|  |
| --- |
| **第1页** |

ltrharvest

|  |
| --- |
| **第2页** |

1引言

本文档介绍了LTRharvest，这是一种用于从头预测基因组序列中的LTR逆向转运的软件工具。LTRharvest计算潜在的LTR反向边界的边界位置。转座子位于基因组靶序列的持久索引结构上，即增强后缀数组。通常，基因组目标序列是FASTA中的完整染色体DNA序列，每个（多个）FASTA格式的DNA序列都可以传递给软件。为了进行预测，LTRharvest实现了几个过滤器。这些被连续应用于序列数据以拒绝不符合序列，长度或距离特征的候选对象LTR逆转座子。由于这些功能大部分是特定于物种的，因此可以切换每个过滤器开启或关闭，并且对于某个LTR反转录转座子模型的参数化是免费的。

LTRharvest用C编写，它基于GenomeTools库。LTRharvest被一部分单个名为gt的二进制文件调用。LTRharvest在增强的后缀数组索引上运行，该索引已存储在文件上。该索引需要由程序后缀程序构造，该程序后缀程序也是GenomeTools二进制gt。

2用法

2.1 LTRharvest的选项

由于LTRharvest是gt的一部分，因此LTRharvest的调用如下。

gt ltrharvest -index indexname [options]

所有可能选项都有简短的一行描述。表1中给出了option选项。所有选项只能指定一次。

2.2输入选项

-index indexname

指定增强后缀数组索引的名称。索引必须包含表suf，lcp，des和tis。如何使用程序Suffixerator可以在第3节中找到。

|  |
| --- |
| **第3页** |

表1：按类别分类的LTRharvest选项概述输入选项

-index指定增强后缀数组索引的名称

-range在输入序列中指定要搜索LTR逆转座子候选物的范围

输出选项

-out以多种FASTA格式为预测指定输出文件名

-gff3以GFF3格式指定用于预测的输出文件名

-outinner为预测内部区域的多个FASTA文件指定outputfilename

筛选选项

-seed指定最小的种子长度，以进行精确的最大重复

-minlenltr 指定每个LTR的最小长度

-maxlenltr 指定每个LTR的最大长度

-mindistltr指定LTR起始位置的最小距离

-maxdistltr指定LTR起始位置的最大距离

-similar在[1..100％]范围内指定相似度阈值

-mintsd 指定每个TSD的最小长度

-maxtsd 指定每个TSD的最大长度（需要-mintsd）

-motif 指定回文基序，该基序由2个核苷酸startmotif + 2个核苷酸endmotif组成，

-motifmis指定主题中最大不匹配数（需要-motif）

-vic指定核苷酸数（在左侧和右侧）将在5'和3'边界附近搜索TSD和/或图案

-overlaps指定否|最佳|全部对齐方式

-xdrop 为种子扩展指定xdrop值

-mat 指定用于种子扩展名的matchscore

-mis 为种子扩展指定不匹配分数

-ins 指定用于种子扩展的insertscore

-del 为种子扩展指定deletescore杂项选项

-v详细模式

-longoutput附加主题/ TSD输出（需要-mintsd或-motif）

|  |
| --- |
| **第4页** |

选项-index是必需的。

-range x y

在输入序列中指定要搜索LTR反转录转座子候选物的范围。

也就是说，如果x = 1000和y = 10000，则仅当候选人在排名1000之后开始时才被报告，并且在各自的序列坐标中位于位置10000之前。如果x = y = 0（默认值），则搜索整个序列。

2.3输出选项

结果以表格形式在stdout上报告，并且可以使用表示法轻松地写入文件

> resultfile，如下所示：

gt ltrharvest -index indexname [options] > resultfile

另外，以下选项将额外的信息写入文件

-out outputfile

指定将要写入预测的文件的名称。每个预测将是由单个FASTA条目表示。

-outinner outputfile

指定内部区域（预测顺序的文件名LTR序列）将被写入。每个预测将由一个单独的FASTA表示。

-gff3 outputfile

指定将要写入预测的文件的名称。每个预测将是由单个GFF3条目表示。

2.4过滤器选项

过滤器选项提供了排除不需要的序列，长度或距离功能。如果用户未设置特定选项，则该选项的默认值为设置，但选项 -mintsd，-maxtsd和-motif除外。因此，如果未设置任何过滤器选项。用户将进行LTR逆转座子的预测，而无需搜索目标位点重复（TSD）或特定的LTR起始-终止基序。

-seed Lex

指定确切最大重复次数的最小长度。只有那些重复的规格确定的最小长度在寻找候选对的过程中进行分析。精确最大长度低于该阈值的重复将不再考虑。Lex必须为正整数。如果用户未选择此选项，则默认将Lex设置为30。

-minlenltr Lmin

指定每个LTR的最小长度。Lmin必须指定为正整数。如果这用户未选择该选项，则Lmin默认设置为100。

|  |
| --- |
| **第5页** |

-maxlenltr Lmax

指定每个LTR的最大长度。Lmax必须指定为正整数。如果这用户未选择该选项，则默认情况下Lmax设置为1000。

-mindistltr Dmin

指定LTR起始位置的最小距离。必须将Dmin指定为正数整数。如果用户未选择此选项，则默认情况下Dmin设置为1000。

-maxdistltr Dmax

指定LTR起始位置的最大距离。必须将Dmax指定为正数整数。如果用户未选择此选项，则默认情况下Dmax设置为15000。

-similar similaritythreshold

指定两个LTR之间的最小相似度值。论点相似度阈值必须从[0,100]范围内选择，并且表示百分比。如果未选择此选项用户将默认值设置为85％。

-mintsd TSDminlen

如果选择此选项，将使用以下命令搜索目标站点重复项（TSD）

TSDminlen的最小TSD长度。如果用户未选择此选项，则搜索将执行最小TSD长度为4的TSD。如果设置了此选项，但没有最大值TSD长度由-maxtsd选项指定，然后将最大TSD长度设置为20。

-maxtsd TSDmaxlen

此选项需要选项-mintsd。如果选择此选项，则搜索目标站点重复（TSD）将以最大TSD长度TSDmaxlen进行。

-motif expr

在开始处指定2个核苷酸作为起始基序，并指定2个核苷酸作为结束基序和每个LTR的结尾。仅回基序序列等于其互补序列，可向后读取允许使用，例如tgca键入核苷酸，没有任何空格将它们分开。如果没有选择此选项，则将不检查候选对是否包含主题。如果设置了此选项，但是-motifmis选项未指定允许的不匹配数，然后搜索将会进行精确的图案设计。

-motifmis n

该选项需要选项-motif。指定允许的不匹配数论据n。如果未设置此选项，则将搜索确切的图案。这非负整数n必须从[0,3]范围中选择。

-vic l

指定左侧和右侧（附近）的核苷酸位置的数目，这将在LTR的5'和3'预测边界附近搜索TSD和一个基序反转录转座子。此选项只有一个作用，如果选择-mintsd或选项-motif是切换到的，如果用户未选择此选项，则默认值为60。

|  |
| --- |
| **6页** |

-overlaps nojbestjall

指定有关嵌套和重叠的LTR反转录转座子预测的输出。如果选择参数no，则不会在其中报告嵌套或重叠的预测输出。如果选择了最佳参数，则在两个或多个嵌套或重叠的情况下预测，仅是其间具有最高相似性的LTR反转录转座子预测将被报告。如果参数都被选中，那么所有LTR反转录转座子预测是否存在嵌套和重叠的预测。如果这个选项未由用户选择，则默认设置参数为best的选项。

2.5对齐选项

X-drop扩展过程允许搜索退化的LTR。对齐选项有助于提供控制最大重复种子扩展的机会。如果有特定的对齐选项不是由用户设置的，此选项的默认值将被设置。

-xdrop X

指定xdrop值X，以在两个方向上扩展种子重复以实现匹配，不匹配，插入和删除。参数X必须为正整数或0。只要涉及匹配，不匹配，插入注释和删除的分数小于T-X，其中T表示看到的最大分数，如果用户未选择此选项，则默认情况下X设置为5。

-mat score

为X-drop扩展过程指定正匹配分数。如果未选择该选项由用户设置，默认值为2。

-mis score

为X-drop扩展过程指定负的不匹配得分。如果不是这个选项由用户选择，默认值为-2。

-ins score

为X-drop扩展过程指定否定插入得分。如果不是这个选项由用户选择，默认值为-3。

-del score

为X-drop扩展过程指定否定的删除分数。如果不是这个选项

由用户选择，默认值为-3。

2.6其他选项

-v

此选项启用详细模式。在运行期间的处理将被打印到标准输出。这包括一长串已开启的清单或关闭选项。

|  |
| --- |
| **第7页** |

-longoutput

此选项还会将有关检测到的TSD和/或主题的信息打印到stdout，如果用户已经选择了对TSD和/或主题的搜索。此选项需要选项-mintsd和/或-motif。

-help

LTRharvest将在stdout上显示所有选项的摘要，并以退出代码0终止。

3. 例子

在本节中，将介绍LTRharvest的示例应用程序。在小节3.1中，给出了使用LTRharvest的不同选项的信息。然后，第3.2小节举例说明了LTR逆转座子在整个酿酒酵母基因组中的定位。在3.3小节中，显示了LTRharvest输出的聚类过程。此步骤不属于LTRharvest由Vmatch进行。

3.1使用LTRharvest的不同选项

作为目标序列文件，我们使用某些FASTA文件chr02.19970727.fsa.gz，其中包含S. cerevisiae基因组序列，第2号染色体。请注意，该文件是gzip格式的压缩文件（因为.gz的结尾）。此格式可以由程序Suffixerator处理。首先，我们创建增强的后缀数组。我们使用选项-tis，-suf和-lcp ， -des, -ssp和-sds，因为LTRharvest需要相应的表。当我们处理DNA序列时我们指定-ify -dna。

$ gt suffixerator -db chr02.19970727.fsa.gz -indexname chr02.19970727.fsa -tis

-suf -lcp -des -ssp -sds -dna

# dna=yes

# indexname="chr02.19970727.fsa"

.

.

.

# TIME overall 1.37

现在我们可以将索引用于LTRharvest。第一个示例调用将仅使用默认参数用于过滤器和对齐选项，而无需搜索TSD或LTR起始结束基序。

$ gt ltrharvest -index chr02.19970727.fsa

# args=-index chr02.19970727.fsa

# predictions are reported in the following way

# s(ret) e(ret) l(ret) s(lLTR) e(lLTR) l(lLTR) s(rLTR) e(rLTR) l(rLTR)

sim(LTRs) seq-nr

# where:

# s = starting position

# e = ending position

# l = length

# ret = LTR-retrotransposon

6

# lLTR = left LTR

# rLTR = right LTR

# sim = similarity

# seq-nr = sequence number

259532 265448 5917 259532 259863 332 265117 265448 332 99.40 0

427672 430021 2350 427672 428170 499 429522 430021 500 91.60 0

29632 35597 5966 29632 29969 338 35259 35597 339 98.82 0

220989 226919 5931 220989 221339 351 226574 226919 346 97.15 0

每条非注释行表示一个LTR后退,整个LTR逆转座子的起始和结束位置进行转座子预测，左侧LTR实例和正确的LTR实例。此外，对于这些元素中的每一个，报告了相应的元素长度以及两个LTR的相似度百分比。最后每行整数表示输入序列号，LTR反转录转座子预测发生在。输入序列号从0开始计数。使用可选参数-v调用LTRharvest可提供有关enabled和disabled选项以及有关增强的后缀数组索引和时间/空间的其他信息。此外，将选项-out和-outinner指定为运行结果的两个FASTA文件。

$ gt ltrharvest -index chr02.19970727.fsa -v -out pred-chr02.fsa

-outinner pred-inner-chr02.fsa

# args=-index chr02.19970727.fsa -v -out pred-chr02.fsa -outinner

pred-inner-chr02.fsa

# user defined options and values:

# verbosemode: On

# indexname: chr02.19970727.fsa

# outputfile: pred-chr02.fsa

# outputfile inner region: pred-inner-chr02.fsa

# xdropbelowscore: 5

# similaritythreshold: 85.00

# minseedlength: 30

# matchscore: 2

# mismatchscore: -2

# insertionscore: -3

# deletionscore: -3

# minLTRlength: 100

# maxLTRlength: 1000

# minLTRdistance: 1000

# maxLTRdistance: 15000

# overlaps: best

# minTSDlength: 0

# maxTSDlength: 20

# palindromic motif:

# motifmismatchesallowed: 4

# vicinity: 60 nt

# predictions are reported in the following way

# s(ret) e(ret) l(ret) s(lLTR) e(lLTR) l(lLTR) s(rLTR) e(rLTR) l(rLTR) sim(LTRs)

seq-nr

# where:

# s = starting position

# e = ending position

# l = length

# ret = LTR-retrotransposon

# lLTR = left LTR

# rLTR = right LTR

# sim = similarity

# seq-nr = sequence number

259532 265448 5917 259532 259863 332 265117 265448 332 99.40 0

427672 430021 2350 427672 428170 499 429522 430021 500 91.60 0

29632 35597 5966 29632 29969 338 35259 35597 339 98.82 0

220989 226919 5931 220989 221339 351 226574 226919 346 97.15 0

此外，我们还使用-mintsd，-maxtsd，-motif和-motifmis 。

$ gt ltrharvest -index chr02.19970727.fsa -mintsd 5 -maxtsd 20 -motif tgca

-motifmis 0

# args=-index chr02.19970727.fsa -mintsd 5 -maxtsd 20 -motif tgca -motifmis 0

# predictions are reported in the following way

# s(ret) e(ret) l(ret) s(lLTR) e(lLTR) l(lLTR) s(rLTR) e(rLTR) l(rLTR) sim(LTRs)

seq-nr

# where:

# s = starting position

# e = ending position

# l = length

# ret = LTR-retrotransposon

# lLTR = left LTR

# rLTR = right LTR

# sim = similarity

# seq-nr = sequence number

259532 265448 5917 259532 259863 332 265117 265448 332 99.40 0

29632 35590 5959 29632 29963 332 35259 35590 332 99.70 0

220996 226911 5916 220996 221329 334 226575 226911 337 97.33 0

最后，如果我们对TSD的顺序和长度以及主题，我们选择选项-longoutput。这也是指定所有过滤器和用户的对齐选项。

$ gt ltrharvest -index chr02.19970727.fsa -seed 30 -xdrop 5 -mat 2 -mis -2 -ins -3

-del -3 -minlenltr 100 -maxlenltr 1000 -mindistltr 1000 -maxdistltr 15000

-similar 90.0 -overlaps all -mintsd 5 -maxtsd 20 -motif tgca -motifmis 0 -vic 60

-longoutput

# args=-index chr02.19970727.fsa -seed 30 -xdrop 5 -mat 2 -mis -2 -ins -3 -del -3

-minlenltr 100 -maxlenltr 1000 -mindistltr 1000 -maxdistltr 15000 -similar 90.0

-overlaps all -mintsd 5 -maxtsd 20 -motif tgca -motifmis 0 -vic 60 -longoutput

# predictions are reported in the following way

# s(ret) e(ret) l(ret) s(lLTR) e(lLTR) l(lLTR) TSD l(TSD) m(lLTR) s(rLTR)

e(rLTR) l(rLTR) TSD l(TSD) m(rLTR) sim(LTRs) seq-nr

# where:

# s = starting position

# e = ending position

# l = length

# m = motif

# ret = LTR-retrotransposon

# lLTR = left LTR

8

# rLTR = right LTR

# TSD = target site duplication

# sim = similarity

# seq-nr = sequence number

259532 265448 5917 259532 259863 332 gtaat 5 tg..ca 265117 265448 332

gtaat 5 tg..ca 99.40 0

29632 35590 5959 29632 29963 332 ataat 5 tg..ca 35259 35590 332

ataat 5 tg..ca 99.70 0

220996 226911 5916 220996 221329 334 ggaat 5 tg..ca 226575 226911 337

ggaat 5 tg..ca 97.33 0

|  |
| --- |
| **第10章** |

3.2对S.cerevisiae基因组的预测

作为目标序列文件，我们使用多个FASTA文件chrAll.19971001.fsa.gz，其中包含拥有所有16个释放的染色体序列，可能由Kim等人使用。在他们对逆转座子的全面调查中。首先，我们创建增强的后缀数组。我们使用选项-tis，-suf和-lcp ， -des ，因为LTRharvest需要相应的表。此外，我们指定选项-dna，因为我们正在处理DNA序列。

$ gt suffixerator -db chrAll.19971001.fsa.gz -indexname chrAll.19971001.fsa -tis

-suf -lcp -des -sds -dna

# dna=yes

# indexname="chrAll.19971001.fsa"

.

.

.

# TIME overall 108.35

现在我们可以将索引与LTRharvest一起使用。除了过滤器和对齐选项之外，我们还选择option -out，用于将预测的LTR反转录转座子序列打印到文件中。

$ gt ltrharvest -index chrAll.19971001.fsa -seed 100 -minlenltr 100 -maxlenltr 1000

-mindistltr 1000 -maxdistltr 15000 -xdrop 5 -mat 2 -mis -2 -ins -3 -del -3

-similar 90.0 -overlaps best -mintsd 5 -maxtsd 20 -motif tgca -motifmis 0

-vic 60 -longoutput

-out pred-chrAll.fsa

# args=-index chrAll.19971001.fsa -seed 100 -minlenltr 100 -maxlenltr 1000

-mindistltr 1000 -maxdistltr 15000 -xdrop 5 -mat 2 -mis -2 -ins -3 -del -3

-similar 90.0 -overlaps best -mintsd 5 -maxtsd 20 -motif tgca -motifmis 0

-vic 60 -longoutput -out pred-chrAll.fsa

# predictions are reported in the following way

# s(ret) e(ret) l(ret) s(lLTR) e(lLTR) l(lLTR) TSD l(TSD) m(lLTR) s(rLTR) e(rLTR)

l(rLTR) TSD l(TSD) m(rLTR) sim(LTRs) seq-nr

# where: 160239 166163 5925 160239 160575 337 ggttc 5 tg..ca 165827 166163 337

ggttc 5 tg..ca 100.00 0

29632 35590 5959 29632 29963 332 ataat 5 tg..ca 35259 35590 332

ataat 5 tg..ca 99.70 1

844407 850335 5929 844407 844744 338 gaaat 5 tg..ca 849998 850335 338

gaaat 5 tg..ca 100.00 15

56452 62375 5924 56452 56788 337 gttat 5 tg..ca 62039 62375 337

gttat 5 tg..ca 100.00 15

|  |
| --- |
| **第11页** |

3.3 LTRharvest输出的序列聚类

除了LTRharvest进行的预测过程外，还对预测序列进行了聚类分析被推荐。在这里，我们从Vmatch中选择单连锁聚类分析程序软件工具（不是GenomeTools二进制gt的一部分），以显示此任务的执行方式可以完成。程序需要根据预测的序列构造索引mkvtree是Vmatch的一部分。

$ mkvtree -db pred-chrAll.fsa -dna -pl -allout -v

reading file "pred-chrAll.fsa"

...

.

.

.

overall space peak: main=2.61 MB (10.01 bytes/symbol), secondary=0.53 MB

现在，我们可以对预测的序列使用vmatch（属于Vmatch的一部分）。以下命令计算所有（数据库）序列的集群在pred-chrAll.fsa中。每个簇都有一个唯一的簇号，后跟序列号包含在群集中。

$ vmatch -dbcluster 95 7 Cluster-pred-chrAll -p -d -seedlength 50

-l 1101 -exdrop 9 pred-chrAll.fsa

# args=-dbcluster 95 7 Cluster-pred-chrAll -p -d -seedlength 50

-l 1101 -exdrop 9 pred-chrAll.fsa

# 3 clusters

# 45 elements out of 46 (97.83%) are in clusters

# 1 elements out of 46 (2.17%) are singlets

# 1 cluster of size 2

# 1 cluster of size 3

# 1 cluster of size 40

0: 32 41 42 36 39 13 40 43 37 30 44 38 18 21 17 34 35 31 12 20

33 23 26 27 6 28 3 29 25 1 24 9 19 2 7 22 4 10 5 0

1: 8 15 11

2: 14 45

|  |
| --- |
| **第12章** |

[3] MI Abouelhoda，S。Kurtz和E. Ohlebusch。用增强的后缀数组替换后缀树。

离散算法杂志，2：53-86，2004年。

[4] D. Ellinghaus，S。Kurtz和U. Willhoeft。LTRharvest，高效，灵活的软件

重新检测LTR逆转座子。BMC生物信息学，2008年9月18日。

[5] G. Gremme。G ENOME T OOLS基因组分析系统。http://genometools.org。

[6] JM Kim，S。Vanguri，JD Boeke，A。Gabriel和DF Voytas。转座因子和

基因组组织：完整的Sac-

酿酒酵母基因组序列。Genome Research，8（5）：464-478，1998。

[7] S. Kurtz。V MATCH大型序列分析软件。http://www.vmatch.de。

11