第 38 卷 第 24 期
Vol.38 No.24

计 算 机 工 程
Computer Engineering

2012 年 12 月
December 2012

·开发研究与设计技术· 文章编号：1000—3428(2012)24—0258—04 文献标识码：A 中图分类号：TN911.73

NRDPT：下一代测序数据的处理方法

张 骏[1]，殷 陶[1]，陈玲慧[2]

(1. 上海交通大学计算机科学与工程系生物智能与信息实验室，上海 200240；
2. 台湾交通大学多媒体工程研究所，中国台湾 新竹 30050)

摘 要：随着下一代基因测序技术的发展，其数据处理面临着越来越高的要求和挑战，基因测序软件厂商的基本数据处理程序已经不能满足实验要求。为此，提出一种新的数据处理方法(NRDPT)。在第 3 代基因测序原理的基础上，根据边缘信息和霍夫变换重新设计图像的处理步骤，给出新的基因簇定位算法，并设计两步配准算法。实验结果表明，与传统的直接配准方法相比，该方法能使配准速度提升约 9 倍。

关键词：下一代测序；基因测序；开源软件；图像处理；图像配准；图像分析；信号处理

NRDPT: Processing Approach for Next Generation Sequencing Data

ZHANG Jun[1], YIN Tao[1], CHEN Ling-hwei[2]

(1. Biological Intelligence and Information Lab, Department of Computer Science and Engineering, Shanghai Jiaotong University, Shanghai 200240, China; 2. Institute of Multimedia Engineering, National Chiao Tung University, Hsinchu 30050, China)

【Abstract】With the development of the Next Generation Sequencing(NGS) technology, the default raw data processing pipeline faces more and more challenge. This paper presents a new data processing pipeline——NRDPT. NRDPT designs a new raw data processing algorithm, based on edges and Hough transform. In addition, a high performance 2-step registration algorithm is proposed. The experimental results show that, based on the same precision, the 2-step speeds up the classical algorithm by 9 times.

【Key words】Next Generation Sequencing(NGS); DNA sequencing; open source software; image processing; image alignment; image analysis; signal processing

DOI: 10.3969/j.issn.1000-3428.2012.24.061

1 概述

下一代基因测序技术相对于 Sanger sequencing[1]，使并行化的基因测序技术成为了可能。这种测序方法的原理是将 DNA 切成很多个小片段，然后对这些片段进行分别测序，最后借助计算机将这些片段整合起来[2-4]。该技术在近来得到了显著的发展，如 Genome Analyzer(Illumina, San Diego, USA)、454-FLX(Roche, Basel, Switzerland)和 SOLiD(Applied Biosystems, California, USA)等。此外以 Polonator(Dover System)为代表的测序系统是一种以连接反应进行 DNA 序列分析的技术[5]。基于该技术的测序仪器和反应成本更低，有着广阔的发展空间。

在上述过程中，对于 DNA 片段的每个位置都需要用显微镜和照相机拍出 4 张不同的荧光照片，如果某个 cluster 在某个位置的某张荧光照片中较亮，就可以在一定程度上认为这个 cluster 的 DNA 片段在这个位置上的核苷酸类型是对应的那张照片所识别的。然而在实际的过程中，因为各种干扰的存在[6]，使得从亮度值到达 ATCG 中那个值的这个过程并不是那么容易，已经有很多文章和模型讨论了这个问题[7-8]，并且得到了不错的结果。

随着技术的应用，产生出的 NGS 数据越来越多，诞生了像 BING[9]、Swift[10]等第三方的 NGS 数据处理软件。其中，在 BING 中，为了提高整个测序的通量，并降低图像处理的难度，提出了基于像素的方法，将荧光图中所有的发亮像素点均作为基因簇而进入后续 basecall 步骤，该方法的缺点在于增加了后面步骤的负担，增加了很多冗余信息，并且只支持 Illumina 测序机器所产生的数据。

本文介绍了一个第三方的处理 NGS 原始数据处理系统 NRDPT。相对于 BING 和 Swift，该系统是首个第三方的从 Polonator 类似系统产生原始数据开始进行分析的处理软件。NRDPT 重新设计了图像处理的步骤，采用基于边缘的定位方法，能够有效地定位出基因簇，降低后续处理的数据量。

2 NRDPT 的主要方法

图 1 为本文算法的主要流程。

作者简介：张 骏(1987－)，男，硕士研究生，主研方向：图像处理，机器学习；殷 陶，硕士研究生；陈玲慧，教授

收稿日期：2011-06-23 **修回日期：**2011-11-07 **E-mail：**Fluyd.zh@gmail.com

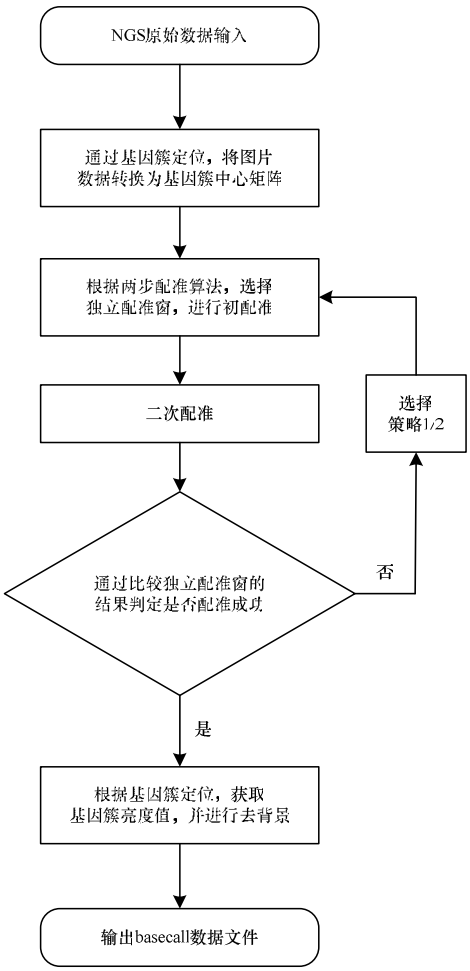


图 1　本文算法流程

本文方法的主要步骤如下:

(1)图像配准

在采集照片的过程中，不同时间采集的照片之间有一定的相对位移，因此，图像配准的准确率将直接影响整个数据处理的成功与否，错误的图像配准结果将直接导致读出的 BEAD 灰度值完全没有意义。

(2)基因簇定位

根据反应原理，每一个基因簇吸附在一个磁珠上，磁珠在荧光图中的表现形式为光斑。本文需要定位每一个基因簇的位置，从而读取到每一个位置的光斑亮度，利用该信息来得到基因序列信息。在图像配准完成后，可以利用配准信息来增强基因簇的定位。

(3)背景去除

在这一部分的处理中，因为不同照片之间采用了不同的曝光时间，其背景亮度不同，所以需要将背景亮度去除。

仔细考量了以上的处理过程后发现，可以用在基因簇定位步骤中得到的基因簇中心矩阵来进行配准。使用标示基因簇中心位置的 0-1 矩阵进行图像的配准工作。首先将参照图和荧光图中的基因簇全部找出来，标记 1 在一个同样大小的全 0 矩阵上，用该矩阵代替灰度图进行配准，这样的配准速度将大大加快，而且可以完成荧光图和参照图之间的配准。此外，还设计了一个两步配准方法，相对传统方法得到了大约 9 倍的效率提升。

这种重复利用中间步骤处理数据的算法大大减少了图像配准所用的时间，并且根据结果部分的展示可以看到，因为利用了 NGS 数据特征，所以该图像配准算法能够更有效地抵抗噪声，准确率也相对较高。

2.1　基因簇定位

2.1.1　参考图处理

参照图中包含了所有的基因簇，而荧光图中只能看到部分基因簇(即在该荧光照射下产生反应的碱基)。对于 2 种不同类型的图片，采取了不同的方法进行分析。

对于参照图，发现 BEAD 的可见形状不规则，但因为采用普通光源的原因，每个 BEAD 周围都有明显的阴影，所以可以采用图像二值化的方法获得 BEAD 的中心区域。关于二值化算法的选择，采用了 Otsu[11]方法，该方法对图像的二维最大类间方差进行优化，自动选取最佳阈值，使得类间的分离性达到最大。

为了对抗曝光不均的影响，将整张图片划分为 $m\times n$ 个小区域，对于每个区域分别求二值结果。对于得到的二值图像，接下来求得所有联通域(如果不是特别说明，本文中提到的联通域均为 8-联通)，计算联通域面积，求得面积在 $p1$ 和 $p2$ 之间的所有联通域(在本文方法中，$p1$ 和 $p2$ 一般取 4 和 40，在 40X 放大倍数下获取的图片 BEAD 直径大概为 8，对应的面积为 50)，进一步求得符合阈值范围的区域的中心坐标，该坐标即近似为 BEAD 的中心坐标。至此，就可以得到标记所有中心位置的矩阵。

2.1.2　荧光图处理

对于荧光图，可以看到基因簇的表现形式跟参照图差异很大，大都表现为规则的圆环或圆斑，而且有着相似的半径。针对这种特征，采取了更为精确的霍夫变换方法[12]来获得圆心，过程见图 2。

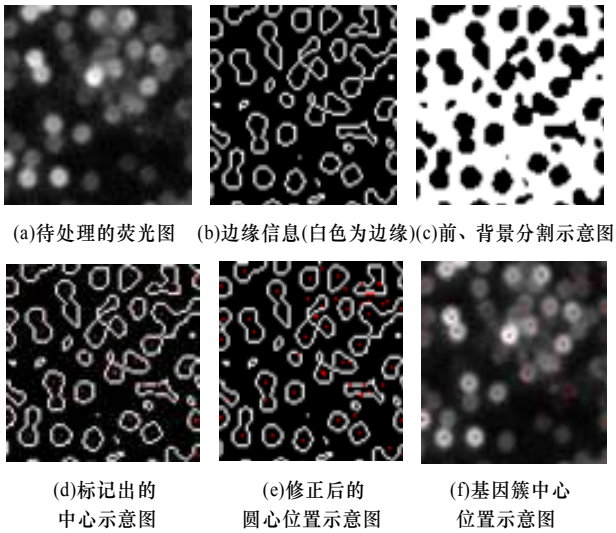


图 2　基因簇定位示意图

霍夫变换是通过将图像坐标中的点转换到目标检测图形(如在本文方法中应用的圆)参数坐标中，在参数参考系中对转换过来的点进行计数，求得区域最大的点，该点

在图像坐标中就对应着待检测图形(相当于用图像坐标中的点进行投票，求得票数最高的待检测图形位置)。霍夫变换的一个副作用在于，因为阈值选择较低，有可能对于同一个圆得到若干个相距较近的圆心，而阈值选择过高的话，又会在图像质量不好的时候失去很多数据，此外，还有可能在背景中获得错误边缘。为了应对霍夫变换的问题，进行了如下 2 点改进：

(1)在霍夫变换的应用中，首先需要得到图片的边缘，在本文所述的方法中，采用了高斯拉普拉斯(LoG)边缘提取算法[13]，它先对图片进行平滑处理，然后通过寻找图像灰度值二阶微分中的过零点来检测边缘(如图 2(b)所示)，保证边缘的闭合性。在图片的闭合边缘结果中，找到它的最大联通域并标记其为背景(如图 2(c)所示)，从而可以避免产生在背景中获得的错误圆心。

(2)为了应对这个缺陷，将霍夫变换得到的圆心矩阵进行 1 个单位的形态学膨胀操作，求得联通域的中心点，将该中心点作为圆心，从而滤去伪圆心的干扰。

2.2 图像配准

在获得了参照图和荧光图中的基因簇中心位置后，即可以进行图像配准。

对于同一个位置，一共有参照图 $\boldsymbol{IM}_{\text{ref}}$ 和荧光图 $\boldsymbol{IM}_{\text{test}}$，对每一个荧光图，采用以下误差矩阵来计算相对位移。其中，$\boldsymbol{IM}_{\text{ref}}$ 和 $\boldsymbol{IM}_{\text{test}}$ 均为 0-1 矩阵，$\boldsymbol{E}_{\text{SAD}}$ 的大小为 $K\times K$，K 是允许测试的最大偏移：

$$\boldsymbol{E}_{\text{SAD}}(x_0,y_0)=\sum_{x=1,y=1}^{M,N}\frac{\boldsymbol{IM}_{\text{ref}}(v_{x,y})-\boldsymbol{IM}_{\text{test}}(v_{x+x_0,y+y_0})}{M\times N} \tag{1}$$

其中，(x_0,y_0)为荧光图相对参照图的位移；M、N 是图片的长和宽。

对于结果准确性的判定，采用以下方法，同时提取图片上 BEAD 数量最高的 2 个同样大小的区域，对其进行分别配准，如果 2 次配准结果相差小于 p，那么就认为该配准是成功的，否则，有以下 2 种策略可以选择(可通过参数选择)：

(1)继续加大配准的位移范围，增大配准窗的面积再进行一次，直到配准成功或者到达设定的最大值，返回配准失败。

(2)再选择一个独立的配准窗进行一次配准，如果结果与之前 2 次之间的相差小于 p，那么就认为这次配准是成功的，否则可以采取策略(1)。

根据对基因簇荧光图的本身性质分析，设计了一个两步的图像配准过程，有效提高了配准过程的效率。

在荧光图片中，绝大多数的基因簇之间并没有出现层叠情况，所以，可以认为 2 个 BEAD 中心之间的距离大于等于基因簇的直径，这种系数矩阵之间的配准，如果按照一般的配准方法，实际上增加了很多不必要的计算。

步骤 1

首先降低基因簇中心矩阵的分辨率，对于每一个基因簇的中心，将其坐标除以 d(因为基因簇的半径为 4 像素左右，为了增加容错性，一般将 d 取为 3)：

$$\boldsymbol{IM}'(x',y')=\boldsymbol{IM}(x'\times d,y'\times d) \tag{2}$$

这样可以将图片面积降低为原来的 1/(d^2)，使用这个图片 $\boldsymbol{IM}'$ 进行位移量的搜索。

步骤 2

因为之前将坐标都除以了 d，所以在得到了位移之后，首先需要将得到的位移值乘以 d 以获得在原图 $\boldsymbol{IM}$ 中的大概位移：

$$(\tilde{x}_t,\tilde{y}_t)=(x_t'\times d,y_t'\times d) \tag{3}$$

在此时得到的 $(\tilde{x}_t,\tilde{y}_t)$ 基础上，再对 $\boldsymbol{IM}_{\text{test}}$ 和 $\boldsymbol{IM}_{\text{ref}}$ 的 X、Y 方向进行范围为[$-d$, d]的配准(式(1))，将得到的偏移量加入 $(\tilde{x}_t,\tilde{y}_t)$，从而得到精确的偏移值 (x_t,y_t)：

$$(x_t,y_t)=(\tilde{x}_t+\Delta_x,\tilde{y}_t+\Delta_y) \tag{4}$$

下面分析以上改进算法的时间复杂度，容易得到：

$$O\left(\left(\frac{M}{d}\right)^2\times\left(2\times\frac{s}{d}+1\right)+M^2\times(2m+1)^2\right) \tag{5}$$

而传统配准算法时间复杂度为：

$$O\left(M^2\times(2\times S+1)^2\right) \tag{6}$$

其中，S 为一个方向上的偏移，即在上文中提到的 K=2×S+1，M 是边长。

当 d=3 时，代入式(5)和式(6)进行推导，可以得出算法效率提高了 80 倍左右(矩阵运算中的每步操作均算为单元操作)，但是在实际实现中，对于矩阵运算有整体加速效果，即在当前应用下可以忽略每次矩阵运算中大小的影响，这样可以得到实际的效率提高应该在 9 倍左右(80/9)，随后的实验部分也证明了这点。

图像配准之后，可以用配准的图像中的信息来增强 BEAD 定位的结果。

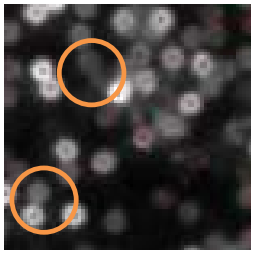
2.3 基因簇位置信息的增强

在此时得到的基因簇定位信息中，可能有如下误差来源：(1)基因簇中心位置不准确；(2)噪声，该位置不存在基因簇。

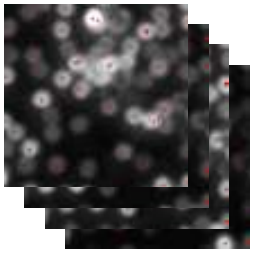
对于每一个基因簇中心，在一次拥有 t 个周期的实验中，产生了 4t 张图，理论上，在每个周期的 4 张图中，该基因簇中心在荧光图中将会至少出现一次(它总归会是某个碱基)。那么可以假设：一个基因簇中心在 4t 张图片中理论上至少会出现 t 次。

基于以上事实，用经过配准的 4t 张图片来对于某个 BEAD 的位置进行投票操作。将所有的基因簇中心矩阵叠加起来(式(7))，然后求得结果中的 8-联通域，对于每一个联通域，求得其最大位置坐标，继而得到该位置为圆心的统计次数(图 3)：

$$M=\sum_{i=1}^{T}M_{A,i}+M_{T,i}+M_{C,i}+M_{G,i} \tag{7}$$



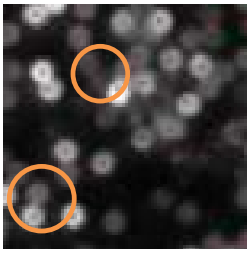
(a)单张结果



(b)增强定位示意图

?	?	?	?	?
?	0	7	8	?
?	1	18	1	?
?	6	1	4	?
?	?	?	?	?

(c)加和矩阵



(d)最终增强结果

图 3　基因簇定位的增强示意图

在图 3 中，图 3(a)使用单张荧光图得到基因簇中心(标记的位置为找到的基因簇中心，圈出的部分中有漏掉的基因簇)。图 3(b)将配准后位置对其的基因簇中心矩阵(0-1 矩阵)加和，得到图 3(c)。图 3(d)求矩阵图 3(c)的局部最大值位置而得到增强的结果，可以看到在图 3(d)中成功地标记出了在图 3(a)中失去的信息。

2.4　亮度值获取和背景去除

在亮度值提取的过程中，会出现以下问题：

(1)基因簇光点的亮度不均匀。

(2)在同一张图片中，不同区域的试剂浓度不同会导致同一张图片上的亮度不均匀。

(3)在不同曝光量下导致图片之间亮度不存在可比性。

在不同循环时，荧光剂量和曝光时间的不同导致的亮度不同，需要算出背景参考值以供碱基识别。而在同张图片中，同时获得被处理基因簇周围(一般取以该基因簇为中心 50 像素单位正方形)的背景均值，在得到基因簇亮度之后将这个背景均值减掉。

3　实验结果与分析

在本文中展示的所有实验运行在一台 Intel 酷睿 2 双核 T7500 2.4 GHz 的 CPU 以及 3 GB 内存的笔记本型电脑上。使用软件 Matlab R2010a。基因数据来自深圳华因康基因科技有限公司的 PSTAR-II 高通量基因测序仪。测序内容为已知的模板序列。

3.1　基因簇定位准确度实验

在这个实验中，选取了一次 8 碱基长度测序中的一个位置的数据，并截取了 200×200 像素大小的测试框，实际记录了可以看见的基因簇数目和程序找到的数目，如表 1 所示，其中，碱基位置为一个 DNA 链上的不同位置。从表 1 可以看到，NRDPT 所采用的基于边缘的基因簇定位算法有着极高的精确度(97.83%)和极低的标准差(0.012 2)。

表 1　基因簇定位精确度

碱基位置	定位基因簇数量	错误定位数量	精确度/(%)	精确度标准差	精确度平均值/(%)
1	147	4	97.30		
2	137	6	95.60		
3	172	2	98.80		
4	171	3	98.20	0.012 2	97.83
5	174	4	97.70		
6	160	5	96.90		
7	168	1	99.40		
8	159	2	98.70		

3.2　图像配准实验

为了验证两步配准方法的有效性，本文设计了它和利用基因簇中心矩阵进行直接配准方法的比较实验。

表 2 为 2 种配准方法的时间性能比较，可以看到，两步配准方法在单张图片的配准中，比传统方法提高了近 13 倍，而在多张图片的测试中(160 张，包含 5 个位置的 8 个碱基长度基因片段实验)提高了 8.67 倍。

表 2　配准方法时间性能比较

方法	单张图片处理	多张图片处理(160)
两步配准方法(1)/s	9.89	903.67
直接方法(2)/s	131.78	7 835.54
(2)/(1)	13.32	8.67

表 3 是配准方法的准确度比较。NRDPT 根据同一张图片上不同位置窗的配准结果之间的差距来判断是否成功配准。对于首次配准失败的图片，首先采用策略(1)进行二次匹配(选择第 3 个配准窗)，如果继续失败，则继续采用策略(2)进行(增大配准窗)。可以看到，尽管两步配准方法在首次配准的成功率上低于传统方法，但是在采用策略(1)后，已经超过了传统方法的成功率。表 3 的最后一列为人工对比最后配准结果的准确度，可以看到，无论是两步配准方法还是传统方法，都达到了 100%的配准成功率。

表 3　配准方法准确度比较　(%)

方法	1 次配准(p=2)	策略(1)	策略(1)+策略(2)	配准准确度
两步配准方法	78.13	99.375	100	100
直接方法	81.88	98.750	100	100

根据以上实验，可以看到 2 种方法在本文设计的策略下均能实现 NGS 数据的完全配准，而其中两步配准法平均比传统方法快了将近 9 倍。

4　结束语

下一代基因测序技术目前的飞速发展令人鼓舞和振奋，可以预见在不远的未来将会大规模地应用于临床医疗领域，而数据处理在其中扮演着重要的角色，如何迅速地传递、存储和处理这大量数据将会成为一个挑战，本文介绍的 NRDPT 是一种新的基于边缘的处理方法，能提高数据处理的抗噪声能力和处理速度，从而增加了基因测序的通量。

(下转第 265 页)