information available at ScienceDirect
Contents lists available at ScienceDirect
Computers in Biology and Medicine
journal homepage: www.elsevier.com/locate/cbm

1. Introduction
Analysis of genetic linkage has been highly successful in mapping the genes responsible for Mendelian diseases. Many disorders with an underlying single-gene mode of inheritance have been identified by positional cloning. In the past decade, attempts have been made to extend this approach to multifactorial disorders and other health-related traits. It has proved difficult, however, to find strong and replicable linkages since a significant feature of a complex disease is the modest contribution of each susceptibility gene to its onset [1,2]. Therefore, diagnosis of complex diseases using single genes is limited in efficacy.

Recent progress in genotyping techniques enables us to use short tandem repeats (STR) or single nucleotide polymorphisms (SNP) as allelic markers of complex diseases [3,4]. It is possible by combining several genetic markers to improve the diagnostic power to meet the requirements for clinical diagnosis. Consequently, it is necessary to analyze the frequency indices and to evaluate the joint action of multiple genetic markers.

Multivariate logistic regression is the first choice for multivariate analysis [5]. However, it requires original data and thus is not suitable for theoretical illustration. In this paper we have chosen the decision tree analytical method [6,7], which solves these problems satisfactorily and lays a theoretical foundation for research and the clinical application of genetic markers.

2. Models and method

2.1. Setting of decision tree
We assume three genetic markers with those expected under the hypothesis of panmixia (Hardy–Weinberg equation), X1, X2 and X3, the frequencies of which are PX1, PX2 and PX3, respectively, and establish the decision tree shown in Fig. 1.

Calculations are performed using the VB Program as shown in Fig. 2 and VB Program Code is listed in Appendix A.

We can suppose three observation indices X1, X2 and X3, whose positive rates are Ptx1, Ptx2 and Ptx3, respectively, in a test group and Pcx1, Pcx2 and Pcx3, respectively, in the control group, thus decision tree of the test group can be set with Ptx1, Ptx2 and Ptx3 and decision tree of control group can be set with Pcx1, Pcx2 and Pcx3. We can work out Pt(i) and Pc(i) (i = 1, 2, 3, . . . ), respectively, according to probability of multiplication principle. The Pt(i) represents the probability calculated by the test group decision tree and Pc(i) represents the probability calculated by control group decision tree, equivalent to P1, P2 and P3 as shown in Fig. 1.

The algorithm is Pt(1) = Pt(1) × X1 × X2 × X3, Pt(2) = Pt(1) × X1 × (1 − X2) × X3, Pt(3) = Pt(1) × (1 − X1) × X2 × X3, Pt(4) = Pt(1) × (1 − X1) × (1 − X2) × X3, Pt(5) = Pt(1) × (1 − X1) × (1 − X2) × (1 − X3), Pt(6) = (1 − Pt(1)) × X2 × (1 − X3), Pt(7) = (1 − Pt(1)) × (1 − X2) × X3,
Determining diagnostic power

The analysis indicates that diagnostic power increases to 0.5 when 23 single genetic markers with diagnostic power 0.10 are combined. Medium diagnostic power (Youden's index 0.7) can be obtained by combining four low-power diagnostic items. High diagnostic power (Youden's index 0.9) can be obtained by combining either eight low-power items or four medium-effect ones. Details are listed in Table 1.

3. Results

2.2. Determination of diagnostic power

Youden's index is chosen as the measure of diagnostic power [8].

\[
Y = \text{Sen} + \text{Spe} - 1, \quad \text{Sen}(+) = \sum_{i=1}^{m} \text{Pt}(i), \quad P > 0.5, \\
\text{Spe}(-) = \sum_{i=1}^{m} \text{Pc}(i), \quad P \leq 0.5
\]

where \( Y \) is Youden's index, \( \text{Sen} \) is diagnosis sensitivity, that is, rate of (correct) diagnosis in the disease group; \( \text{Spe} \) is diagnosis specificity, that is, rate of (false) diagnosis in the control group. If the rate of one genetic marker is 0.2 in the disease group and 0.1 in the control group, for instance, then Youden's index of this marker is \([0.2+(1-0.1)]-1 = 0.1\). If \( Y < 0.5 \), the diagnostic power is low; if \( Y > 0.7 \), it is medium; if \( Y > 0.9 \), it is high.

4. Discussion

During the past 10 years, microsatellites have been the most widely used markers in linkage and association studies because of their high degree of heterozygosity. More recently, SNP have become more popular in association studies. In the course of the Human Genome Project, more than 1.68 million SNPs have been identified. Mapping the genetic basis underlying common multifactorial diseases such as cancer through whole genome association studies has attracted much attention in recent years [9,10]. The discovery and characterization of STR or SNP markers throughout the human genome and the development of genotyping technology is making diagnosis with genetic markers more technically feasible for complex diseases.

Multivariate analysis provides a method to study multiple problems with several factors and indices. Its basic principle is to use a group of samples with clear classification to find out the discriminant function, or discriminant principle, in accordance with several specified classification indices; whose purpose is using the function formula, or discriminant principle, to determine classification. The classification of unknown units identifies where it belongs; by analyzing the classification index of any sample. For enumeration index it usually chooses logistic regression to establish discriminant function. However, logistic has strict requirements to original data and each sample index must be full and each variable must be absolute. Also, complex diseases require different genome sign it is associated with. Therefore, decision tree analysis method is used in this essay to analyze the combination effect of several enumeration indices.

This article uses distribution of frequencies of different genetic indices in test group and control group to establish the decision tree of both test group and control group. The evaluation of frequencies and genetic indices is based on properties of the sample (classified) comparing the probability of decision trees calculated from the disease group and the control group, hence we can evaluate the combined effects of several genetic indices once we know (or estimate) the distribution of different genetic indices in the disease group and control group.

Since the clinical manifestation of the majority of complex diseases is diversified, typical diseases do not occur frequently, it is very difficult to obtain high-quality data of this kind of samples. In this case, the use of logistic analysis would not yield good result. Youden's index is the common index for evaluating experiment diagnosis effectiveness. It is the sum of positive rates of the positive group (Sen) and the negative rates of the negative group (Spe) minus 1, therefore it is suitable for evaluating indices of the combined effect. The coincidence index \((\text{coincidence} = (\text{Sen}+\text{Spe})/2)\) has the similar function with Youden's index; however, Youden's index analysis is more conducive to understanding the clinical significance of this article.
Table 1

Estimation of diagnostic power obtained by combining genetic markers.

<table>
<thead>
<tr>
<th>Single genetic markers (Y)</th>
<th>Frequency (%)</th>
<th>Number of markers combined (Y)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Disease group</td>
<td>Control group</td>
</tr>
<tr>
<td>0.1</td>
<td>20</td>
<td>10</td>
</tr>
<tr>
<td>0.2</td>
<td>30</td>
<td>10</td>
</tr>
<tr>
<td>0.3</td>
<td>40</td>
<td>10</td>
</tr>
<tr>
<td>0.4</td>
<td>50</td>
<td>10</td>
</tr>
<tr>
<td>0.5</td>
<td>60</td>
<td>10</td>
</tr>
<tr>
<td>0.6</td>
<td>70</td>
<td>10</td>
</tr>
<tr>
<td>0.7</td>
<td>80</td>
<td>10</td>
</tr>
<tr>
<td>0.8</td>
<td>90</td>
<td>10</td>
</tr>
</tbody>
</table>

The analysis shows that at least 23 poorly distinguishing genetic markers (diagnostic power 0.1) have to be combined to meet the requirements of low diagnostic power with Youden’s index 0.5 (Table 1). Only medium diagnostic power or better is usually considered to have clinical value. Medium diagnostic power (Youden’s index 0.7) can be obtained by combining four low power diagnostic items (Table 1). This implies that about 100 (23 × 4) STR or SNP, differing in gene frequency between the disease and control groups by 10%, must be combined to reach the requirement for clinical diagnosis. We can make a basic evaluation of the joint actions of genetic markers through Table 1. The diagnostic power obtained by combining genetic markers with various individual diagnostic powers can also be determined by the decision tree method. Usually, the combination of a low-diagnostic power genetic marker and a high-diagnostic power one is no more effective than the high-diagnostic power marker alone. Readers can experiment directly to explore the use of combinations of genetic markers for diagnosing complex diseases.

Conflict of interest statement

I, the undersigned author, certify that I have no commercial associations that pose a conflict of interest in connection with the submitted article. I declare that the above statement is true on behalf of all the authors related to this study.

Appendix A.

VB Program Code:

Private Sub cmd_cal_Click()
    For i = 1 To 2 ^ iTotalNum
        dTestResult(i) = 1
        dCompResult(i) = 1
    Next
    For i = 1 To iTotalNum
        dTestResult(i) = dTestResult(i) * dTestDataP(i)
        dCompResult(i) = dCompResult(i) * dCompDataP(i)
        dTestResult(i) = dTestResult(i) * (1 - dTestDataP(i))
        dCompResult(i) = dCompResult(i) * (1 - dCompDataP(i))
    Next
    For i = 1 To iTotalNum
        If Round(dTestResult(i)/(dTestResult(i) + dCompResult(i)), 3) > 0.5 Then
            yang1 = yang1 + dTestResult(i)
            yin1 = yin1 + dCompResult(i) + dTestResult(i)
            Else
            yin2 = yin2 + dCompResult(i) + dTestResult(i)
        End If
    Next
    yang = Round(yang1/yang2, 3)
    yin = Round(yin1/yin2, 3)
    y = Round(yang1 + yin1, 3) - 1
    dtestsum = 0:dcompsum = 0
    For i = 1 To iShowCols
        dtestsum = dtestsum + CDbl(ResultGrid.TextMatrix(i, 1))
dcompsum = dcompsum + CDbl(ResultGrid.TextMatrix(2, i)) Next i
ResultGrid.TextMatrix(0, lShowCols+1) = “total”
ResultGrid.TextMatrix(1, lShowCols+1) = dtestsum
ResultGrid.TextMatrix(2, lShowCols+1) = dcompsum
Exit Sub
errHandle:
MsgBox “error!”,. vbCritical
End Sub
Private Sub cmd_Sure_Click()
iTestNum = DataGrid.Rows-2
iTotalNum = iTestNum
For i = 1 To iTestNum
    dTestDataP(i) = CDbl(DataGrid.TextMatrix(i, 1))
    dCompDataP(i) = CDbl(DataGrid.TextMatrix(i, 2))
Next
For i = 1 To 2 ^ iTTestNum
    dTestResult(i) = 1
    dCompResult(i) = 1
Next
bSave = True
End Sub

References

Liu Hui (1960), male, born in Li, Hebei province, Han nationality, China, graduated from Dalian Medical College, MS. He is a Professor in the College of Medical Laboratory, Dalian Medical University, Dalian, China. His Research direction is molecular foundation on complex disease and bioinformation.