TISSUE ANTIGENS
IMMUNE RESPONSE GENETICS

Tissue Antigens ISSN 0001-2815

LETTER TO THE EDITOR

Decision tree classifier makes genotyping more intuitive and more efficient

H. Lee, B. Wang, X. Wu, H. Zhang & F. Xu

The Key Laboratory of Biomedical Information Engineering of Ministry of Education, School of Life Science and Technology, Xi'an Jiaotong University, Xi'an 710049, P.R.China

Key words: decision tree classifier; genotyping; human leucocyte antigen; ID3

With the fast development of life science research, a large number of species have been sequenced and more genomic sequence variants have been discovered, which leads to an increasing demand of genotyping for functional DNA sequences research. Genotyping refers to the process of determining the genetic constitution of an individual by examining its DNA sequence with a biological assay. The majority of assays require three basic steps: polymerase chain reaction (PCR), allele discrimination and allele detection. There are four popular allele discrimination methods: primer extension, hybridization, ligation and enzymatic cleavage. The major methods of allele detection include mass spectrometry, fluorescence and chemiluminescence (1). The human major histocompatibility complex is a genomic region on chromosome 6p21.3, and highly polymorphic human leukocyte antigen (HLA) genes in this locus perform the crucial function of antigen presentation. Medical research has found that more precise HLA matching between donor and recipient reduces immunological complications, and increases the survival rate in transplantation, particular for bone marrow transplantation (2). In addition, HLA typing results can also be used as important forensic evidence in paternity identification and criminal identification (3). PCR-SSP (sequence-specific primer) and PCR-SBT (sequence-based typing) are the most commonly used genotyping methods in this locus at present. Compared to PCR-SSP, PCR-SBT can genotype with high-resolution and can be used for discovering new alleles by sequences alignment. However, conventional PCR-SBT method has limited capability for resolving sequences of heterozygous samples in diploid genomes, resulting in ambiguous genotyping results. Some special sequencing methods such as pyrosequencing (4) and sequencing after cloning (5) can effectively solve the problem. A single-nucleotide polymorphism (SNP) is a base substitution of one nucleotide with another nucleotide (A, T, C or G). Genotypes of samples can be determined by detecting a set of SNPs, so genotyping based on SNPs is widely used.

With the development of artificial intelligence method, many classification algorithms such as Artificial Neural Networks (6), Support Vector Machine (7) and Decision Tree Classifier (DTC; (8–11)), have been applied in biomedical data mining domain. DTC is a nonparametric classification method, which contains a multistage approach of breaking up a complex decision into a union of several simpler decisions. In some representative DTC algorithms, ID3 is widely used for its discrete attributes, while C4.5 is used for both discrete and continuous attributes. The faster C5.0 algorithm based on binary tree gradually made DTC more specific (12–15). Figure 1A shows an example of one genotyping DTC building: multiple sequence alignment (MSA) helps extracting the feature SNPs from labeled sequences, then decision tree can be built up to make genotype classification. For complex structure such as HLA genes which comprise hundreds of SNPs, they also can be used to build genotyping DTC.

The ID3 algorithm was first brought forward by Quinlan, which chooses minimal entropy as the selection criterion of decision attribute. In this study, ID3 algorithm was used to make DTC. In order to apply algorithm of ID3 to make decision tree, node structure is constructed which contain attributes of (A, T, C, G and '-') the column position in MSA file of SNP and the level of node. When recursively called ID3 procedure on each node, the complete decision tree would be generated. A package based on MATLAB platform was developed [http://code.google.com/p/dtbg; Genotyping Decision Tree Classifier Builder (GDTCB)]. It has three main functions, which are extracting feature SNPs from the MSA file of training allelic variants, building DTC with feature SNPs, and the identification of the allelic variant by comparing genotype identity in SNPs of target sequence with DTC from root node to leaves.

Before genotyping, heterozygous sequence should be transformed into homozygous sequence pairs; a program in the package can help performing this function. The base types of feature SNPs in target sequence can be obtained by pairwise alignment of the target sequence and the reference consensus sequence, which was generated by MSA of HLA allelic variants database. Tree structure of genotyping DTC is saved in MATLAB's MAT format, so it can be directly loaded and viewed in MATLAB workspace. A practical genotyping decision tree structure drawing program was also developed, which can depict decision tree in detail in AUTOCAD's standard file format DWG. The drawing process and the final drawing result can be found in the supplement files S1, S2, S3.

The EBI IMGT/HLA database (http://www.ebi.ac.uk/imgt/hla/) provides the latest HLA genotyping sequences. We

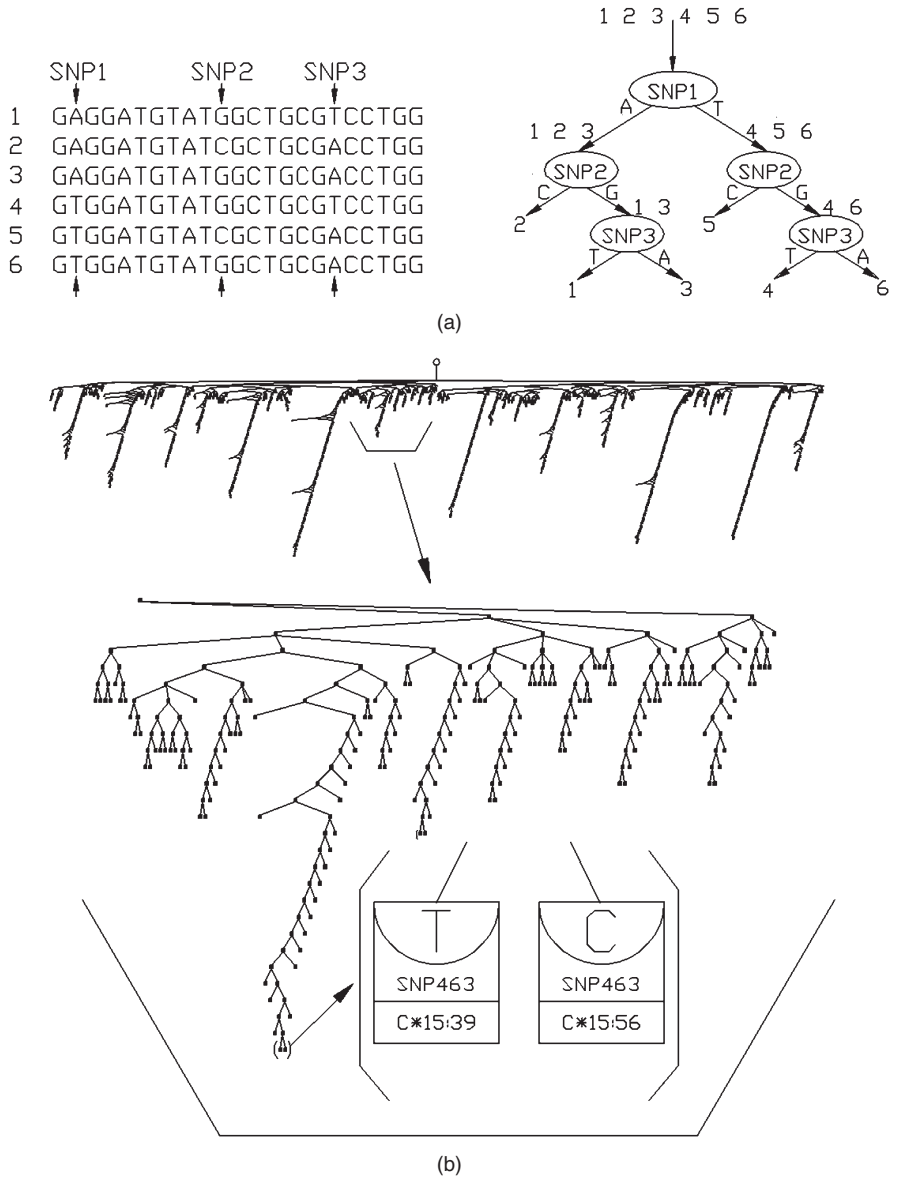

Figure 1 The building of genotyping Decision Tree Classifier. (A) Principle (B) an example of highly polymorphic gene HLA-C.

downloaded MSA files (with the filename extension '.msf') of some highly polymorphic HLA class I genes HLA-A, HLA-B and HLA-C from the database (ftp://ftp.ebi.ac.uk/pub/databases/imgt/mhc/hla/, release 3.6.0). After feature SNPs had been extracted from MSA files, we trained genotyping decision tree with SNPs data. Then we can use the decision tree to make sequence genotyping. The decision trees and the parameters for three genes are shown in Table 1, and the structure for gene HLA-C genotyping is shown on Figure 1B.

For each typing, DTC only make X (X = depth × K, node branch number K is between 2 and 5) comparisons at most,

Table 1 Statistics of three HLA class I genes after genotyping DTC building

			Genotyping DTC			First level of DTC SNP	
Gene	Alleles	SNPs	Node number	Depth	Width	Position in MSA file	Genotype
HLA-A	1729	1067	3280	145	1715	98	A, T, C
HLA-B	2329	1010	4407	80	2319	391	A, T, G, C
HLA-C	1291	1101	2469	87	1280	423	A, T, C

DTC, Decision Tree Classifier; HLA, human leukocyte antigen; MSA, multiple sequence alignment; SNP, single-nucleotide polymorphism.

so the typing efficiency is very high. In addition, some hidden information which included in the relationship between alleles, and differences in SNPs between alleles can be excavated and included in the decision tree. In conclusion, DTC make genotyping more intuitive and more efficient.

DNA sequencing based genotyping is undoubtedly the most accurate method, particular for polymorphism site enriched sequences. But when used for frequent genotyping analysis, it is hard to improve efficiency due to complex sequencing result. In this study, the known subtypes of sequences are used to extract polymorphism sites, to train the decision tree and to construct the decision tree. As a result, a small number of most informative SNPs were included in the decision tree to make genotyping. Using this method, decision trees related to HLA-A, HLA-B and HLA-C genes are constructed; they possess the ability of high-resolution genotype, and can be used for new genotype discovery. This method can be used to decrease biology experiment requirements, due to only the SNPs included in the decision tree are needed to be tested by experiments. This method also can be used for other genomic sequence research to reduce the genotyping experiments. Furthermore, it can assist other genotyping required applications such as organ transplantation, paternity testing or DNA evidence verification.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (No. 81102085/H2401, 81071299/H1103), Scientific Research Foundation of Shaanxi Provincial Office of Health, P.R. China (No. 2010E06, 2010D21).

Conflict of interest

The authors have declared no conflicting interests.

Correspondence

Xiaoming Wu
The Key Laboratory of Biomedical Information Engineering of Ministry of Education
School of Life Science and Technology
Xi'an Jiaotong University
Xi'an 710049
P.R.China
Tel: +86 29 8266 3454
Fax: +86 29 8266 3454
e-mail: wxm@mail.xjtu.edu.cn

doi: 10.1111/j.1399-0039.2012.01901.x

References

1. Kim S, Misra A. SNP genotyping: technologies and biomedical applications. *Annu Rev Biomed Eng* 2007: **9**: 289–320.
2. Persijn GG, Cohen B, Lansbergen Q, et al. Effect of HLA-A and HLA-B matching on survival of grafts and recipients after renal transplantation. *N Eng J Med* 1982: **307**: 905–8.
3. Lu HL, Wang CX, Wu FQ, et al. Paternity identification in twins with different fathers. *J Forensic Sci* 1994: **39**: 1100–2.
4. Ramon D, Braden M, Adams S et al. Pyrosequencing trade mark: a one-step method for high resolution HLA typing. *J Transl Med* 2003: **1**: 9.
5. Cox ST, McWhinnie AJ, Robinson J, et al. Cloning and sequencing full-length HLA-B and -C genes. *Tissue Antigens* 2003: **61**: 20–48.
6. Dayhoff JE, DeLeo JM. Artificial neural networks. *Cancer* 2001: **91**: 1615–35.
7. Ban HJ, Heo JY, Oh KS, et al. Identification of type 2 diabetes-associated combination of snps using support vector machine. *BMC Genet* 2010: **11**: 26.
8. Salzberg S, delcher AL, Fasman KH, et al. A decision tree system for finding genes in DNA. *J Comput Biol* 1998: **5**: 667–80.
9. Markey MK, Tourassi GD, Carey EF Jr. Decision tree classification of proteins identified by mass spectrometry of blood serum samples from people with and without lung cancer. *Proteomics* 2003: **3**: 1678–9.
10. Xie Q, Ratnasinghe L, Hong H, et al. Decision forest analysis of 61 single nucleotide polymorphisms in a case–control study of esophageal cancer; a novel method. *BMC Bioinformatics* 2005: **6**: s4.
11. Swaney DL, McAlister GC, Coon JJ. Decision tree-driven tandem mass spectrometry for shotgun proteomics. *Nat Methods* 2008: **5**: 959–64.
12. Quinlan JR. Discovering rules by induction from large collections of examples. In: Michie D, ed. *Expert Systems in the Micro Electronic Age. Edinburgh, UK: Edinburgh University Press, 1979.*
13. Quinlan JR. Induction of decision trees. *Mach Learn* 1986: **1**: 81–106.
14. Quinlan JR. Improved use of continuous attributes in c4.5. *J Artif Intell Res* 1996: **4**: 77–90.
15. Breiman L, Friedman JH, Olshen RA, et al. *Classification and Regression Trees. New York: Chapman&Hall (Wadsworth, Inc.), 1984.*

Supporting Information

The following supporting information is available for this article:

Movie S1. Decision tree classifier build process of gene HLA-G.

Movie S2. Decision tree classifier profile for gene HLA-C.

Movie S3. Decision tree structure represented in AutoCAD format.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

