软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2017,28(10):2640−2653 [doi: 10.13328/j.cnki.jos.005137]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于读分割最优匹配的 indels 识别算法*

王春宇[1], 潘 俊[1], 郭茂祖[2], 刘晓燕[1], 刘 扬[1], 刘国军[1]

[1](哈尔滨工业大学 计算机科学与技术学院,黑龙江 哈尔滨 150001)

[2](北京建筑大学 电气与信息工程学院,北京 100044)

通讯作者: 郭茂祖, E-mail: guomaozu@bucea.edu.cn

摘 要: 高通量测序技术的发展,极大地推动了基因组结构变异识别的研究.当前,该领域主要使用覆盖度、读分割或片段组装方法来识别变异,但目前的方法识别结果不够准确,敏感度高,对基因组结构变异的信息(如变异序列、变异坐标等)挖掘不充分.插入和删除类型的结构变异统称为 indels,在基因组结构变异中最为常见.为此,针对 indels 的精确识别,提出了基于读分割和动态规划的最优序列匹配算法(optimal split-read matching algorithm,简称 OSRM). OSRM 算法能将异常读片段以最少的空位打断比对到参考序列上.首先,建立异常读片段与特定参考序列的匹配得分矩阵;然后,建立回溯路径矩阵;最后,用以变异特点设计的得分公式对每条路径进行最优匹配筛选,输出精确识别的 indels 坐标及序列.实验结果显示,该方法对小中型的 indels 有很高的识别性能.此外,与读分割法的经典算法 Pindel 进行了比较,证实 OSRM 算法在小中型的 indels 识别方面有更好的效果,可识别更复杂的情况.

关键词: 结构变异;拷贝数变异;动态规划;读分割;精确识别

中图法分类号: TP181

中文引用格式: 王春宇,潘俊,郭茂祖,刘晓燕,刘扬,刘国军.基于读分割最优匹配的 indels 识别算法.软件学报,2017,28(10): 2640−2653. http://www.jos.org.cn/1000-9825/5137.htm

英文引用格式: Wang CY, Pan J, Guo MZ, Liu XY, Liu Y, Liu GJ. Indels detection algorithm based on optimal split-read matching. Ruan Jian Xue Bao/Journal of Software, 2017,28(10):2640−2653 (in Chinese). http://www.jos.org.cn/1000-9825/5137.htm

Indels Detection Algorithm Based on Optimal Split-Read Matching

WANG Chun-Yu[1], PAN Jun[1], GUO Mao-Zu[2], LIU Xiao-Yan[1], LIU Yang[1], LIU Guo-Jun[1]

[1](School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

[2](School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing 100044, China)

Abstract: The development of next-generation high-throughput DNA sequencing techniques has greatly promoted the research of structural variations (SVs) detection. Current genetic structure variation detection methods are mainly base on depth of coverage, pair-end mapping clusters, or sequence assembly, some of them are known to be not accurate or too sensitive. What's more, some methods are not able to recognize the specific position and sequence of structural variation. Insertions and deletions (indels) are the most common forms of genome structure variations. This paper puts forward an optimal split-read matching algorithm (OSRM) using dynamic programming. OSRM breaks an abnormal read into several reads in a least quantity. First, a score matrix of the abnormal read and the corresponding referenced sequence is created. Then a matrix of backtracking path is established. Next, a formula designed according to the characteristics of structural variation is used to elect the optimal backtracking path matrix. And finally the split-read and referenced sequence are matched in an optimal arrangement by which the accurate position and sequence of found indels are outputted. Experiments

* 基金项目: 国家自然科学基金(61402132, 61571163, 61532014)
Foundation item: National Natural Science Foundation of China (61402132, 61571163, 61532014)
收稿时间: 2016-06-13; 修改时间: 2016-09-02; 采用时间: 2016-09-28; jos 在线出版时间: 2016-10-11
CNKI 网络优先出版: 2016-10-12 16:26:39, http://www.cnki.net/kcms/detail/11.2560.TP.20161012.1626.010.html

prove that the performance of algorithm is excellent. In addition, compared with Pindel which is the best in split-read methods, OSRM can offset its defection in detecting small and medium indels while also be able to detect more complex situation.

Key words: structural variation; CNV (copy number variants); dynamic programming; split-read; accurate detection

随着国际千人基因组计划的进行,基因组的研究进入了一个新的阶段,推动了基因组遗传多样性的研究.随着第二代测序技术的迅速发展,也为采用计算方法研究基因组学扩展了新疆域[1].遗传变异是基因组遗传多样性研究中非常重要的一部分内容,包括结构变异(structure variation,简称 SV)、拷贝数变异(copy number variants,简称 CNV)[2]、单核苷酸多态性(single-nucleotide polymorphism,简称 SNP)[3]等.

在全基因组关联分析中,SNP 常被用于研究表型特征与基因间的关系[4],以推断出它们在遗传学中的相关性[5].而对于表型的多样性,SV 能提供更多的证据[6].进一步的研究表明,SV 与许多不同类型的人类疾病和动植物表型都有很大的相关性[7],所以 SV 在生物性状的差别中起着非常大的作用.其中,以人类为例,基因组中小型 indels 的数量居于第二,仅少于 SNP 的数量,而这些小型 indels 常发生在基因组中关键的位置,影响人类重要的表型特征和疾病[8].近几年,对于 indels 的研究越来越热门,尤其是小型的 indel,如 2014 年,Zhang 分析了 PAX7 基因上大小为 31bp 的 indel 多态性与鸡的性状特征的关联性[9],Lyu 分析了 KLF15 基因一个 2bp 的 indel 与鸡的生长以及躯体的特征关联性[10].2016 年,Shi 等人研究了在 SMAD3 基因上的一个 17bp 的 indel 会改变牛的转录水平而影响表型特征[11].另外,Zhang 等人研究了在 DGAT2 基因上一个 13bp 的 indel 多态性与猪的背膘厚度以及瘦肉率具有关联性[12].

基因组结构变异类型繁多,主要包括删除(deletion)、插入(insertion)、倒位(inversion)、重复(duplication)、移位(translocation)等多种形式.基于测序的基因组结构变异检测方法目前主要有 4 种,包括读对法(read-pair)[13]、读深法(read-depth)[14,15]、读分割法(split-read)以及片段组装法(sequence assembly)[16].

本文主要研究基因组结构变异中的 indels,是除 SNP 之外发生最多的变异种类.因读分割法能够识别出更准确的变异信息,所以本文提出了 OSRM(optimal split-read matching algorithm)算法.该算法利用动态规划策略,以打断数最少的方式将读片段(read)比对到指定范围的参考序列来确定变异.动态规划的优势在于可以降低序列比对时的敏感性,不易受到序列复杂情况的影响.相对于 Pindel[17]模式匹配串算法进行单个碱基的查找匹配,虽精确但是匹配方式过于严格,会忽略某些情况下的变异,且其不能识别复杂变异结构.为测试 ORSM 的算法性能,与读分割方法中的经典算法 Pindel 进行了比较,结果表明:本文算法在识别小型的 indels 上具有更大优势,可以弥补 Pindel 的缺陷.

1 读分割方法介绍以及相关概念

1.1 高通量测序

高通量测序技术,又称为下一代或者第二代测序技术,是对传统桑格测序技术的一次变革,能同时对几十万甚至几百万条 DNA 片段进行序列测定.如图 1 所示:将待测样本的 DNA 完整序列打断成 300bp~800bp 的片段,然后读取这些片段一端或者两端的短序列,其长度根据平台不同可达几十至上百 bp 左右[18].

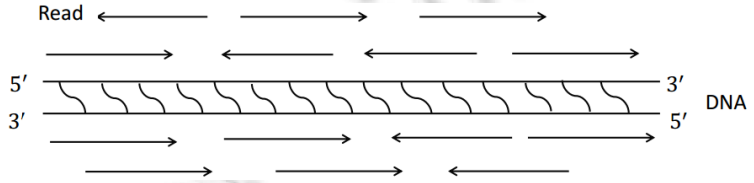


Fig.1 High-Throughput sequencing technology

图 1 高通量测序技术

1.2 全基因组重测序

全基因组重测序是对已知基因组序列的物种进行不同个体的基因组测序,如图 2 所示:将重测序样品的

read 比对到该生物的参考序列上,识别出它们之间的差异,在此基础上对个体或群体进行差异性分析,并寻找与重要性状相关的基因差异或者基因变异[19].

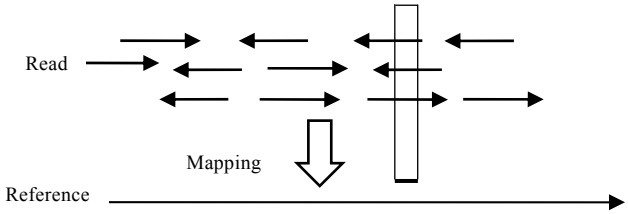


Fig.2 Whole genome sequencing

图 2 全基因组重测序

1.3 Read以及pair-end read

Read 是在高通量测序中 DNA 片段两端被仪器读取的短序列.根据平台技术的不同,当仪器可测得同一条 DNA 片段两端的两个 read 时,称它们为 pair-end read,如图 3 所示.其中,*表示未知序列.同时,这种测序称为双端测序.本文算法主要利用 pair-end read 数据,因为这样成对的 pair-end read 之间的距离是已知的.

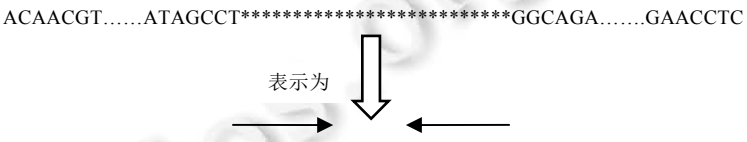


Fig.3 Pair-End read diagram

图 3 Pair-End read 示意图

1.4 Split read

在基因组重测序后,再将 pair-end read 序列数据比对到参考基因组序列上,如图 4 所示,若某些 read 上发生了 indels 或其他变异,将导致 read 不能完整匹配,只能分割成多个片段分别匹配到参考序列上.于是,这些 read 成为读分割,它的匹配结果是包含空位的[20].

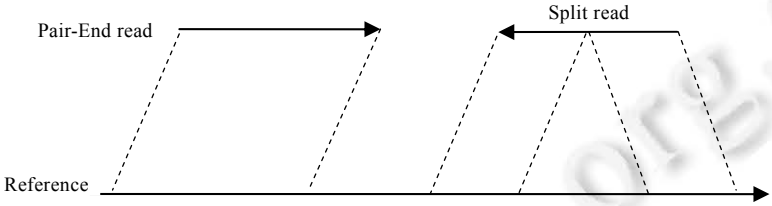


Fig.4 Split read

图 4 Split read

2 相关工作与分析

2.1 短语及符号说明

- One end mapped pair:表示比对到参考序列之后,只有一端 read 能正常匹配,而另外一端 read 不能够正常匹配的 pair-end read;
- Anchor read:表示 one end mapped pair 中匹配上的那条 read,这条 read 可以作为计算变异的坐标参考点,并且能够确定变异发生的方向;
- Insert size:表示建库时 DNA 打断的片段长度,也是 pair-end read 加上中间未知片段的长度;
- Basic set:表示碱基的字符集,由 A,T,C,G,N 组成,其中,A,T,C,G 代表 4 种碱基,N 表示未测得的碱基.

令 S 表示读分割序列,$|S|$表示 S 的长度,$S[i]$为 S 中第 i 个字符($0<i\leqslant|S|$);令 R 表示参考基因组序列,则用二元组(i,j)表示 S 的第 i 个字符与 R 的第 j 个字符匹配;$(i,-)$和$(-,j)$分别表示对应的字符与空格匹配.$\sigma(x,y)$为字符 x 与字符 y 匹配的得分函数;$F(i,j)$为 S 前缀 $S[1,\ldots,i]$与 R 前缀 $R[1,\ldots,j]$的最佳相似性得分;而 $F[i][j]$为得分矩阵,其单元由 $F(i,j)$决定.$T[i][j]$为回溯矩阵的元素,用来记录回溯路径;回溯矩阵中,用特定字符表示 $T[i][j]$可移动到的位置,具体为:

- '\','|'和'–'分别表示 $T[i-1][j-1]$,$T[i-1][j]$和 $T[i][j-1]$;
- '%'表示 $T[i][j-1]$或 $T[i-1][j]$或 $T[i-1][j-1]$;
- '@'表示 $T[i][j-1]$或 $T[i-1][j-1]$;
- '$'表示 $T[i-1][j]$或 $T[i-1][j-1]$.

2.2 Indel结构变异的识别

读分割方法识别 indel 结构变异的过程如下,先从 pair-end read 中挑出不能正常比对到参考序列的 one end mapped pair,pair-end read 上的序列是由 basic set 构成,将其异常匹配的一端打断一处或者是几处变成 split-read,然后比对到参考序列,再利用 one end mapped pair 中的 anchor read 作为参考点,并依据该 one end mapped pair 的 insert size 计算出变异片段发生的位置和方向.

2.3 现有算法的问题

本文主要研究的内容是基于读分割思想的基因组变异结构识别算法.Pindel 是目前主要的 indels 检测方法,它利用 Pattern Growth 算法进行序列局部比对查找,这种方法能够达到单碱基对的识别精度.但是,由于该算法会对单个碱基进行一一查找比对,虽然精确度非常高,但同时敏感度会受到很大的影响.因为生物序列自身特点和测序技术本身的限制,一个 read 可能不仅发生一种变异.若一个 pair-end read 上发生了两个删除变异或删除与 SNP 同时发生,那么 Pindel 精确查找方式无法处理.具体的例子在第 4.2 节的实验中有详细分析.此外,根据 Pindel 的算法特点,其更适应于查找大型的删除与中等尺寸的插入变异,而不是专注于小型 indels 的发现.

本文基于经典的动态序列比对算法——Needleman-Wunsch 算法[21]和 Smith-Waterman 算法[22]进行了改进,结合读分割的思想,提出了一种以最优的得分形式将异常的 pair-end read 比对到参考序列,从而动态识别出 read 中发生 indels 的算法.

2.4 经典序列比对算法分析

本文根据 indels 识别问题,改进了双序列比对经典算法 Needleman-Wunsch 和 Smith-Waterman,在改进的方法中,同样采用两者的动态规划方法,但将两者的全局比对思想与局部比对思想相结合.

Needleman-Wunsch 算法在匹配时允许出现空位和错配,得到两个长度相等的序列,对其中匹配的字符进行加分,对空位及错配字符进行罚分,最终得到的比对序列是计分最高的全局最优序列.匹配的罚分函数定义为 $\sigma(x,y)$,如公式(1)所示.

$$\begin{cases}\sigma(i,-)=-1\\ \sigma(-,j)=-1\\ \ldots\\ \sigma(i,j)=\begin{cases}1, & \text{if } S[i]=R[i]\\ -1, & \text{if } S[i]\neq R[i]\end{cases}\end{cases} \tag{1}$$

得分函数 $F(i,j)$的定义如公式(2)所示.

$$F(i,j)=\max\{F(i-1,j-1)+\sigma(i,j),F(i-1,j)+\sigma(-,j),F(i,j-1)+\sigma(i,-)\} \tag{2}$$

Smith-Waterman 算法采用的也是动态规划的思想,所以与 Needleman-Wunsch 算法一样匹配得分函数 $\sigma(x,y)$.$F(i,j)$的定义如公式(3)所示.

$$F(i,j)=\max\{F(i-1,j-1)+\sigma(i,j),F(i-1,j)+\sigma(-,j),F(i,j-1)+\sigma(i,-),0\} \tag{3}$$

结构变异识别主要是为了找出 read 与参考序列的差别,而 indels 是 read 与参考序列比对时发生的插入和

删除差异,这正符合双序列比对的特点,也是本文基于两种经典比对算法进行改进的初衷.但经典比对算法是为求出全局与局部的最大相似片段,这与需要查找基因序列变异片段的目的有一定的区别.

首先,对于 Needleman-Wunsch 算法,read 会完整地映射到参考序列.由于 Needleman-Wunsch 算法的特点,在回溯的时候,会优先将回溯路径上能够匹配的字符或者是短序列先进行匹配,所以 read 序列最右边的字符会比对到参考序列靠后的位置上.一般情况下,read 的长度达到几十或者几百时,因为参考序列都会很长,即使能完全匹配,也会出现分散匹配到参考序列的现象.而 Smith-Waterman 算法只寻找最相似的子串,在有 indels 发生时,将 read 打断成多个短序列映射到参考序列,短序列会根据在局部最大的相似度比对到参考序列.

3 结构变异识别算法

3.1 数据预处理

在检测结构变异时,首先要获得异常匹配的 read 作为源数据.可以将 pair-end read 映射到参考基因组序列,当一端发生异常匹配时,另一端可以作为提供坐标参考点以及方向,能够缩小搜索的空间范围,也便于更精确、快速地找到变异发生的位置.首先使用比对工具将 pair-end read 映射到参考基因组序列上,本文使用 BWA 软件,BWA 的高效性适合海量的高通量测序数据,并支持带有空隙和有限错配的序列比对方法.由于高通量测序技术的限制,测序原始数据中的错误不可避免,因此需要使用支持错配与空隙的比对工具,本文算法也专门针对这些问题进行了设计.映射完之后,得到 SAM/BAM 文件,其格式如下所示,只需根据其中第 2 个参数挑选出异常的 read 作为源数据[23].

@1002 105 1 1003 37 100M 1453 550 CTCCAGACTACCTGCGAGTTGTGCGTCCAGTGCAT
GGTTCTATTGACGTACCTAGGTTTAGGGTGCCCCAGAACTGGAACAGGAGAGCACCGTTCCCTGG
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
XT:A:U NM:i:0 SM:i:37 AM:i:37 X0:i:1 X1:i:0 XM:i:0 XO:i:0 XG:i:0

3.2 建立得分矩阵

首先要初始化矩阵,与 Smith-Waterman 算法一样,建立大小为$(|S|+1)\times(|R|+1)$的得分矩阵 $F[i][j]$,将首行与首列初始化为 0,表示字符与空格的比对,即 $F[i][0]=F[0][j]=0$.然后再由 $F(i,j)$填充矩阵,如公式(4)所示.

$$F(i,j)=\begin{cases}F(i-1,j-1)+1, & \text{if } S[i]=R[i]\\ \max\{F(i-1,j),F(i-1,j-1),F(i,j-1)\}, & \text{if } S[i]\neq R[i]\end{cases} \tag{4}$$

此时不设置$\sigma(x,y)$罚分函数,也可认为$\sigma(x,y)$都为 0,意思是不对空位和错配罚分.在填充矩阵时,只统计相同的字符.若以序列 ATCCGAC 和 GATCGTACGACCC 为例,可得如表 1 所示得分矩阵.

Table 1 OSRM algorithm score of matrix

表 1 OSRM 算法得分矩阵

		A	T	C	C	G	A	C
	0	0	0	0	0	0	0	0
G	0	0	0	0	0	1	1	1
A	0	1	1	1	1	1	2	2
T	0	1	2	2	2	2	2	2
C	0	1	2	3	3	3	3	3
G	0	1	2	3	3	4	4	4
T	0	1	2	3	3	4	4	4
A	0	1	2	3	3	4	5	5
C	0	1	2	3	4	4	5	6
G	0	1	2	3	4	5	5	6
A	0	1	2	3	4	5	6	6
C	0	1	2	3	4	5	6	7
C	0	1	2	3	4	5	6	7
C	0	1	2	3	4	5	6	7

3.3 建立回溯矩阵

算法的重点在于回溯方法,之所以不直接使用经典算法,是因为它们在回溯时都默认选择了同一特点的路径.从上述的得分矩阵可以看出,最优的路径应该是从得分最高的元素开始回溯,这与 Smith-Waterman 算法一样.由于在基因变异识别时,理想情况下是将 read 完全比对到参考序列上,所以回溯仍然要从 $F[|S|][|R|]$开始到 $F[0][0]$.路径回溯时,会有多条分支,主要是由 $F[i][j]$在 $S[i]$与 $R[i]$相等的单元处,经典算法中,只要遇到相等的元素即匹配.实际上,一些情况下可先不匹配,取决于 $F[i][j]$的上和左的格子是否与左上相等:如果相等,就可以往该方向回溯.所以,在填充回溯矩阵时,出现多个分支也要保留,以便后续筛选出最优.定义回溯矩阵为 $T[i][j]$,具体如公式(5)、公式(6)所示.

- 当 $S[i]$等于 $R[i]$时:

$$T[i][j]=\begin{cases}'\%', \text{ if } F[i-1][j-1]+1=F[i][j-1] \text{ and } F[i-1][j-1]+1=F[i-1][j]\\ '@', \text{ if } F[i-1][j-1]+1=F[i][j-1] \text{ and } F[i-1][j-1]+1\neq F[i-1][j]\\ '\$', \text{ if } F[i-1][j-1]+1=F[i][j-1] \text{ and } F[i-1][j-1]+1=F[i-1][j]\\ '\backslash', \text{ if } F[i-1][j-1]+1=F[i][j-1] \text{ and } F[i-1][j-1]+1\neq F[i-1][j]\end{cases} \tag{5}$$

- 当 $S[i]$不等于 $R[i]$时:

$$T[i][j]=\begin{cases}'\backslash', \text{ if } \max\{F[i-1][j],F[i-1][j-1],F[i][j-1]\}=F[i-1][j-1]\\ '|', \text{ if } \max\{F[i-1][j],F[i-1][j-1],F[i][j-1]\}=F[i-1][j]\\ '-', \text{ if } \max\{F[i-1][j],F[i-1][j-1],F[i][j-1]\}=F[i][j-1]\end{cases} \tag{6}$$

通过以上公式,可以将上述的得分矩阵转化为表 2 所示的回溯矩阵.

Table 2 OSRM algorithm score of matrix (I)

表 2 OSRM 算法回溯矩阵(I)

		A	T	C	C	G	A	C
	n							
G	|	\	\	\	\	\		
A	|	\					\	
T	|	|	\					\
C	|	|	|	\	@			@
G	|	|	|	|	\	\		
T	|	|	$	|	\	|	\	\
A	|	$	|	|	\	|	\	
C	|	|	|	$	\		|	\
G	|	|	|	|	|	\		|
A	|	$	|	|	|	|	\	-
C	|	|	|	$	$	|	|	\
C	|	|	|	$	$	|	|	$
C	|	|	|	$	$	|	|	$

具体回溯的多条路径见表 3.可以很明显地看出:存在 3 条回溯路径,都能使得序列比对上.将这些多条路径都打印出来,序列比对结果分别如情形(1)~情形(3)所示.

–	A	T	C	–	–	–	C	G	A	C	–	–
G	A	T	C	G	T	A	C	G	A	C	C	C

情形(1)

–	A	T	C	–	–	–	C	G	A	–	C	–
G	A	T	C	G	T	A	C	G	A	C	C	C

情形(2)

–	A	T	C	–	–	–	C	G	A	–	–	C
G	A	T	C	G	T	A	C	G	A	C	C	C

情形(3)

Table 3 OSRM algorithm score of matrix (II)

表 3 OSRM 算法回溯矩阵(II)

		A	T	C	C	G	A	C
	0	0	0	0	0	0	0	0
G	0	0	0	0	0	1	1	1
A	0	1	1	1	1	1	2	2
T	0	1	2	2	2	2	2	2
C	0	1	2	3	3	3	3	3
G	0	1	2	3	3	4	4	4
T	0	1	2	3	3	4	4	4
A	0	1	2	3	3	4	5	5
C	0	1	2	3	4	4	5	6
G	0	1	2	3	4	5	5	6
A	0	1	2	3	4	5	6	6
C	0	1	2	3	4	5	6	7
C	0	1	2	3	4	5	6	7
C	0	1	2	3	4	5	6	7

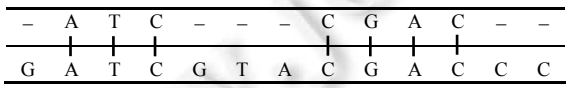
这些结果是所有能一一匹配上的路径,都已被识别出来,只是一些字符的位置有所不同.而且可以看出:其中一条正是 Needleman-Wunsch 算法得到的回溯路径,但是并不是预期的结果.之后,对这些比对结果进行筛选.

3.4 最优比对结果筛选

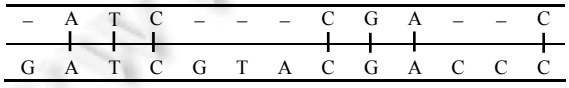
读分割思想本身就是要分割 read 将其比对到参考序列上去,在 Pindel 算法中,由于它是从 read 的两端开始向中间寻找最长以及最短的唯一子串,所以它只能识别出 read 上发生的一个变异.也就说,只能分割一个空位去映射.当然,一些情况下也有可能 read 发生两个或者两个以上的小型 indels,但是由于生物发生的变异本身就是很小的概率,所以在查找变异时,尽量以最少的分割数比对到参考序列上.

首先,需要遍历当前的每条路径.对于路径的遍历,可以把此问题看作是图的遍历,采取邻接表的结构进行存储以及遍历.首先,从矩阵的最后一个单元开始回溯,每到一个新矩阵单元,就给它赋予一个唯一的 ID 号,作为邻接表上的节点 ID 标识,其 ID 号的值为$(i+1)\times(|S|+1)+j+1$,其中的$|S|+1$ 即为矩阵的列数,最终得到一个如图 5 所示的邻接表结构.

从最后一个节点开始,对邻接表递归地根据图的深度优先的顺序将所有路径进行遍历.为了缩短算法的运行时间,无需遍历所有路径,遍历过程中直接对 read 打断的空位进行筛选.即算法预设最大的连续空格数,一般设连读空格数阈值为大于等于 3.算法需指定 read 上允许发生的最多删除变异数.当算法找到 read 中空位数量大于预设值时,即连续的空位数量大于阈值,便停止遍历当前路径.那么,筛选结果如下所示.



筛选结果(1)



筛选结果(2)

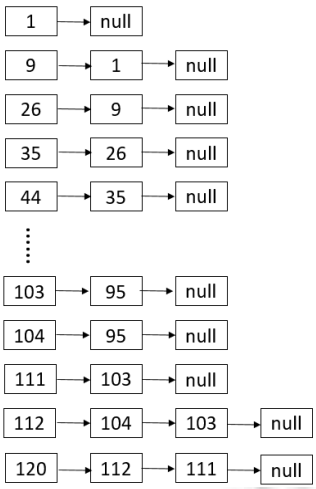


Fig.5 Adjacency list

图 5 邻接表

最后,利用得分公式(7)筛选最优结果:定义读分割分成 m 段,每一段序列中的碱基个数为 $S_i(0\leqslant i\leqslant m-1)$,读分割上连续空格段个数为 b,参考序列上的连续空格数是 r;之后,对于每一个比对结果进行计算,如公式(7)所示.

$$\frac{\sum_{0}^{m-1}2^{s_i}}{b\times r} \tag{7}$$

取最大得分者为最优比对结果,公式(7)中的分母 $b\times r$ 在算法中可保证打断数最少,分子可让读分割上的碱基排序更紧密,尽量保证读分割的完整性.因此,筛选结果(1)的得分为 8,而筛选结果(2)的得分为 17/3,因此,筛选结果(1)为最优结果.

–	A	T	C	–	–	–	C	G	A	C	–	–
G	A	T	C	G	T	A	C	G	A	C	C	C

3.5 ORSM算法伪代码

算法. 基于读分割方法的智能最优序列匹配基因变异算法.

输入:单端异常 mapping 的双末端序列集 $RP(RP_1,RP_2,\ldots,RP_n)$,参考基因序列集 REF,搜索范围为 $l_{ref}(L_{\min}<l_{ref}<L_{\max}$,其中,$L_{\min}$ 和 $L_{\max}$ 为欲查找的片段);

输出:每个读分割在参考基因序列上的比对结果.

1. 定义单端异常 mapping 的 pair-end read 映射对为 $URP(URP_1,URP_2,\ldots,URP_n)$
2. 定义 URP_i 结构体为 $URP_i(i_m,l_m,s_m,i_u,l_u,s_u)$
3. 定义回溯矩阵为 $F[i][j]$,定义 read 长度为 l_{ref}
4. 构建一个 SV 集合,用来存储多个读分割映射到的同一个位置的变异点信息
5. For 所有 URP_i 以及对应的长度为 l_{ref} refleng 的参考序列 REF_i
6. 初始化得分矩阵即 $F[i][0]=F[0][j]=0$
7. For i=0 to l_{ref}
8. For j=0 to l_{ref}
9. 利用得分递归公式填充得分矩阵
10. 同时利用递归公式填写回溯矩阵
11. End for

12. End for
13. 对回溯矩阵上的路径的点进行标识,并建立邻接表进行存储
14. 对邻接表进行深度遍历
15. For 每一条路径
16. 累积其连续空位段数
17. If 连续空位段数大于给定的阈值
18. Continue
19. If 该路径上 read 的连续空位段数最小
20. 保存该路径对应的比对结果
21. Else
22. Continue
23. End for
24. For 结果集的每个比对结果
25. 利用筛选公式进行得分计算,记录比对结果中的得分为最大值
26. End For
27. 根据异常对正常映射的一端对找到的变异结果进行坐标的计算,以及序列的提取,并且保存到变异结构体中
28. End for
29. 保存 smap 中的坐标与序列到 sv 结构体中

3.6 变异识别结果提取

在找到了最佳的序列比对之后,还需要将其变异发生的信息进行提取,利用 pair-end reads 的正常映射的一端可以得到变异端发生的方向,并且利用 Insert size 以及 read 两端序列的长度,可以计算出变异发生的具体坐标,如图 6 所示.

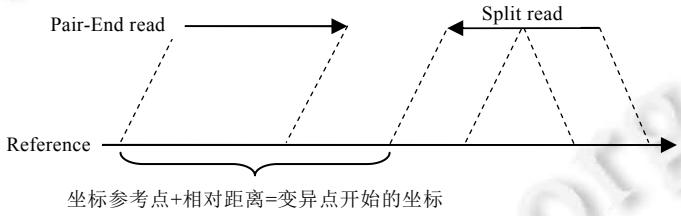


Fig.6 Coordinate calculation of deletions variation

图 6 删除变异坐标计算

插入变异同理可求:在确定完 read 上的变异之后,由于 read 以一定的覆盖率进行重复覆盖,所以还需对这些同一个位点的变异进行聚集,将变异分类到同一个位点上,位点上发生的变异覆盖率越高,说明该变异的可信度越高.

3.7 Pair-End read未知区域的deletions识别算法

另外,本文利用 OSRM 算法,针对 pair-end read 之间的变异识别方法进行了设计.由于测序的技术限制,pair-end read 中间的序列片段是未知的,当变异发生在此区域时,OSRM 算法只能识别出此处发生了变异,而无法求出具体的序列以及坐标,如图 7 所示.

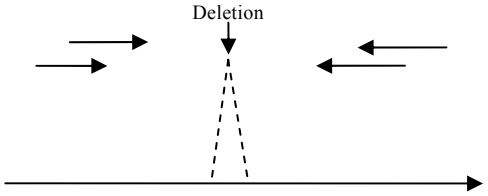


Fig.7 Deletion of pair-end read's unknown area

图 7 Pair-End read 未知区域的删除

由于读分割法的限制,即 read 打断的断点位置不能在边缘(一般离边缘点不能小于 10bp 左右),所以一些情况下,图 8 所示的两个 pair-end read 之间的异常 read 也可能无法准确识别此删除变异.为此,采用本文算法可进一步识别这种变异.

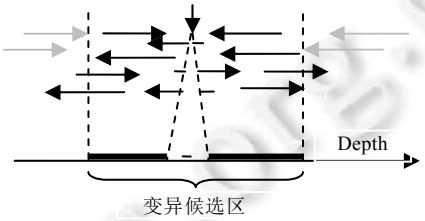


Fig.8 Split-Reads generated from deletion

图 8 删除发生时产生的读分割

首先,将两个 pair-end read 之间的最小公共未知区域作为变异存在的候选区;之后,利用 mapping 工具找到候选区域上的那些异常和正常的 read,利用本文的动态规划算法进行映射;然后,求出候选区的每个碱基的覆盖深度,求出覆盖深度变化最大的两个点,确定为基因变异的片段,这是因为发生变异的区域不能正常比对,而非变异区的深度仍然保持很高,如图 8 所示.

3.8 Pair-End read中删除变异识别算法

算法. 基于读分割法的 pair-end read 未知区域的删除变异识别算法.

输入:OSRM 算法识别后的结果中,异常端能完整比对到参考序列上而坐标超过插入长度允许范围内的双末端序列集(RP_1,RP_2,…,RP_n),即,图 8 中灰色的双末端读对以及参考基因序列集 *REF*;

输出:pair-end read 未知区域的是否为删除变异的判定结果以及变异发生的坐标、序列等信息.

1. 定义 pair-end read 为 $URP(URP_1,URP_2,\ldots,URP_n)$
2. 定义 URP_i 结构体 $URP_i(i_m,l_m,s_m,i_u,l_u,s_u)$,其中,$i_m$,$l_m$,$s_m$ 为正常映射端的变量,分别为坐标、长度以及序列;i_u,l_u,s_u 为异常映射端的变量
3. 对于每一个 URP_1
4. 取该变异簇中发生变异区域长度最小的区域作为候选变异区域,开始点和结束点为 i_m 和 i_u
5. For 原始 read 集上的每个 read
6. If i_m<read 左端坐标 or i_u<read 右端
7. 利用本文最优匹配算法将其以最优方式比对到参考序列上
8. For 比对得到的结果中每个 read
9. 求得比对上的碱基的坐标
10. 将变异候选区相应范围的碱基深度加 1
11. End For
12. For i from i_m to i_u
13. 求得深度变化最大的两个点,作为变异的边界点

14. End For

15. 保存两点的坐标和两点之间的序列到 sv 结构体中

3.9 OSRM算法复杂度分析

在建立矩阵时,需要建立二维矩阵.设定 read 的序列长度为 M,参考序列的长度为 N,则需要 $O(MN)$的空间复杂度.对于哈希表的建立,首先,链表的个数是根据回溯路径上的节点个数来决定的,而节点个数最多的时候是整个矩阵上的所有节点,即 MN 个,而对于链表的节点个数最多为 3 个,所以空间复杂度为 $O(3MN)$,因而总的空间复杂度为 $O(MN)+O(3MN)=O(MN)$,与 Needleman-Wunsch 算法以及 Smith-Waterman 算法的空间复杂度一样.

4 算法性能与实验结果分析

4.1 实验结果

为了检测本算法的性能以及识别变异的表现,本文实验利用模拟数据生成测序片段,根据现有的对人体 x 染色体的变异信息,在人类 21 号染色体上随机放置了插入和缺失的变异,并且随机置入一定几率的 SNP 和测序错误率,本算法未涉及到其他类型的变异检测,所以并未生成其他类型的变异.如表 4、表 5 所示.

读分割中能够具体输出变异序列以及坐标的表现最好的是 Pindel 算法,所以本文算法主要与 Pindel 算法进行对比,对于删除的识别实验结果见表 6.

Table 4 Deletions of simulation data

表 4 模拟删除变异数据

类型	<10bp	10~50bp	50~100	100~1000	1000~100000
Deletions	1 000	500	300	150	50

Table 5 Insertions of simulation data

表 5 模拟插入变异数据

类型	2~5bp	5~10bp	10~15bp	15~25bp	25~30bp
Insertions	500	400	300	200	100

Table 6 Results of the deletions

表 6 删除识别实验结果

类型	<10bp	10~50bp	50~100	100~1000	1000~100000
OSRM	921	458	269	2	0
Pindel	803	397	238	127	45

对于删除的识别准确率展示如图 9 所示.

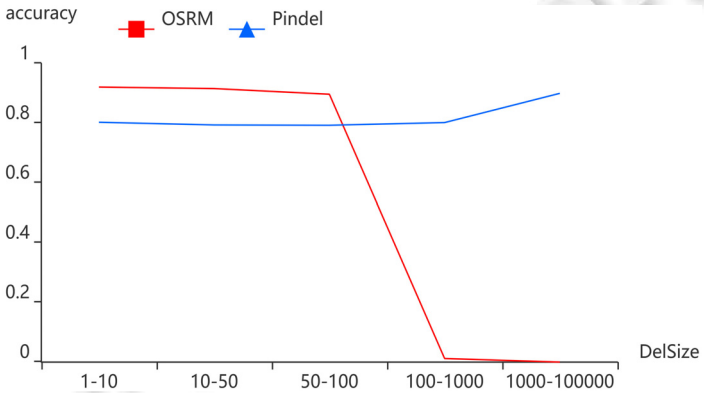


Fig.9 Accuracy of deletions detection

图 9 删除识别准确率

可以明显地看出:在小于 100bp 左右的小中型删除变异识别中,本文算法的识别数量更高;而对于更大的尺寸,由于本文的算法时间复杂度过高,效率低,不进行识别.

对于插入变异的识别结果见表 7.

Table 7 Results of the insertion

表 7 插入识别实验结果

类型	2~5bp	5~10bp	10~15bp	15~20bp	20~25bp
OSRM	445	256	221	130	12
Pindel	432	234	260	104	3

对于插入识别准确率如图 10 所示.

可以看到:本算法插入的识别率与 Pindel 算法不相上下,都是小中型的插入识别准确度较高.

由于本算法在大型的删除识别复杂度过高,为了测试本算法的性能,对不同尺寸的删除变异进行平均搜索的时间进行了测试,如图 11 所示.可以看到:当变异尺寸大于 100bp 时,本算法的时间消耗急剧升高,这也是本算法的不足之处.

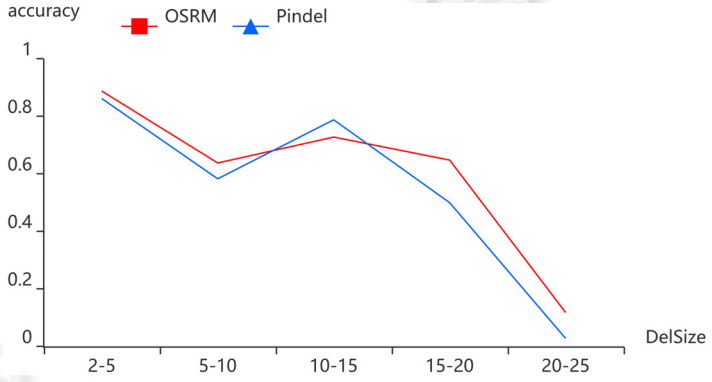


Fig.10 Accuracy of insertion detection

图 10 Insertion 识别准确率

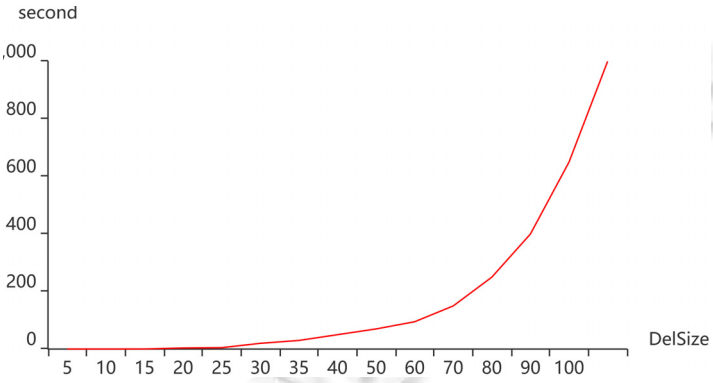


Fig.11 Speed of OSRM

图 11 OSRM 识别速度

4.2 实验结果分析

综上,我们对实验结果进行了分析,本文能够识别更为复杂的情况,即,在一个 read 上发生 1 个以上的删除或者插入的情况.这种情况发生得较少,但本算法也能够精确地识别.相对于 Pindel 算法来说,由于 Pindel 算法的特

点,从 read 的两端开始寻找最大和最小的唯一子串,使其一次只能识别 read 上的一个变异,所以不能识别出一些复杂的事件,虽然这些事件发生的几率很小.并且测试发现:本文算法可以发现 read 上发生的超过一个 indels 的复杂事件.另外,本算法的错配优势在识别 indels 的同时能够允许有 SNP 的存在,图 12、图 13 中的圆圈表示了 SNP 的存在.本算法能够识别出图 12 和图 13 所示的事件,更复杂的情况本算法也能识别.但是这些事件发生的概率很小,这里不再展示.

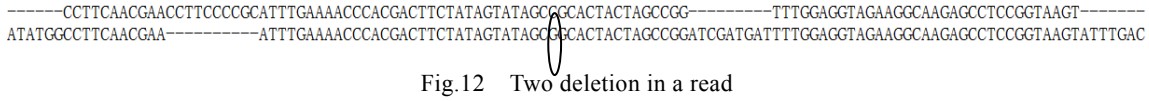


Fig.12 Two deletion in a read

图 12 一个 read 发生两个删除

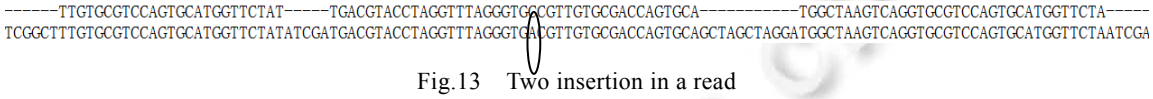


Fig.13 Two insertion in a read

图 13 一个 read 发生两个插入

5 总 结

本文分析了经典的序列比对算法,并且基于基因组结构变异识别方法领域中的读分割法,提出了一种最优序列的匹配方法.将异常 read 分裂比对到参考序列上,智能匹配识别出变异的片段.并且在模拟数据上测试了算法的性能,对识别变异的准确度进行了测试,并且与 Pindel 算法进行了比较.实验结果表明:本算法更适应于小中型的 indels,并且还能识别更复杂的事件.另外,本算法的不足之处在于:当搜索的变异尺寸过大时,算法时间复杂度过高,大大降低了搜索效率.后续的工作若能降低算法的复杂度,将会极大地提高算法的识别性能.

References:

[1] Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, Seifert M, Borodina T, Soldatov A, Parkhomchuk D, Schmidt D. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science, 2008, 321(5891):956−960. [doi: 10.1126/science.1160342]

[2] Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, Teague B, Alkan C, Antonacci F, Haugen E. Mapping and sequencing of structural variation from eight human genomes. Nature, 2008,453(7191):56−64. [doi: 10.1038/nature06862]

[3] McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, Wysoker A, Shapero MH, de Bakker PI, Maller JB, Kirby A, Elliott AL. Integrated detection and population-genetic analysis of SNPs and copy number variation. Nature Genetics, 2008,40(10): 1166−1174. [doi: 10.1038/ng.238]

[4] Cao J, Schneeberger K, Ossowski S, Günther T, Bender S, Fitz J, Koenig D, Lanz C, Stegle O, Lippert C, Wang X. Whole-Genome sequencing of multiple Arabidopsis Thaliana populations. Nature Genetics, 2011,43(10):956−963. [doi: 10.1038/ng.911]

[5] Platt A, Horton M, Huang YS, Li Y, Anastasio AE, Mulyati NW, Ågren J, Bossdorf O, Byers D, Donohue K, Dunning M. The scale of population structure in Arabidopsis Thaliana. PLoS Genet, 2010,6(2):e1000843. [doi: 10.1371/journal.pgen.1000843]

[6] Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L, Taillon BE. Paired-End mapping reveals extensive structural variation in the human genome. Science, 2007,318(5849):420−426. [doi: 10.1126/science.1149504]

[7] Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, Yamrom B, Yoon S, Krasnitz A, Kendall J, Leotta A. Strong association of De Novo copy number mutations with autism. Science, 2007,316(5823):445−449. [doi: 10.1126/science.1138659]

[8] Mullaney JM, Mills RE, Pittard WS, Devine SE. Small insertions and deletions (INDELs) in human genomes. Human Molecular Genetics, 2010,19(R2):R131−R136. https://dx.doi.org/10.1093hmgddq400

[9] Zhang S, Han RL, Gao ZY, Zhu SK, Tian YD, Sun GR, Kang XT. A novel 31-bp indel in the paired box 7 (PAX7) gene is associated with chicken performance traits. British Poultry Science, 2014,55(1):31−36. [doi: 10.1080/00071668.2013.860215]

[10] Lyu SJ, Tian YD, Wang SH, Han RL, Mei XX, Kang XT. A novel 2-bp indel within Krüppel-like factor 15 gene (KLF15) and its associations with chicken growth and carcass traits. British Poultry Science, 2014,55(4):427−434. [doi: 10.1080/00071668.2014.921886]

[11] Shi T, Peng W, Yan J, Cai H, Lan X, Lei C, Bai Y, Chen H. A novel 17 bp indel in the SMAD3 gene alters transcription level. Archives Animal Breeding, 2016,59(1):151−157. [doi: 10.5194/aab-59-151-2016]

[12] Zang L, Wang Y, Sun B, Zhang X, Yang C, Kang L, Zhao Z, Jiang Y. Identification of a 13bp indel polymorphism in the 3'-UTR of DGAT2 gene associated with backfat thickness and lean percentage in pigs. Gene, 2016,576(2):729−733. [doi: 10.1016/j.gene.2015.09.047]

[13] Korbel JO, Abyzov A, Mu XJ, Carriero N, Cayting P, Zhang Z, Snyder M, Gerstein MB. PEMer: A computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. Genome Biology, 2009,10(2):1. [doi: 10.1186/gb-2009-10-2-r23]

[14] Yoon S, Xuan Z, Makarov V, Ye K, Sebat J. Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Research, 2009,19(9):1586−1592. [doi: 10.1101/gr.092981.109]

[15] Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Research, 2011,21(6):974−984. [doi: 10.1101/gr.114876.110]

[16] Gnerre S, MacCallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S, Berlin AM. High-Quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc. of the National Academy of Sciences, 2011,108(4):1513−1518. [doi: 10.1073/pnas.1017351108]

[17] Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: A pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics, 2009,25(21):2865−2871. [doi: 10.1093/bioinformatics/btp394]

[18] Mardis ER. The impact of next-generation sequencing technology on genetics. Trends in Genetics, 2008,24(3):133−141. [doi: 10.1016/j.tig.2007.12.007]

[19] Ng PC, Kirkness EF. Whole genome sequencing. Methods in Molecular Biology, 2010,628:215−226.

[20] Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. Nature Reviews Genetics, 2011,12(5):363−376. [doi: 10.1038/nrg2958]

[21] Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology, 1970,48(3):443−453. [doi: 10.1016/0022-2836(70)90057-4]

[22] Smith TF, Waterman MS. Identification of common molecular subsequences. Journal of Molecular Biology, 1981,147(1):195−197. [doi: 10.1016/0022-2836(81)90087-5]

[23] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The sequence alignment/map format and SAMtools. Bioinformatics, 2009,25(16):2078−2079. [doi: 10.1093/bioinformatics/btp352]

王春宇(1979－),男,辽宁宽甸人,博士,副教授,CCF 专业会员,主要研究领域为计算生物学,机器学习.

刘晓燕(1963－),女,博士,副教授,主要研究领域为计算生物学.

潘俊(1991－),男,硕士,主要研究领域为计算生物学.

刘扬(1976－),男,博士,副教授,CCF 专业会员,主要研究领域为机器学习,计算生物学.

郭茂祖(1966－),男,博士,教授,博士生导师,CCF 专业会员,主要研究领域为生物信息学,机器学习.

刘国军(1979－),男,博士,讲师,CCF 专业会员,主要研究领域为机器学习,计算生物学.