

# DNA 质量筛选算法研究

张 曦,樊晓桢,康继昌,徐 然

(西北工业大学 软件与微电子学院,陕西 西安 710072)

**摘 要:** DNA 测序过程中,每个碱基具有一个质量值,其值都会与合格质量有所偏差。为了保证测序的准确性,偏差过大的碱基须淘汰,于是原来 100 bp 的 DNA 序列根据淘汰碱基的位置而缩短成若干序列片段了。最后选取其中最长的碱基片段,即 DNA 质量筛选。传统质量筛选算法虽然思路简单,但合格碱基过于分散。为了解决这一问题,上海生命科学院介绍了“滑窗算法”,能选取出 60% 以上的碱基片段,大大提高了碱基有效片段的长度。在此基础上,文中提出了“双侧变值滑窗算法”。经测试表明,文中算法能选取出 70% 以上的碱基片段,进一步提高了碱基片段有效长度和 DNA 测序的准确性。

**关键词:** DNA 质量筛选;滑窗算法;碱基有效片段;双侧变值

**中图分类号:** TP301.6

**文献标识码:** A

**文章编号:** 1673-629X(2015)07-0041-04

**doi:** 10.3969/j.issn.1673-629X.2015.07.009

## Research on a Selective Algorithm for DNA Quality

ZHANG Xi, FAN Xiao-ya, KANG Ji-chang, XU Ran

(School of Software & Microelectronics, Northwestern Polytechnical University,  
Xi'an 710072, China)

**Abstract:** In the process of DNA sequencing, each base has a quality value which is less than the standard value. In order to ensure the accuracy of the sequencing, the base whose quality value is much less than the standard value will be picked out. The longest stretches of DNA base sequence will be selected in 100 bp, which is called DNA quality selection. Although the traditional algorithm is easy, the qualified bases will be such fragments as can't be used. In order to solve the problem, an algorithm called slipping window was presented by Shanghai Institute of Life and Science, which can select over 60% base fragments, largely improving the length of the base effective fragments. Based on this, an improved algorithm called value changing in both sides of slipping window is proposed. The test result shows it can select over 70% base fragments, further improving the effective length of base fragments and precision of DNA sequencing.

**Key words:** selection of DNA quality; slipping window algorithm; effective fragments of base; value changing in both sides

## 0 引 言

上海生命科学研究院利用进口的基因检测仪提供了人类碱基部分片段,谋求对人类 DNA 进行更快更有效的测序。在测序过程中,基因检测仪首先将 DNA 序列拆分成大量的长度约为 100 bp 的 READ, 然后进行比对<sup>[1]</sup>。

实际测序中,每一个碱基具有一个质量值,不同的用户根据其需要来设定“质量合格值”。为了保证应用需求<sup>[2-4]</sup>,质量值与“质量合格值”偏差过大的碱基应该淘汰,从而选取出符合要求的最长碱基片段,即为正式比对时的有效片段,称为 DNA 质量筛选<sup>[5]</sup>。因

此,被筛选处理后的 READ 长度应该小于等于 100 bp。根据不同的疾病诊断应用,碱基质量的合格值是不同的。文中以合格值为 53 的唐氏综合症筛选检查<sup>[6-7]</sup>为例来研究碱基质量的筛选算法。

传统的筛选方法,就是通过基因检测仪,得出每一个碱基的质量,然后将这 100 bp 碱基序列的每个碱基质量与合格值依次进行比较。若其质量不小于合格值,则定义为合格可用;反之则定义为不合格而被淘汰<sup>[8-10]</sup>,最终选取 100 bp 碱基中最长的合格片段。

这样的碱基质量筛选方法虽然思路简单,且易于通过软件编程实现,但由于每次只是以一个单位的碱

收稿日期:2014-08-22

修回日期:2014-11-26

网络出版时间:2015-06-23

基金项目:国家“863”高技术发展计划项目(2003AA001018)

作者简介:张 曦(1988-),男,硕士研究生,研究方向为软件工程集成电路设计;樊晓桢,教授,博士生导师,研究方向为布尔逻辑基因比对;康继昌,教授,博士生导师,研究方向为布尔逻辑基因比对。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20150623.1051.043.html>



度的碱基质量值并求和,将所得结果和合格值的  $L$  倍 ( $53 \times L$ ) 相比,若不小于该数值,则视为在这段碱基中第一个非补充碱基的碱基质量为合格,记录合格碱基质量与类型。否则即视为不合格。随后将该序列向后滑动一个 bp。

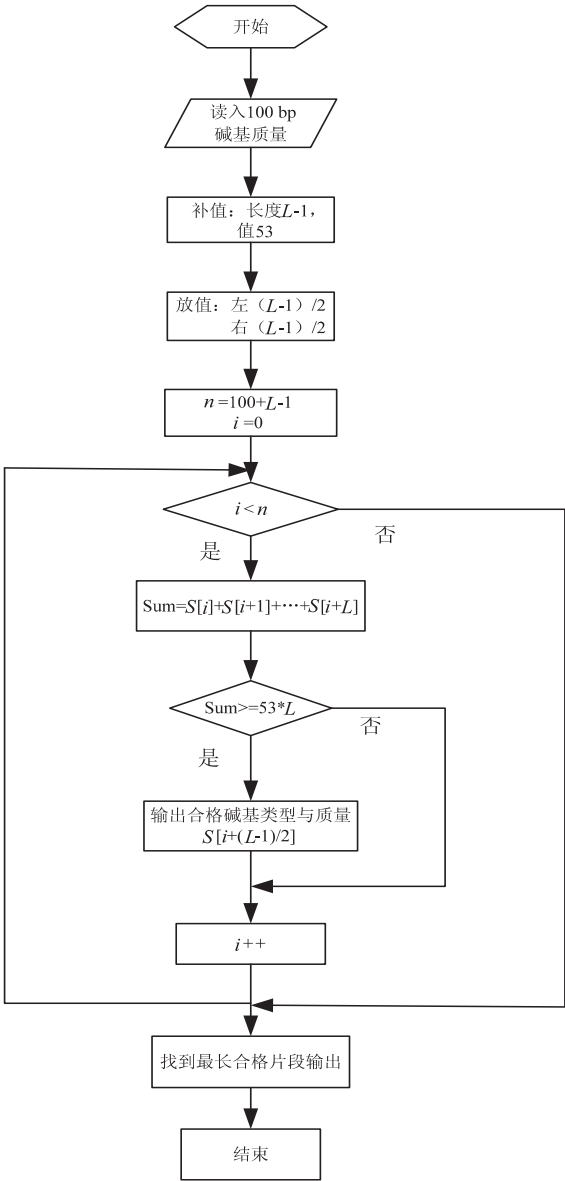


图 2 滑窗算法流程图

(4) 依次循环上述过程,直到碱基片段末尾。将合格的碱基整理并找到最长合格片段,即为 100 bp 中符合要求的碱基片段。

图 3 展示了在取  $L$  长度为 7 的滑窗算法下筛选质量合格碱基的过程。

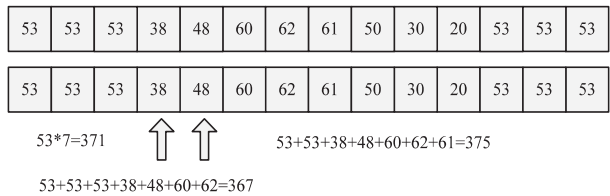


图 3 滑窗算法的质量筛选过程

图 3 中,滑窗算法的质量筛选详细过程为:  
第一步:在原本 8 bp 的碱基序列中进行补值,补值长度为 6,补值内容是碱基的合格质量值 53。再将这 6 个补充的质量值左右各半放置在原来碱基序列的两侧,形成新的 14 bp 的碱基序列,如图中第一行所示;

第二步:从新序列的开头算起,依次取 7 个碱基质量值求和,即:

$53 + 53 + 53 + 38 + 48 + 60 + 62 = 367$

$367 < 53 \times 7 = 371$ ,也就是说,补充后的质量总值小于合格碱基的质量总值,则视为第一个碱基质量 38 不合格。此时向右滑动一个单位,再依次取 7 个碱基质量值求和,即:

$53 + 53 + 38 + 48 + 60 + 62 + 61 = 375$

$375 > 53 \times 7 = 371$ ,也就是说,补充后的质量总值大于合格碱基的质量总值,则视为第二个碱基质量 48 合格。此时再向右滑动一个单位。

依次进行上述过程,直至碱基片段的末尾,最终找到最长的合格碱基片段就完成了质量筛选的过程。

1.3 性能分析与对比

滑窗算法优于传统的碱基质量筛选算法之处在于,它回避掉了由于单个碱基与合格值比较产生的合格碱基离散度过大的问题。因为在质量比较过程中加入了补充的碱基,这使得多个碱基的总质量被提高,从而增大了其中单一碱基通过合格值达到合格的概率,所以通过该算法筛选出来的碱基片段远远长于传统算法。这种算法是在忽略较少的质量偏差基础上,保护了 100 bp 碱基中合格碱基的最长完整度,适合于 DNA 测序工作。图 4 为通过滑窗算法后得到的一组 100 bp 碱基质量筛选情况。

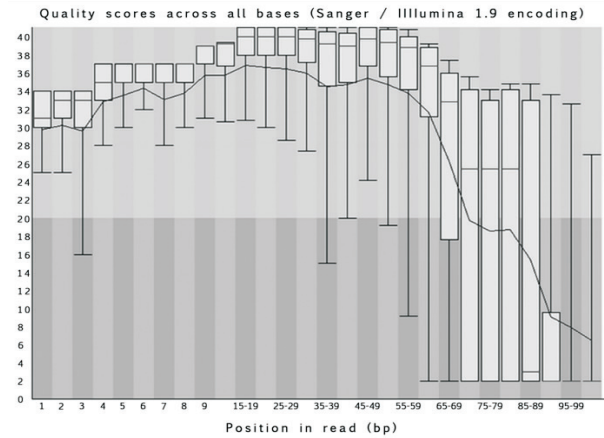


图 4 100 bp 碱基质量筛选后结果

图中横坐标表示每一个碱基在 100 bp 碱基中所处的位置,纵坐标表示碱基的质量值。每一段碱基最底端横线表示未进行滑窗算法前原有的质量值,上面的轮廓线表示通过滑窗算法计算后提高的质量值。碱

基质量沿纵坐标方向由上到下分为三个区域,依次表示为碱基质量值较优、一般、较差区域。观察图 4 可以发现,通过滑窗算法后,使较长的碱基片段达到了第一区域,从而实现了最终碱基质量筛选的目标。表 1 为传统碱基质量筛选算法与滑窗算法的对比情况。

表 1 传统碱基质量筛选算法和滑窗算法的结果对比

筛选算法	碱基长度/bp	最长有效片段/bp	筛选时间/s
传统筛选	100	12	1.4
滑窗算法	100	66	3.0

2 滑窗算法进一步的改进

滑窗算法的目的是通过滑窗操作使更多的连续碱基段被采用,从而最大限度地保护了 READ 的连续性。该算法的核心步骤在于补充的合格碱基质量值的放置情况。目前使用滑窗算法时通常是将补充值分别平均放置在 DNA 序列的首尾两边,即将其左右各半放置,这样平衡整个 100 bp 的碱基,使得碱基序列对称,从而方便碱基质量筛选。然而,在实际筛选时,补充值并非一定要左右各半放置在 DNA 序列的两侧。通过观察 Fastq 文件中第二和第四部分,可以大致找到 100 bp 的碱基质量值合格较多、片段长度较长的区域,因此补充值在碱基序列左右放置的个数并不局限,可在质量值较低的区域增加补充值数量以提升筛选通过碱基的概率。即若 100 bp 碱基偏合格部分集中在整体左侧,则将补充部分放置在整体右侧偏多一些,这样可以更有效地平衡 100 bp 碱基质量,从而使更多碱基达到合格标准,反之亦然,称为“双侧变值滑窗算法”。

“双侧变值滑窗算法”的主要思路是放置在 100 bp 碱基序列左右两侧的补充值数量是可变的。根据不同的情况,灵活地补值能继续增加 DNA 序列有效片段长度。表 2 为滑窗算法与改进的双侧变值滑窗算法的对比情况。

表 2 滑窗算法和双侧变值滑窗算法结果对比

筛选算法	碱基长度/bp	最长有效片段/bp	筛选时间/s
滑窗算法	100	60	2.8
补值右侧的算法	100	69	2.8
补值左侧的算法	100	70	3.1

由表 2 可知,根据不同的碱基质量分布,选择改进的变值算法在运行时间相同的前提下,最长有效片段均比原滑窗算法得出片段提高了约 10%,从实验结果上证明了该改进算法可行,但其理论根据涉及概率论等知识,尚不明确,是文中有待深化研究之处。

3 结束语

针对 DNA 测序中碱基质量筛选算法的不足,文中设计和实现了有效可靠的碱基质量筛选算法—滑窗算法、双侧变值滑窗算法。它们均采用在碱基左右两侧补充合格质量值的思路,提高了碱基质量达到合格值的概率,在处理速度差别不大的前提下,大大增加了碱基有效片段的长度。

参考文献:

[1] Chiang J, Studniberg M, Shaw J, et al. Hardware accelerator for genomic sequence alignment[C]//Proceedings of the 28th IEEE EMBS annual international conference. New York, USA:IEEE,2006.

[2] 刘超,马志强,刘帅.生物信息学中的双序列比对算法[J].长春工程学院学报:自然科学版,2006,7(3):55-57.

[3] 王勇献,王正华.生物信息学导论—面向高性能计算的算法与应用[M].北京:清华大学出版社,2011.

[4] Sachdeva V, Kistler M, Speight E, et al. Exploring the viability of the cell broadband engine for bioinformatics applications[C]//Proc of international parallel and distributed processing symposium. [s. l.]:[s. n.],2007.

[5] Krane D E, Raymer M L. 生物信息学概论[M]. 孙啸,陆祖宏,谢建明,译.北京:清华大学出版社,2004.

[6] 徐琳,李晓民,谭光明,等.面向 FPGA 的 RNA 二级结构预测并行算法研究[J].计算机学报,2006,29(2):233-238.

[7] 李方洁,刘希玉,陈洁.基于改进蚁群算法的 DNA 双序列比对[J].南京师大学报:自然科学版,2010,33(4):148-152.

[8] Needleman B, Wunsch C D. A general method applicable to the search for similarities in the amino acid sequence of two proteins[J]. Journal of Molecular Biology, 1970, 48: 443-453.

[9] Kanopoulos N, Hallenbeck J J. A first-in, first-out memory for signal processing applications[J]. IEEE Trans on Circuits and Systems, 1986, 33(5):556-558.

[10] 张阳,窦勇,夏飞.生物信息学双序列比对算法加速器设计与实现[J].计算机科学与探索,2008,2(5):519-528.

[11] Keong W K C, Schmidt B. Parallel DNA sequence alignment on the cell broadband engine[C]//Proc of workshop on parallel computational biology. [s. l.]:[s. n.],2007.

[12] 王进科,冯萍,康继昌,等.基于布尔逻辑的双序列比对协处理器的设计与实现[J].西北工业大学学报,2011,29(1):1-5.

[13] 杨焯,刘娟.第二代测序序列比对方法综述[J].武汉大学学报:理学版,2012,58(5):463-470.

DNA质量筛选算法研究

作者：[张曦](#)，[樊晓桢](#)，[康继昌](#)，[徐然](#)，[ZHANG Xi](#)，[FAN Xiao-ya](#)，[KANG Ji-chang](#)，[XU Ran](#)  
作者单位：[西北工业大学 软件与微电子学院, 陕西 西安, 710072](#)  
刊名：[计算机技术与发展](#)  
英文刊名：[Computer Technology and Development](#)  
年，卷(期)：2015(7)

引用本文格式：[张曦](#).[樊晓桢](#).[康继昌](#).[徐然](#).[ZHANG Xi](#).[FAN Xiao-ya](#).[KANG Ji-chang](#).[XU Ran](#) [DNA质量筛选算法研究](#)  
[期刊论文]-[计算机技术与发展](#) 2015(7)