华南农业大学学报 2016,37(3):115-120
Journal of South China Agricultural University
http://xuebao.scau.edu.cn
doi:10.7671/j.issn.1001-411X.2016.03.018

邓继忠,林伟森,甘四明,等.一种二倍体片段测序中SNP检测系统的构建[J].华南农业大学学报,2016,37(3):115-120.

一种二倍体片段测序中SNP检测系统的构建

邓继忠[1],林伟森[1],甘四明[2],黄华盛[1,3],李 梅[2],金 济[1],何明昊[1]
(1 华南农业大学 工程学院,广东 广州 510642;
2 中国林业科学研究院 热带林业研究所,广东 广州 510520;3 广东科技学院,广东 东莞 523083)

摘要:【目的】开发基于模式识别方法的二倍体片段测序中单核苷酸多态性(Single nucleotide polymorphism,SNP)自动检测系统,提高检测的准确性。【方法】采用LabWindows/CVI 9.0开发平台,结合Matlab函数库编程,以二倍体PCR片段测序的.ab1或.scf格式文件作为源数据,首先分离出碱基G、A、T和C,进行一维离散小波滤波,再对各碱基处的波形进行典型特征提取,最后运用基于反向传播神经网络的分类器完成SNP识别和判断。【结果】系统界面友好、运行稳定。SNP等级分为6级,允许用户对可疑的SNP进行人工修正,对尾叶桉*Eucalyptus urophylla*的26个测序序列143个SNP的测试中检测准确率、假阳性率和假阴性率均明显优于之前的类似软件。【结论】本文所构建的SNP自动检测系统准确性高,不需参考序列,可用于二倍体PCR片段测序中SNP的高效检测。

关键词:二倍体;测序;SNP检测;模式识别;尾叶桉
中图分类号:TP391.4;Q523.8 **文献标志码:**A **文章编号:**1001-411X(2016)03-0115-06

Development of an automatic system for SNP detection in diploid fragment sequencing

DENG Jizhong[1], LIN Weisen[1], GAN Siming[2], HUANG Huasheng[1,3], LI Mei[2], JIN Ji[1], HE Minghao[1]
(1 College of Engineering, South China Agricultural University, Guangzhou 510642, China;
2 Research Institute of Tropical Forestry, Chinese Academy of Forestry, Guangzhou 510520, China;
3 Guangdong University of Science and Technology, Dongguan 523083, China)

Abstract:【Objective】This study aims to develop a pattern-recognition based system for automatic single nucleotide polymorphism (SNP) detection in diploid fragment sequencing and improve the detection accuracy. 【Method】The LabWindows/CVI 9.0 platform and Matlab environment were combined for analyzing .ab1 or .scf files generated in diploid PCR fragment sequencing. Firstly, four bases G, A, T and C were separated for eliminating noise through one-dimensional discrete wavelet filtering, following with extraction of typical features of each base position (peak) from a fluorescence curve. A classifier based on back-propagation neural network was then used for SNP recognition and diagnosis. 【Result】This established system was characterized by friendly interface, stable operation and manual modification accessibility. It classified the SNP reliability into six grades. Performance test with 143 SNPs of 26 sequencing fragments from *Eucalyptus urophylla* demonstrated that our system outperformed three previously reported software packages in detecting accuracy, false positive and false negative rates. 【Conclusion】Our system has a high rate of accuracy without the need for a reference sequence. It could be used for efficient SNP detection in diploid PCR fragment sequencing.

收稿日期:2015-08-26 **优先出版时间:**2016-04-15
优先出版网址:http://www.cnki.net/kcms/detail/44.1110.s.20160415.1554.002.html
作者简介:邓继忠(1963—),男,副教授,博士,E-mail:jz-deng@scau.edu.cn;通信作者:甘四明(1970—),男,研究员,博士,E-mail:Siming_Gan@126.com
基金项目:863计划(2013AA102705);国家自然科学基金(31270702,31070592)

Key words: diploid; sequencing; SNP detection; pattern recognition; *Eucalyptus urophylla*

单核苷酸多态性(Single nucleotide polymorphism,SNP)是指生物体基因组中单个碱基的颠换或转换导致的DNA变异,是最普遍和最广泛的基因组变异形式,如拟南芥 *Arabidopsis thaliana* 3个品系间存在超过82万个SNP,平均约1.4 kb就存在1个SNP[1]。SNP目前已广泛应用于生物、医学和农学等研究领域[2-3]。

传统的PCR片段测序中,测序厂家所提供的软件只能识别各波峰位置的单个碱基,而对二倍体个体内存在SNP的双峰处的低峰碱基不能有效检测。已有一些第三方软件可用于二倍体物种PCR片段测序的SNP自动检测,如Mutation Surveyor(http://www.softgenetics.com/MutationSurveyor.html)、novoSNP[4]和PolyPhred[5]等。但是,这些软件均需参考序列,具有局限性,不能有效用于一些序列的分析,如表达序列标签(Expressed sequence tag,EST)和候选基因的测序中的内含子区域,而且操作上较为繁琐。这些软件的检测准确性均偏低[6]。此外,PolyPhred[5]还需要至少8个测序文件以保证足够的准确性,不利于少量样品的检测。针对二倍体PCR片段测序中如何进行SNP的自动、有效检测,本文基于模式识别的方法构建了无需参考序列进行SNP自动检测的计算机系统,并验证了系统的准确性。

1 材料与方法

1.1 系统开发平台

本研究基于LabWindows/CVI(以下简称CVI)平台,以CVI 9.0作为系统前台进行主界面的设计和基本功能菜单的实现。CVI是基于C语言的虚拟仪器软件开发平台,其功能强大,具有灵活的交互式编程方法和丰富的用户接口资源[7-8],且运行速度快、界面控件丰富。

采用MATLAB 2012函数库编程实现模式识别的分类器构建。MATLAB具有编程简单、仿真能力强和易于扩展移植等优点[9],提供内部神经网络工具箱,且可外接用户自行开发的模式识别算法或软件包,有利于用户进行各种模式识别算法的测试与建模。本文主要通过MATLAB实现模式识别的分类器结构的建立,利用反向传播神经网络(Back-propagation neural network, BPNN)进行权值和阈值的训练,再将训练好的BPNN结构移植至CVI环境用作SNP分类器。

系统的运行环境为Windows XP Professional、Windows 7或Window 8。

1.2 系统功能模块和界面的设计

系统主要包括测序数据导入、碱基分离、噪声处理、SNP识别、数据显示与存储、人工校正等功能模块。测序数据导入可打开扩展名为.ab1或.scf的测序文件。碱基分离是根据测序数据中G、A、T和C 4种碱基的不同荧光颜色而将它们分别提取出来。噪声处理通过一维离散小波变换滤除噪声,便于后续的数据特征提取。SNP识别主要根据测序峰图的波形特征,结合二倍体内SNP表现为双峰的现象,提取各碱基位置的典型波形特征作为BPNN的输入向量,输出SNP属性分数,再参照判别标准给出SNP等级。数据显示可在运行窗口直观显示测序峰图和序列以及SNP的位置、双峰碱基、属性分数和等级,识别结果可存为.txt文件。人工校正允许用户对误判和漏判的SNP进行删除、添加和更改碱基等手动操作。系统的工作流程见图1。

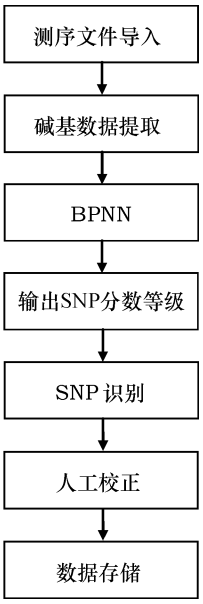


图1 SNP检测流程图

Fig. 1 A flow chart of the SNP detection process

系统界面包括主操作区、原始峰图显示区、SNP识别结果区和人工校正区等4个功能区域(图2)。

主操作区位于系统界面的上部,如图2所示,包含面板选项卡、菜单栏、快捷键和信息提示栏。面板选项卡可提供多个检测面板,各检测面板的序列分析数据相互独立,以便进行多个测序文件的检测。菜单栏包括文件、编辑、滤波选择、识别方法和帮助等一级菜单,各一级菜单可下拉显示二级菜单。文件的二级菜单包括原始测序文件导入、SNP检测结

果保存等。编辑的二级菜单包括 *X* 轴和 *Y* 轴缩放以及隐藏/显示碱基 G、A、T 和 C。滤波选择的二级菜单包括小波变换等方法。识别方法的二级菜单包括 BPNN 等多种方法。快捷键工具栏列出常用功能,包括文件打开、SNP 检测(按默认的滤波和模式识别方法)、增加面板、关闭当前面板、峰图和结果显示窗口 *X* 轴和 *Y* 轴缩放等。信息提示栏显示文件路径等。

原始峰图显示区位于界面右中区域,见图 2,不同颜色代表碱基 G、A、T 和 C 显示测序的峰图,可通过主操作区的快捷按钮选择性地显示一种或几种碱基类型,也可进行 *X* 轴和 *Y* 轴的缩放。

SNP 识别结果区位于界面下部区域,见图 2,界面右下区域显示 SNP 识别结果,包括碱基位置、碱基类型、双峰 SNP、SNP 分数和 SNP 等级等。界面左下区域列表总结 SNP 检测的结果,点击某一行可与右下窗口相应位置的 SNP 进行快速定位。

人工校正区位于界面左中区域(图 2),允许用户对任一位置碱基进行更改、删除和增加等操作。

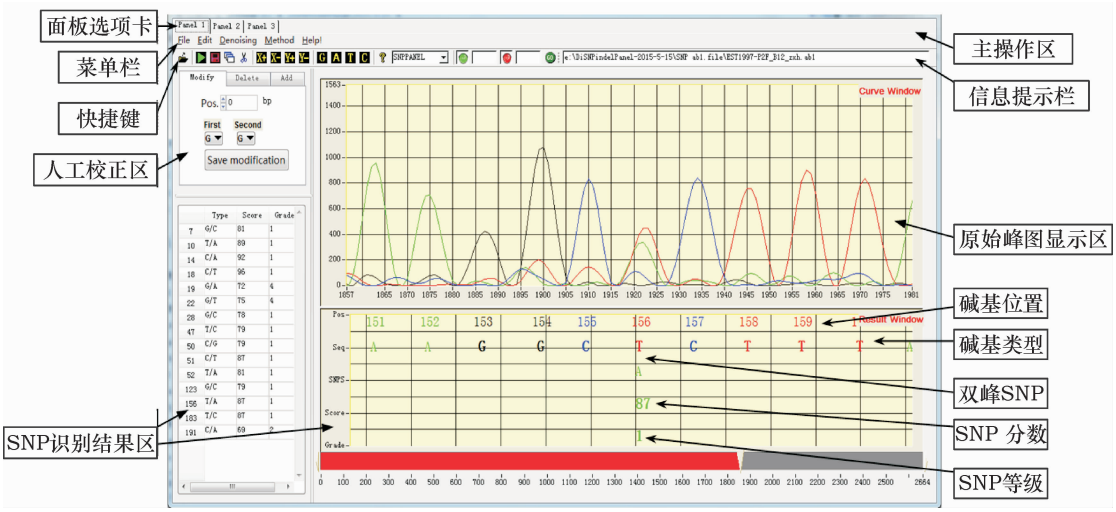


图 2　SNP 检测系统的界面

Fig. 2　A typical interface of the SNP detection system

1.3　系统主要功能的实现

1.3.1　*测序数据的导入和碱基分离*　本系统可导入. ab1 或. scf 格式的测序文件,根据文件内容分离出各种碱基。以 . ab1 文件为例说明。按照美国 Applied Biosystems 公司的 ab1f 文件格式, . ab1 文件是由一个指定的文件夹存储文件信息,包括文件名称、碱基类型和碱基数量等。文件 7 ~ 34 字节是'DirEntry'结构体,指向文件夹所在的位置,其结构为:

Struct DirEntry { sInt32 name; sInt32 number; sInt16 elementtype; sInt16 elementsize; sInt32 numelements; sInt32 datasize; sInt32 dataoffset; sInt32 datahandle; },

其中,number 显示通道编号(对应 4 种碱基),elementsize 显示碱基所占空间大小,numelements 显示碱基数量, dataoffset 显示碱基位置的偏移量。根据这些信息即可将各通道的碱基和荧光数据提取出来。

1.3.2　*滤波的实现*　分离出的碱基荧光信号通常含有噪声,影响 SNP 分析的准确性,可通过小波分解进行噪声滤除。小波分解具有良好的时频分析特性,在时域和频域内都具有突显信号局部特征的能力,并且能够在多种分辨率下观测数据,是噪声滤除的有效方法[10-11]。

本系统调用了 CVI 中 Signal Processing Toolkit 7.0.2 工具箱所含的小波工具包的函数,一维离散小波的滤波过程如下:分离 4 个通道的碱基数据;读取各通道的碱基数据长度,获得滤波参数;调用 SptAllocCoeffWFBD 函数,分配小波滤波工具包的滤波参数结构;调用 SptReadCoeffWFBD 函数,读取小波工具包参数;调用 SptDiscreteWaveletTransform (filterData, dataLenght, analysisFilter, SptSymmetric, NULL, NULL, scales, shift, dwt, &outputSize, lengths)函数,进行一维离散小波分解,滤除后的数据存储于变量 dwt 中,用于后续分析。

1.3.3　*SNP 识别分类器的构建*　利用 BPNN 构建 SNP 识别分类器。BPNN 是一种反向传播且能修正误差的多层映射模式,由输入层、隐含层和输出层及各层神经元之间连接组成[12]。SNP 识别可以视作从波形特征输入到 SNP 类型输出的非线性映射问题。

基于文献[13],选择了 SNP 处双峰的波峰距离、高度比值和起伏度比值作为特征参数,归一化处理

后作为 BPNN 输入层的特征向量,即输入层的神经元数为3。BPNN 输出层为 SNP 类别的诊断结果,可能是双峰 SNP 位点或单峰非 SNP 位点,即输出层神经元为1;输出层中设定范围0 ~ 100为1个位点的属性分数,并结合属性分数和周边杂峰将 SNP 可能性分为6个等级,等级越小越可能为 SNP。

MATLAB 神经网络工具箱中调用 BPNN 构建 SNP 分类器的过程具体如下:调用 net = newff[input, output, 10, ('tansig', 'purelin'), 'trainlm'],初始建立适于 SNP 识别的 BPNN 分类器;调用 net. trainParam. goal = 0. 01,设置误差,当网络精度达到误差则停止训练,训练好的权值和阈值存于 MATLAB 中 net 结构体的 net. IW、net. LW 和 net. b 元素;调用 net = train (net, input, output),CVI 中重新塑造 BPNN 即可用于 SNP 检测。

1.4 数据处理

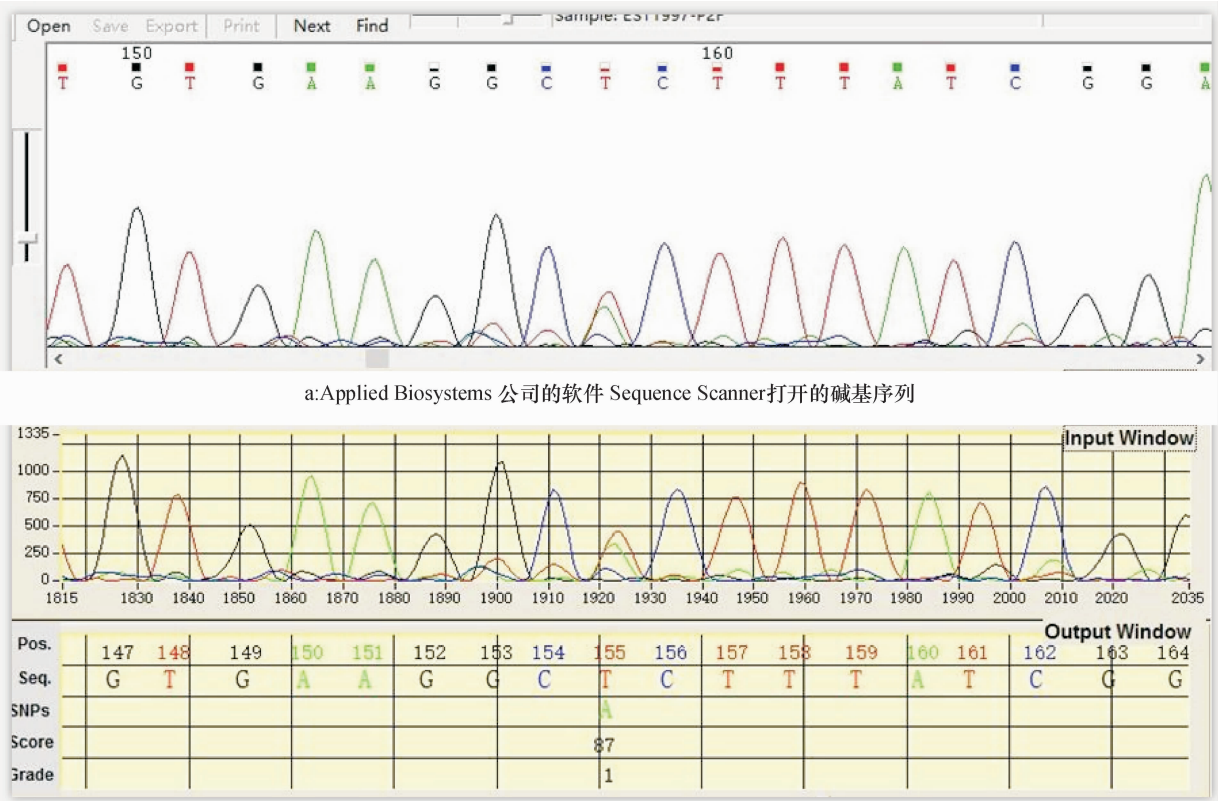
因 SNP 检测可能存在假阳性和假阴性,人工校正的碱基删除、添加和更改的结果可直接显示在当前面板。SNP 检测结果的保存格式为. txt 文件。

利用尾叶桉 *Eucalyptus urophylla* 26个测序文件的143个 SNP 进行系统测试,并与 Mutation Surveyor (http://www. softgenetics. com/MutationSurveyor. html)、novoSNP[4] 和 PolyPhred[5] 的结果进行比较,分析软件的有效性。所用26个测序文件为尾叶桉 SNP 开发的 EST 重测序文件[14]。

2 结果与分析

2.1 系统主要功能的实现

系统导入测序文件后能正确地进行碱基判读,判读结果与测序结果较为一致。图3显示了本系统对1个序列的判读与测序结果一致。但对测序质量较差或者杂峰较多的区域,测序结果亦不准确,与本系统的判读结果可能有异。图3B 表明本系统能可靠地识别双峰碱基位置的 SNP,155 bp 处的 SNP 判读为 T/A,分值为87,等级为1。并且,其他较低的杂峰被有效排除,未被误判为 SNP。



a:Applied Biosystems 公司的软件 Sequence Scanner打开的碱基序列

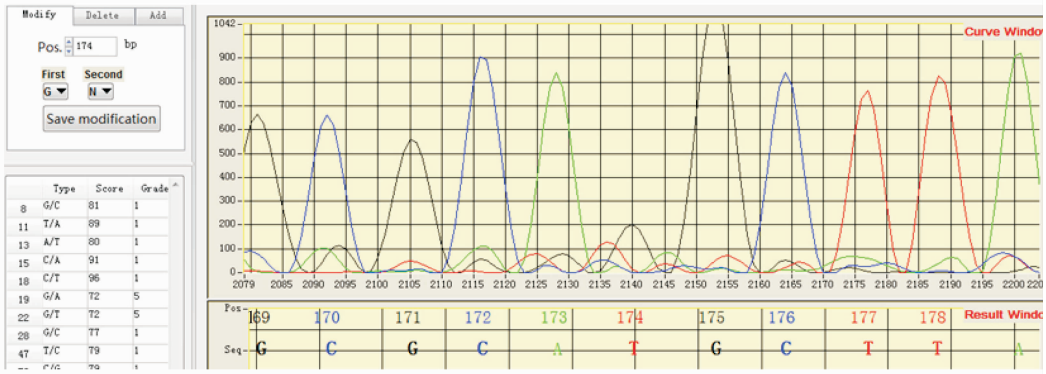
b:本系统判读的序列,155 bp 处双峰、SNP 为 T/A

图3 本系统导入测序文件后的判读序列与测序比较

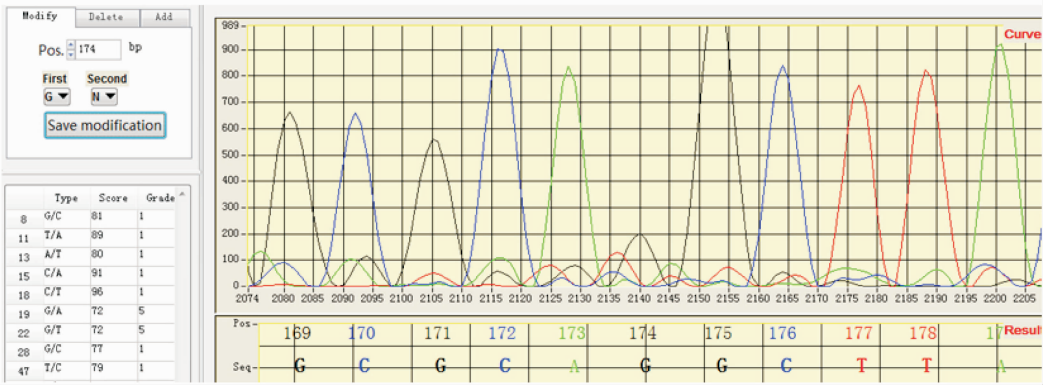
Fig. 3 Base sequence obtained in our system being compared to that of sequencer

人工校正可对特定位置的第1和/或第2碱基进行修改、删除和添加操作。图4显示了对174 bp 处误读的'T'修改为'G',表明系统可有效地进行人工校正。

数据存储的. txt 文件见图5,第1行为文件名,数字代表显示碱基位置,识别序列上面一行为碱基序列,下面一行显示识别的 SNP 的第2碱基。测序前20个碱基一般是较杂乱的峰,SNP 可靠性低。



a:174 bp处‘T’为误读，需修改



b:校正后 174 bp 处更正为‘G’

图 4　人工校正的有效性

Fig. 4　Efficient manual correction of a misidentified base

```
zxh_2008-06-24_EST1997-P2F_B12_zxh.ab1
1 A G A A A A G A A T A T A C A A A C G C A G G C T G A G A A A C T A T G C A C C
1           C   A       A       T A   T

41 A A G G T N T C G C A T C A T T T C C T T G C C A T T G C A A A A G C T C T G G
41

81 G T G T G A T G C G T G A A T G T T T T C T C A T G T T T G A A T C T G T C T C
81

121 A G G T T T T G G A G C T G G A T A G C A G G A A T G T G A A G G C T C T T T A
121                                                       A

161 T C G G A G G G C G C A G G C T T A C A T T G A G C T G G C T G A T C T T G A T
161                                   C             A

201 T T G G C T G A A G T T G A C A
201
```

图 5　SNP 检测结果存为. txt 文件

Fig. 5　SNP detection result saved as a . txt file

2.2　SNP 检测的有效性

对于测试的尾叶桉 26 个测序文件,对比另外 3 个软件,本系统具有较高的准确性和误检率。表 1 显示了不同软件在高和低的标准下对 143 个 SNP 的识别准确率、漏判(假阴性)率和误判(假阳性)率,不管是在高还是低的标准下,本文所开发的系统均有最高的 SNP 判读准确率以及最低的假阴性率和假阳性率,PolyPhred 除外,因其未检测到任何 SNP。

表 1　不同软件识别 SNP 的结果对比[1)]

Tab. 1　Comparison of software performance in SNP detection

判别	SNP	DiSNPindel		novoSNP		Mutation Surveyor		PolyPhred	
标准	准确性	SNP 数量	判别率/%	SNP 数量	判别率/%	SNP 数量	判别率/%	SNP 数量	判别率/%
高	准确	138	96.5	12	8.4	30	21.0	0	0
	漏判	5	3.5	131	91.6	113	79.0	143	100
	误判	45	24.6	17	58.6	31	50.8	0	0
低	准确	141	98.6	34	23.8	31	21.7	0	0
	漏判	2	1.4	109	76.2	114	78.3	143	100
	误判	75	42.9	89	72.4	31	50.0	0	0

1) DiSNPindel、novoSNP、Mutation Surveyor 和 PolyPhred 高的标准分别为等级 2、分值 13、中敏感性和等级 2[6],低的标准分别为等级 4、分值 6、高敏感性和等级 4 准确 SNP 的判别率 = 准确 SNP 数量/(准确 SNP 数量 + 漏判 SNP 数量)/ × 100%,漏判 SNP 的百分比 = 漏判 SNP 数量/(准确 SNP 数量 + 漏判 SNP 数量)/ × 100%,误判 SNP 的百分比 = 误判 SNP 数量/(误判 SNP 数量 + 准确 SNP 数量) × 100%。

3 结论与讨论

本文基于CVI 9.0开发平台,结合MATLAB函数库,运用特征提取和模式识别等技术构建了二倍体片段测序的SNP自动检测系统。本系统开发中结合了CVI运行速度快和用户接口资源丰富以及MATLAB函数库多样和用户可自行扩充函数等优点,降低了系统开发的难度,并且检测系统界面友好、易于扩充。系统各模块功能的完整实现和测试的有效性也证明相关技术的合理运用。

高通量的新一代测序技术已经兴起,但传统的PCR片段测序仍非常重要[15]。测序厂家提供的软件只能识别各序列位置的最高峰所对应的碱基,因此双峰位置的低峰碱基的识别需要第三方软件。另外,测序也常出现杂峰,且杂峰可能高于碱基峰而导致测序厂家的软件不能正确读序。本文针对这一问题研发了SNP自动检测系统,可以有效实现双峰判读和杂峰滤除等功能,为二倍体PCR片段测序的SNP判读提供了有力工具。

与功能类似的软件Mutation Surveyor (http://www.softgenetics.com/MutationSurveyor.html)、novoSNP[4]和PolyPhred[5]相比,本系统不需参考序列,既为EST和候选基因的测序中内含子无参考序列的SNP检测提供了便利,也减少了输入参考序列的繁琐操作;并且,本文所开发的检测系统具有最高的SNP判读准确率以及最低的假阴性率和假阳性率,可用于二倍体PCR片段测序中SNP的高效检测。

实际应用中,Mutation Surveyor和novoSNP可能假阳性率极高,如对15个基因171个SNP的检测中的假阳性SNP分别为3 728和505个,显著多于正确判读的SNP[16]。另一方面,任何检测系统的准确率都严重依赖于测序质量,实际应用时需要优化测序体系[4]。

参考文献:

[1] OSSOWSKI S, SCHNEEBERGER K, CLARK R M, et al. Sequencing of natural strains of *Arabidopsis thaliana* wit short reads[J]. Genome Res, 2008,18(12):2024-2033.

[2] 唐立琼,肖层林,王伟平. SNP分子标记的研究及其应用进展[J]. 中国农学通报, 2012,28(12):154-158.

[3] 许家磊,王宇,后猛,等. SNP检测方法的研究进展[J]. 分子植物育种, 2015,13(2):475-482.

[4] WECKX S,DEL-FAVERO J,RADEMAKERS R,et al. novoSNP, a novel computational tool for sequence variation discovery[J]. Genome Res, 2005,15(3):436-442.

[5] MATTHEW S,JAMES S,ROBERTSON P D,et al. Automating sequence-based detection and genotyping of single nucleotide polymorphisms (SNPs) from diploid samples[J]. Nat Genet, 2006,38(3):375-381.

[6] DENG J Z,HUANG H S,YU X L,et al. DiSNPindel: Improved intra-individual SNP and InDel detection in direct amplicon sequencing of a diploid[J]. BMC Bioinformatics, 2015,16:343.

[7] 仇志平,李树军. LabWindows/CVI虚拟仪器软件在测试领域中的应用[J]. 计算机工程与设计,2007,28(22):5544-5548.

[8] 刘君华. 虚拟程序编程语言LabWindows/CVI编程[M]. 北京:电子工业出版社,2001.

[9] 肖伟,刘忠,曾新勇,等. MATLAB程序设计与应用[M]. 北京:清华大学出版社,2005.

[10] BUI T D,CHEN G. Translation-invariant denoising using multiwavelets[J]. IEEE Trans Sig Proc, 1998,46(12):3414-3420.

[11] PAN Q,ZHANG P,DAI G,et al. Two denoising methods by wavelet transform[J]. IEEE Trans Sig Proc,1999,47(12):3401-3406.

[12] MCKEOWN J J,STELLA F,HALL G. Some numerical aspects of the training problem for feed-forward neural nets [J]. Neural Netw,1997,10(9):1455-1463.

[13] 黄华盛. 基于模式识别的二倍体个体内SNP和InDel自动检测[D]. 广州:华南农业大学,2014.

[14] YU X,GUO Y,ZHANG X,et al. Integration of EST-CAPS markers into genetic maps of and *Eucalgptus urophylla* and *E. tereticornis* and their alignment with *E. grandis* genome sequence[J]. Silvae Genet, 2012,61(6):247-255.

[15] STUDER A,ZHAO Q,ROSS-IBARRA J,et al. Identification of a functional transposon insertion in the maize domestication gene *tb1*[J]. Nat Genet,2011,43(11):1160-1163.

[16] NGAMPHIW C, KULAWONGANUNCHAI S, ASSAWAMAKIN A,et al. VarDetect: A nucleotide sequence variation exploratory tool [J]. BMC Bioinformatics, 2008, 9 (S12):9.

【责任编辑 霍 欢】