DOI:10.3969/j.issn.1673-4785.201207015
网络出版地址:http://www.cnki.net/kcms/detail/23.1538.TP.20130125.1524.011.html

采用最小误差阈值分割算法的基因芯片图像分析

尹宁,刘富,张玉
(吉林大学 通信工程学院,吉林 长春 130025)

摘 要:为了能够较好地处理芯片图像,尽可能准确地提取出描述基因样点的数据信息,采用了最小误差阈值的分割算法.该方法在假设目标和背景的分布服从混合正态分布的前提下,设定了最小误差分类目标函数,通过求得使目标函数值最小的最佳分割阈值,实现基因样点和背景图像的分割.针对分割出来的基因样点图像提取特征数据,最后对这些数据进行聚类分析,进而对实验样点进行分类.在实验中应用该方法分析了2组基因芯片图像,基因样点的分类效果较好,验证了该基因芯片分析方法的可行性.

关键词:基因芯片图像;图像分析;最小误差阈值分割;聚类分析
中图分类号:TP391.41 **文献标志码**:A **文章编号**:1673-4785(2013)01-0028-05

Image analysis of gene chip using minimum error threshold segmentation algorithm

YIN Ning, LIU Fu, ZHANG Yu
(College of Communication Engineering, Jilin University, Changchun 130025, China)

Abstract: In order to analyze gene chip image better, along with extract the data information as accurately as possible, to describe the gene sample, this research paper proposes to implement a minimum error threshold segmentation method. Based on the assumption that the distributions of object and background are governed by a mixture normal distribution, this method sets an objective function of minimum error classification. This method also allows for the implementation of the segmentation between gene sample and background image through calculating the optimal segmentation threshold by minimizing the objective function. Next, the feature data from the segment of gene sample image was extracted and a clustering analysis with the data was done to realize the successful classification of the experimental samples. The study examined two groups of gene chip images and analyzed them by using this method in the experiment. The results show that the classification result was better and the feasibility of the analysis method was verified.

Keywords: gene chip image; image analysis; minimum error threshold segmentation method; cluster analysis

随着人类基因组计划(human genome project)的提出以及实施[1],很多生物信息学的相关技术得到了飞速的发展.特别是基因芯片技术由于其强大的基因组信息分析功能,应用于生物科学的众多领域,成为许多研究机构研究的重点.在过去几年中,国外已经出现了一批商业或科研用的基因芯片图像处理与分析软件.而国内成型的软件产品不多,主要软件技术都被国外大型软件公司掌握,如Bluefuse、GenePix、ScanAlyze、QuantArray等.这些软件对芯片图像的处理与分析或者采用手工、半自动的样点定位方法,或者将样点的形状假定为圆形;由于实际的样点图像很少完全是圆形,有的呈现椭圆形,有的呈现空心形,因此这些商用软件读取信号的准确率并不非常可靠.

由于基因芯片图像容易受到制备和扫描过程中

收稿日期:2012-07-03. 网络出版日期:2013-01-25.
基金项目:吉林省科技发展计划资助项目(10100505).
通信作者:刘富. E-mail: liufu@jlu.edu.cn.

玻片不洁、光线不均以及杂交反应不彻底等因素的影响[2],如何滤除噪声,并且完整地保留基因样点的边缘特征是基因芯片图像分析的关键步骤. 目前基因样点分割算法[3]主要有模板匹配[4-5]、阈值分割以及特殊理论(如形态学分割[6]、模糊聚类分割算法[7])等. 通过对一些算法的比较,本文选用了最小误差阈值分割算法对基因样点进行分割处理得到了较好的效果. 尤其是针对基因芯片图像中基因样点模糊、与背景对比度不清晰的情况,分割处理的效果良好,可以完整地分割基因样点并保留其边缘细节特征,大大提高了基因芯片识别的效果.

1　最小误差阈值分割算法

最小误差阈值法[8]是基于 Bayes 理论,由 Kittler 和 Illingworth 提出的,国际上有很多学者对该算法进行了研究,目前已经提出了很多最小误差阈值算法的改进算法以及二维扩展算法等. 通常为了更加清楚地描述该方法,大多选用信息论中的相对熵的概念进行解释.

设 I 为一幅大小为 $M \times N$ 的数字图像,图像上各点的像素值由函数 $f(x,y)$ 来表示,x、y 为该点的横纵坐标值,且 $f(x,y) \in G = \{0,1,\cdots,L-1\}$. 图像的灰度直方图用 $p(g)$ 来表示,它可以看成是由目标和背景 2 个区域像素组成的混合总体的概率密度函数:

$$p(g) = \sum_{i=0}^{1} P_i \times p(g/i).$$

式中:P_i 是子分布的先验概率,$p(g)$ 的 2 个子分布 $p(g/i)$ 分别服从均值为 μ_i、方差为 σ_i 的正态分布:

$$p(g/i) = \frac{1}{\sqrt{2\pi}\sigma_i}\exp\left(-\frac{(g-\mu_i)^2}{2\sigma_i^2}\right), i = 0,1.$$

对于阈值 $t \in G$,最小误差阈值方法给出函数:

$$J(t) = 1 + 2[P_0(t)\ln\sigma_0(t) + P_1(t)\ln\sigma_1(t)] - 2[P_0(t)\ln P_0(t) + P_1(t)\ln P_1(t)].$$

式中:

$$P_0(t) = \sum_{g=0}^{t} h(g), P_1(t) = \sum_{g=t+1}^{L-1} h(g),$$

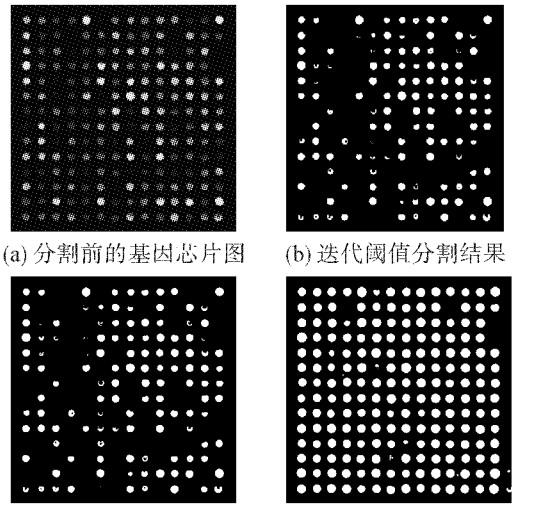
$$\mu_0(t) = \sum_{g=0}^{t} h(g)g/P_0(t), \mu_1(t) = \sum_{g=t+1}^{L-1} h(g)g/P_1(t),$$

$$\sigma_0^2(t) = [\sum_{g=0}^{t} (g-\mu_0(t))^2 h(g)]/P_0(t),$$

$$\sigma_1^2(t) = [\sum_{g=t+1}^{L-1} (g-\mu_1(t))^2 h(g)]/P_1(t).$$

最佳阈值选为使 $J(t)$ 最小化的 t^*,$t^* = \arg\min_{0<t<L-1} J(t)$. 本文根据上述的分割算法原理,对基因芯片图像进行图像分割处理. 图 1 为上述算法与其他常见的基因芯片图像分割算法的实验对比图.

针对基因芯片图像的特点进行仿真实验,综合观察 3 种阈值分割方法得到的图像,可以看出迭代法和最大类间差阈值分割算法(Ostu 算法)[9]较为简单,处理速度快;但是得到的分割图片中,基因样点的缺损比较多,对于与背景灰度靠近的样点往往不能识别. 最小阈值分割算法虽然有些复杂但是能得到较好的效果,能够更完整地分割出基因芯片中的基因样点区域,为后续求得基因样点的平均灰度值提供了很好的支持.



(a) 分割前的基因芯片图　(b) 迭代阈值分割结果

(c) Ostu分割结果　(d) 最小误差阈值分割结果

图 1　几种分割算法的比较

Fig. 1　Comparison of several segmentation algorithms

2　实验结果及分析

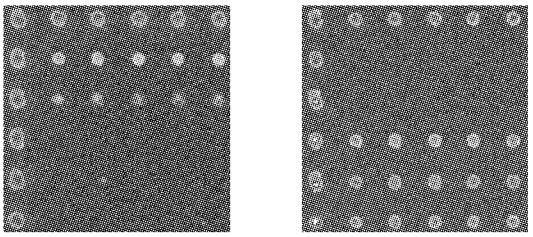
应用上述的最小误差阈值算法,本文构建了一个基因芯片分析体系,主要分为图像预处理、图像识别以及数据提取和分析 3 个步骤. 为了测试基因芯片分析结果,选用凡敏等制备完成的基因芯片(基孔肯亚病毒与辛德毕斯病毒特异性检测基因芯片[10])作为实验样本,具体处理步骤如下.

2.1　图像预处理与识别

通过芯片扫描仪得到的基因芯片图像是彩色图像,如图 2 所示. 为了便于后续处理并且提高图像质量,首先要进行图像预处理,其中包括图像灰度化、自适应中值滤波[11-12]以及适当的对比度增强处理.

图3为图2(a)所示的基因芯片图像预处理之后的结果.

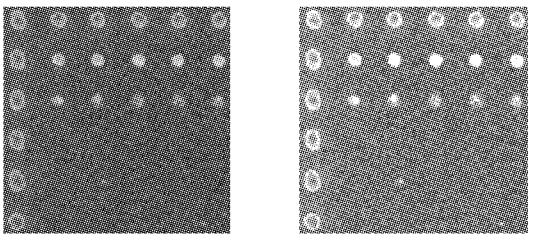
实验得到预处理之后的芯片图像,再经过基因样点网格定位和图像分割2个处理过程之后,就可以把基因芯片图像中的每个基因样点都分离出来.采用基于功率谱的投影网格定位算法[13]以及上面介绍的最小误差阈值分割算法处理芯片图像,得到如图4的结果.



(a)基孔肯亚病毒基因图像 (b)辛德毕斯病毒基因图像

图2 彩色基因芯片图像

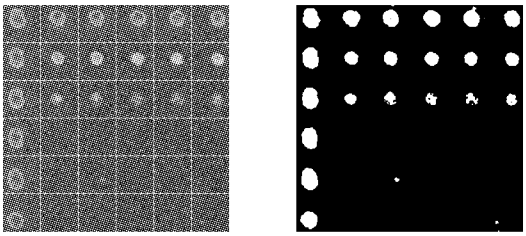
Fig.2 Color cDNA microarray images



(a)滤除噪声的芯片图像 (b)对化度增强的芯片图像

图3 基因芯片图像预处理结果

Fig.3 Pretreatment with cDNA microarray image



(a)网格定位结果 (b)最小误差分割结果

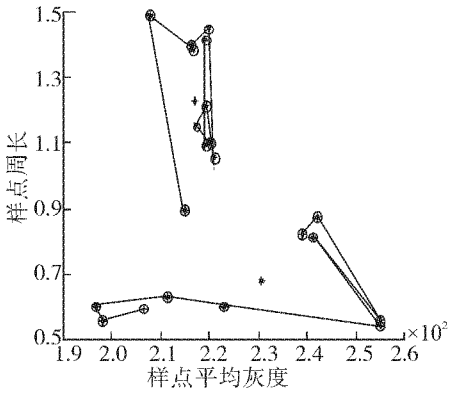
图4 图像识别结果

Fig.4 Image recognition results

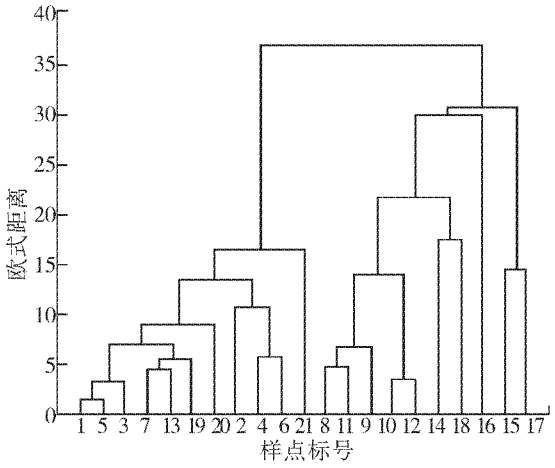
2.2 数据提取和分析

网格定位确定了每个基因样点图像的具体位置,图像分割又把每个小区域内的基因样点与背景成功地分离出来,接下来按照基因样点的分布把图4(b)分割成若干个小图片.

通过观察基因样点的形态和亮度特性,选择应用基因样点的平均灰度、面积、周长以及圆度4个特征参数来描述基因样点.因此,基因样点的图像信息就成功地转化成了数据信息,把这些信息整理成数据集,通过一系列的模糊聚类和层次聚类分析[14]可以成功地把病毒样点和对照样点区分开来.聚类分析的可视化结果如图5所示,其中图5(a)横坐标表示各基因样点的平均灰度,单位为灰阶,纵坐标表示各基因样点的周长,单位是像素点个数;图5(b)横坐标表示样点标号,纵坐标表示各数据集之间的欧式距离.



(a)模糊聚类分析



(b)层次聚类分析

图5 聚类分析结果

Fig.5 Results of cluster analysis

根据最小误差分割算法建立的基因芯片分析体系,能得到符合基因样点信息的分类结果.并且依据现有的一些基因芯片检测图像,按照病毒类别和浓度的不同,把众多基因样点分类得到的结果与已知的基因芯片制备设计的样点分类情况进行比较,即可以计算出测试分类的准确率(正确区分的基因样

点个数/待区分的基因样点总数). 通过多组图片的分类实验,分别计算准确率,最终求得准确率的平均值结果如表 1.

表 1　分类准确率汇总

Table 1　Accuracy summary

病毒名称	病毒与对照样本分类	病毒浓度分类
基孔肯亚	0.928 8	0.669 4
辛德毕斯	0.825 5	0.716 8

从表 1 可以看出,受到基因芯片制备条件和处理算法的影响,每组实验结果的准确率有很大的不同. 总体来看,对于同一组小芯片中不同基因样点的分类结果要好于不同组小芯片中基因样点的分类. 基因样点病毒间差异的分类结果要优于同病毒不同浓度的分类结果. 尤其是针对样点信号和背景图像对比不大的情况,算法的分割效果较好.

3　结束语

上述实验表明,应用最小误差阈值分割算法设计的基因样点识别系统能够成功地把基因芯片中大量的基因样点区分出来,并且计算描述基因样点的特征参量数值. 通过分析这些数据基因芯片系统实现了基因样点分类的功能,但准确率还不是很高,对于基因样点间的细微差别还是无法识别,有待于今后在算法上进一步完善.

同时,由于基因芯片种类较多,制备方法不尽相同,而且基因芯片扫描仪型号各异,因此不同种类基因芯片图像存在很大的差异. 本文基因芯片图像来源有限,很多类型的基因芯片图像还未应用到本文描述的基因芯片识别系统进行分析处理. 今后还需多方搜集实验样本,检测该系统的性能,使其不断完善以及有更广泛的适用范围.

参考文献:

[1] MALEKINEJAD H, SCHOEVERS E J, DAEMEN J, et al. Exposure oocytes to the Fusarium toxins zearalenone and deoxynivalenol causes aneuploidy and abnormal embryo development in pigs[J]. Biology of Reproduction, 2007, 77(5): 840-847.

[2] WIESE K C, EICHER C. Graph drawing tools for bioinformatics research: an overview computer based medical systems[C]//Proceedings of the 19th IEEE Symposium on Computer-Based Medical Systems. Washington, DC, USA: IEEE Computer Society, 2006: 653-658.

[3] 张晶,王黎,高晓蓉,等. 数字图像处理中的图像分割技术及其应用[J]. 信息技术, 2010(11): 33-39.
ZHANG Jing, WANG Li, GAO Xiaorong, et al. The image segmentation technology and its application in digital image processing[J]. Information Technology, 2010(11): 33-39.

[4] CECCARELLI M, ANTONIOL G. A deformable grid-matching approach for microarray images[J]. IEEE Transactions on Image Processing, 2006, 15(10): 3178-3188.

[5] BAJCSY P. An overview of DNA microarray grid alignment and foreground separation approaches[J]. EURASIP Journal on Applied Signal Processing, 2006, 2006(1): 080163.

[6] 王宇,陈殿仁,沈美丽,等. 基于形态学梯度重构和标记提取的分水岭图像分割[J]. 中国图象图形学报, 2008, 13(11): 2176-2180.
WANG Yu, CHEN Dianren, SHEN Meili, et al. Watershed segmentation based on morphological gradient reconstruction and marker extraction[J]. Journal of Image and Graphics, 2008, 13(11): 2176-2180.

[7] 刘华军,任明武,杨静宇. 一种改进的基于模糊聚类的图像分割方法[J]. 中国图象图形学报, 2006, 11(9): 1312-1316.
LIU Huajun, REN Mingwu, YANG Jingyu. An improved image segmentation method based on fuzzy clustering[J]. Journal of Image and Graphics, 2006, 11(9): 1312-1316.

[8] 范九伦,雷博. 二维直线型最小误差阈值分割法[J]. 光电工程, 2009, 31(8): 1801-1806.
FAN Jiulun, LEI Bo. Two-dimensional linear-type minimum error threshold segmentation method[J]. Journal of Electronics & Information Technology, 2009, 31(8): 1801-1806.

[9] 谭优,王泽勇. 图像阈值分割算法实用技术研究与比较[J]. 计算机信息, 2007, 23(24): 233, 298-299.
TAN You, WANG Zeyong. Study on applied technology arithmetic of image threshold segmentation[J]. Microcomputer Information, 2007, 23(24): 233, 298-299.

[10] 凡敏,田明尧,赵权,等. 基孔肯亚病毒和辛德毕斯病毒检测基因芯片的建立[J]. 中国兽医学报, 2012, 32(10): 1493-1497.
FAN Min, TIAN Mingrao, ZHAO Quan, et al. Establishment and application of gene chip for Chikungunya virus and Sindbis virus[J]. Chinese Journal of Veterinary Science, 2012, 32(10): 1493-1497.

[11] 王晓凯,李峰. 改进的自适应中值滤波[J]. 计算机工程与应用, 2010, 46(3): 175-176, 218.
WANG Xiaokai, LI Feng. Improved adaptive median filte-

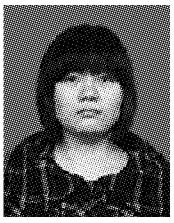
ring[J]. Computer Engineering and Applications, 2010, 46(3): 175-176, 218.

[12]郭海霞,谢凯.一种改进的自适应中值滤波算法[J].中国图象图形学报, 2007, 12(7): 1185-1188.
GUO Haixia, XIE Kai. An improved method of adaptive median filter[J]. Journal of Image and Graphics, 2007, 12(7): 1185-1188.

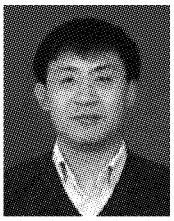
[13]胡园园,孙啸,何农跃,等.基于图像投影的基因芯片图像网格定位[J].生物医学工程学杂志, 2005, 22(4): 668-671.
HU Yuanyuan, SUN Xiao, HE Nongyue, et al. A gene-chip image grid localization method based on profiles of image[J]. Journal of Biomedical Engineering, 2005, 22(4): 668-671.

[14]李明华,刘全,刘忠,等.数据挖掘中聚类算法的新发展[J].计算机应用研究, 2008, 25(1): 13-17.
LI Minghua, LIU Quan, LIU Zhong, et al. New developments of clustering methods in data mining[J]. Application Research of Computers, 2008, 25(1): 13-17.

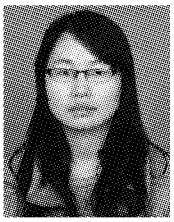
作者简介:



尹宁,女,1985年生,硕士研究生,主要研究方向为图像处理与模式识别.



刘富,男,1968年生,教授,博士生导师,吉林省自动化学会秘书长.主要研究方向为计算机视觉与模式识别、生物信息识别技术等.承担国家重大科技成果转化等项目30余项,获得国家发明专利2项,发表学术论文40余篇.



张玉,女,1989年生,硕士研究生,主要研究方向为图像处理与模式识别.

第8届中国生物识别学术会议(CCBR 2013)

The 8th Chinese Conference on Biometric Recognition (CCBR 2013)

生物识别是模式识别、图像处理、人工智能等学科领域的前沿方向,同时也是保障国家和公共安全的战略高新技术、电子信息产业的新增长点.中国生物识别学术会议从2000年开始在北京、杭州、西安、广州先后成功主办过7届,有力推动了我国生物识别的学科发展和应用推广,同时为国内生物识别学术界和产业界同行提供了一个交流与合作的平台.第8届中国生物识别学术会议(CCBR2013)由山东大学、中国科学院自动化研究所和中国人工智能学会联合主办,将于2013年11月16—17日在济南举行.本届会议向广大科技工作者公开征集优秀学术论文(英文),大会录用的稿件将由Springer出版社的Lecture Notes in Computer Sciences(LNCS)图书系列出版,并被EI和ISTP检索.

重要日期

投稿截止日期:2013年7月5日

录用通知日期:2013年8月20日

会议召开日期:2013年11月16—17日

联系我们

联 系 人:袭肖明

通信地址:山东济南市舜华路中段山东大学计算机学院

电 话:15069056021

邮 箱:ccbr2013@sdu.edu.cn

网 址:http://ccbr2013.sdu.edu.cn