|  |
| --- |
| **第1页** |

|  |
| --- |
| **第2页** |

1引言

本文档介绍了全基因组差异，这是一种用于测量紧密相关的基因组之间的进化距离软件工具。这些距离是成对的基因组之间的Jukes-Cantor校正散度，即它们之间每个碱基的突变数。该距离称为K r，基于所谓的shustrings。计算所有成对的距离都是无对齐的，但是最终的距离具有相同的生物学性，意味着好像是通过多序列比对计算的。本软件只能处理紧密相关的距离，因为K r仅在距离<0.5情况下可靠。

基因组差异是用C语言编写的，它基于GenomeTools库。它单个名为gt的二进制文件调用。可以在不进行任何更改的情况下在32位和64位平台上编译源代码。

2建立genomediff

由于基因组差异是GenomeTools软件套件的一部分，因此Genome必须例如通过GenomeTools主页（http://genometools.org）获得工具，并解压缩到源目录中：

$ tar -xzvf genometools -X.X.X. tar .gz

$ cd genometools -X.X.X

其中XXX表示所需的gt版本。

然后，使用提供的makefile调用make来编译源代码就足够了。如果系统建议使用GenomeTools可执行文件的64位版本支持这一点。传递选项64bit = yes启用64位支持。

选项amalgamation = yes允许编译器使用更好的优化。

$ make 64 bit = yes amalgamation = yes

|  |
| --- |
| **第3页** |

成功编译后，包含基因组差异的GenomeTools可执行文件在未压缩源根目录的bin子文件夹中。然后可以安装供系统范围内使用，如下所示（以root身份执行此操作）：

$ make 64 bit = yes amalgamation = yes install

使用安装目标时，请确保使用与编译步骤相同的选项！如果将prefix = <path>选项附加到此行，则可以将自定义目录指定为安装目标目录，例如

$ make 64 bit = yes amalgamation = yes install prefix =/ home / user /gt

会将gt二进制文件安装在/ home / user / gt / bin目录中。另请参阅自述文件和未安装的源树的根目录中的INSTALL文件以获取更多信息和故障排除建议。

3用法

3.1 genomediff命令行选项

由于genomediff是GenomeTools的一部分，因此它的调用方式如下：

gt genomediff [options] (INDEX | -indexname NAME SEQFILE SEQFILE [...])

其中INDEX是不包含包含基因组的编码序列的文件扩展名的路径进行比较，NAME是要根据给定序列文件构建的编码序列的名称。

表1给出了所有可能选项的简短描述。

清单1：单位文件示例：“ units”部分是必填项，“ genome1 / 2”部分是名称示例，文件名是在命令行上或在索引构建过程中给定的路径。

units = {

genome1 = { "file1.fas", "file2.fas" }，

genome2 = { "path/file3.fas", "file4.fas" }

}

3.2输入文件

工具genomediff可以处理三种类型的准备索引。首先是编码序列，可以由encseq准备。给定一个编码序列，基因组差异将在内存中构建增强的后缀数组，并使用该索引计算K r。第二个是工具后缀生成器（请参阅gt suffixerator -help）和第三个后缀生成的增强后缀数组,使用packedindex工具构建的压缩FM索引（参见gtpackedindex mkindex –help）。不建议使用FM索引，因为计算K r会花费大量时间。

|  |
| --- |
| **第4页** |

表1：genomediff命令行选项

输入选项

-indextype type 指定索引的类型，其中之一：esa | pck | encseq。其中encseq是编码序列和增强的后缀数组仅在内存中。默认值：encseq

-unitfile filename 指定基因组单位，请参见下面的描述。默认值：undefined

输出选项

-indexname name 要构造的encseq的基本名称。默认值：未定义

ESA选项

-mirrored 几乎附加每个序列默认值的反补

-pl n

为存储桶排序建议指定前缀长度：不使用争论; 则合理的前缀长度会自动确定挖出来的 默认值

-dc n

指定差异覆盖值。默认值：0

-memlimit n

指定在索引连接期间要使用的最大内存量指令（以字节为单位，允许使用关键字“ MB”和“ GB”）。

另一种方法是直接指定序列文件的名称。选项--indexname是必需的。在这种情况下，给定名称将用于在磁盘上存储编码序列。文件格式可以是GenomeTools支持的任何序列格式。

无论哪种方式，每个给定的序列文件都将被视为一个基因组单位，无论该文件中的序列数。

为基因组单位赋予文件名以外的其他名称，或将文件组合为单个基因组一个单位可以使用--unitfile选项提供一个单位文件。显示了示例单位文件的格式在清单1中。

3.3输出

标准输出流上的输出由一行与基因组数量或比较的单位。随后是成对距离的二次矩阵，其中每个行由文件名或单位名以及制表符分隔的距离值组成。根据gt调用的选项，可能会有其他输出，其中每行是在标准错误流上以“＃”为前缀，附加输出以“ debug：”为前缀。

|  |
| --- |
| **第5页** |

4例子

本节描述了两个示例方案，第一个是多个基因组的比较组织在单独的多个FASTA文件中，第二个是两个基因组的比较每个文件包含多个文件。

4.1在单独的文件中比较基因组

考虑三个文件基因组1.fas，基因组2.fas和基因组3.fas，每个文件可能包含多个FASTA条目。我们的机器有2个GiB RAM。假设索引构造会需要5 GiB，我们需要将其分成至少三个相等大小的部分，或者限制最大内存要求。

计算这三个基因组的距离矩阵的最简单方法是：

gt genomediff -indexname 3 genomes \

-memlimit 1500 MB \

genome1 .fas genome2 . fas genome3 .fas

--memlimit应该合理地小于可用的主内存。这将在终端上输出距离矩阵，并使用当前目录中的basename 3genomes。为了将结果保存到文件中，请使用终端重定向：

gt genomediff ... > outfile。

该文件OUTFILE可能是这样的：

3

genome1.fas 0.000000 0.115125 0.267473

genome2.fas 0.115125 0.000000 0.293082

genome3.fas 0.267473 0.293082 0.000000

此制表符分隔的表格可与Phylip或R一起使用，以计算出遗传树。如果给定文件的增强后缀数组具有名称3genomes idx在磁盘上已经存在，将是这样的：

gt genomediff -indextype esa 3 genomes\_idx > outfile

要重用现有的编码序列，只需提供其基本名称：

gt genomediff 3 genomes > outfile

4.2比较多个文件中的两个基因组

假设我们有两个分别由多个染色体组成的基因组，它们分别位于不同的文件中。例如，基因组1由g1 chr1.fas和g1 chr2.fas组成，而基因组2的两个文件是相应地命名。unitfile可以这样组织：

|  |
| --- |
| **6页** |

units = {

genome1 = { “g1\_chr1 。FAS” ，“g1\_chr2 。FAS” }，

genome2 = { “g2\_chr1 。FAS” ，“g2\_chr2 。FAS” }

}

在我们的示例中，unitfile的名称为units。现在我们可以这样调用genomediff：

gt genomediff -indexname 2 genomes \

-unitfile units \

g1\_chr1 .fas g1\_chr2 . fas g2\_chr1 .fas g2. chr2 . fas > output

文件输出如下所示：

2

genome1 0.000000 0.115125

genome2 0.115125 0.000000

参考书目

参考

[1] Bernhard Haubold，Mirjana Domazet-Loso和Thomas Wiehe。免对准显示器

密切相关基因组的位置测量。在RECOMB-CG '08中：

比较基因组学国际研讨会，第87-99页，柏林，海德堡，2008年。

施普林格出版社。

[2] B. Haubold，P。Pfaffelhuber，M。Domazet-Loso和T. Wiehe。估计突变

来自未对齐基因组的信息。J.计算机 Biol。，16：1487–1500，“ Oct”，2009年。

[3] Bernhard Haubold，Floyd A. Reed和Peter Pfaffelhuber。无对齐估计

核苷酸多样性。生物信息学，2011，27（4）：449-455。

[4] Gordon Gremme。GenomeTools基因组分析系统。http://genometools.org。

5