从 DNA 序列预测 RFLP 和 TRFLP

seqRFLP 使用说明

Version 1.0.0

丁　琼

张金龙

中国科学院植物研究所生态中心

生物多样性与生物安全研究组

http://www.biodiv.ibcas.ac.cn/

声　明

在使用本软件前，请认真阅读以下协议。

1 使用 seqRFLP 软件意味着用户认同该协议。

2 本软件符合 GPL-2 协议，用户可以免费下载和使用该软件，可以对源代码进行修改。如需使用本软件，请按照如下方式引用:

Qiong Ding and Jinlong Zhang (2010). seqRFLP: Simulation and visualization of restriction enzyme cutting pattern from DNA sequences.. R package version 1.0.0. http://CRAN.R-project.org/package=seqRFLP

3 如有问题，请发邮件至 dingqiong1@gmail.com 或 jinlongzhang01@gmail.com 我们将尽量解答遇到的各种问题，但并不保证每封邮件都能及时回复。

4 由于使用本软件造成的法律问题，软件作者不承担任何责任。

丁　琼

张金龙

2010 年 7 月 3 日

于北京 香山

关于作者:

丁　琼

女，博士，2009 年毕业于中国科学院植物研究所。主要兴趣是土壤菌物/微生物多样性的数据分析，目前从事土壤微生物分子生物学领域的研究。

张金龙

男， 中国科学院植物研究所博士研究生。主要兴趣是大尺度生物多样性数据挖掘，将系统进化信息整合到生物多样性分析中。

目　录

1. seqRFLP 简介

RFLP 和 TRFLP 等技术在生物多样性研究中十分广泛。随着测序技术的日趋普及，在进行 RFLP 或 TRFLP 之前，用户很可能已经获得了 DNA 序列。为了对 RFLP 和 TRFLP 的结果进行更好的分析，特别是选择合适的内切酶，对模糊的电泳条带进行确认，基于 RFLP/TRFLP 及序列分别对物种进行鉴定等显得日趋重要，但是目前还缺乏这样一种已知序列就能获得 RFLP/TRFLP 的分析平台。为此我们编写了 seqRFLP 软件。

seqRFLP 是用来对 DNA 序列进行模拟酶切、模拟 PCR 引物筛选的软件包，目前版本是 1.0.0。用开源的 R 语言 (http://www.r-project.org/)写成，可以在 Linux, Windows, MacOS 等多种计算机平台上运行。seqRFLP 已经被 RCRAN 接收，用户可以在 R 的任何一个镜像下载。本说明假设用户使用 Windows，其他平台用户的使用方法与之类似，均需先安装 R 软件。

2. seqRFLP 的下载和安装

首先下载和安装 R 软件。建议选择较新的版本，如 Windows: R-2.11.1，其下载地址之一为:

http://ftp.ctex.org/mirrors/CRAN/bin/windows/base/R-2.11.1-win32.exe

下载到本地后，双击 R-2.11.1-win32.exe 开始安装，各选项默认即可。安装完成后双击开始>所有程序>R>R 2.11.1

输入命令:

```
install.packages("seqRFLP")
```

R 将提示选择 CRAN,选择距离较近的 CRAN 镜像，将软件包自动下载和安装好。

导入 seqRFLP 程序包:

```
library(seqRFLP)
```

3.查询 seqRFLP 的帮助

seqRFLP 的每个函数都有详细的帮助文件，要查询帮助可输入:

`?seqRFLP` 查看详尽的帮助手册:用户也可以在 CRAN 上下载 pdf 版本的帮助手册。

seqRFLP 符合 GPL-2 协议，用户可以免费使用并拷贝该软件，也可以更改其源代码。

运行 seqRFLP 实例:

```
example(seqRFLP)
```

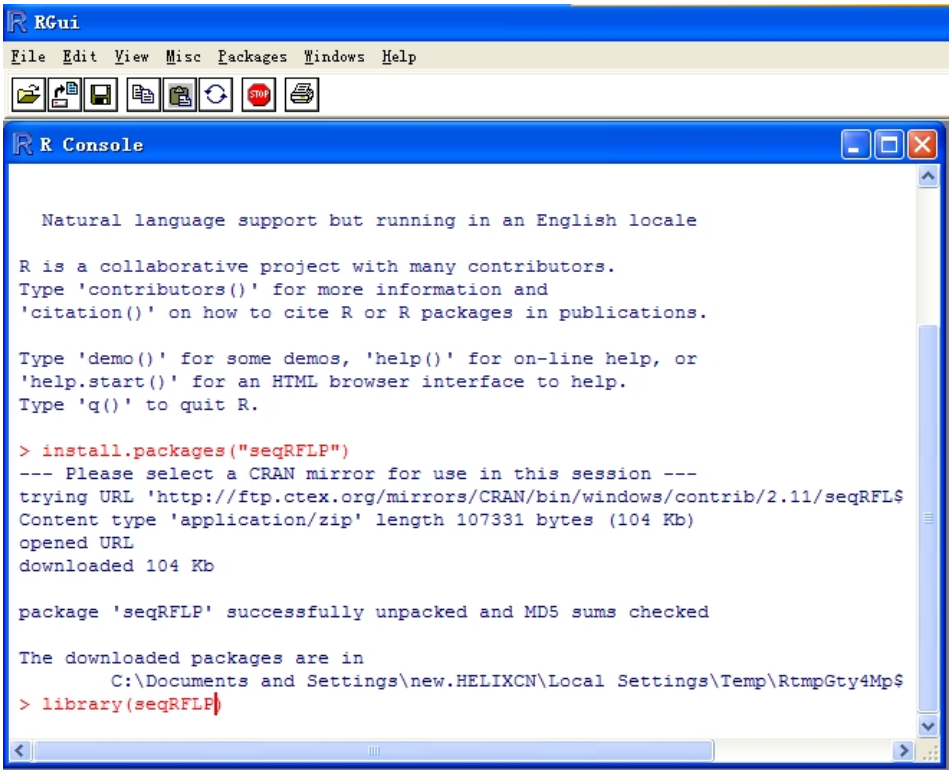

图 1 seqRFLP 软件的下载和安装

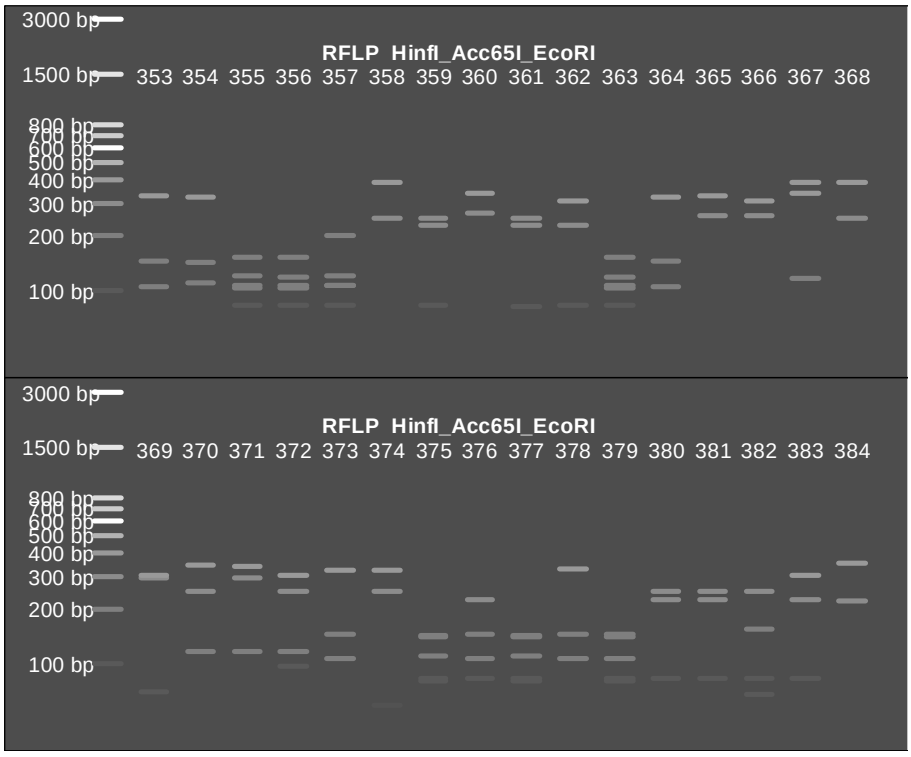

图 2 用“EcoRI”, “Acc65I”, “HinfI”进行模拟酶切的结果

4.主要函数及用法

当前版本(1.0.0)的 seqRFLP 共有 22 个函数，用户可以通过查询其 R 帮助文档，这里介绍的是数据导入导出与分析的主要函数。

file.cat()

描述： 将一个目录下任何纯文本格式的文件拷贝到新生成的目标文件中。

参数： `file.cat(dir = NULL, appendix = ".fas", file = NULL)`

参数说明： `dir` 文件夹

`appendix` 文件后缀

`file` 要生成的文件

实例：

在 `D:/direct/`文件夹下 下有 `100` 个 `fasta` 文件，扩展名为`.fas`，需要将该 `100` 个 `fasta` 文件拷贝到 `result.fasta` 文件中

输入命令：

```
file.cat( "D:/direct", ".fas", file = "result.fasta")
```

就可将该目录下的扩展名为`.fas` 的所有序列文件的内容拷贝到 `result.fasta` 文件中。

read.fasta()

描述：读取 `fasta` 格式的文件

参数：`read.fasta(fil = NULL)`

参数说明： `fil` 要读取的文件路径

例如 `DNA` 序列信息保存在 `D:/data/sequences.fasta`

读取的方法是

```
read.fasta("D:/data/sequences.fasta")
```

read.phy()

描述： 读取 phylip 格式的文件
参数： read.phy(fil = NULL)
参数说明： fil 要读取的文件路径
实例

例如 DNA 序列信息保存在 D:/data/sequences.phy
读取的方法是

```
read.phy("D:/data/sequences.phy")
```

read.nxs()

描述： 读取 nexus 格式的文件
参数： read.nxs(fil = NULL)
参数说明： fil 要读取的文件路径
实例

例如 DNA 序列信息保存在 D:/data/sequences.nex
读取的方法是

```
read.nxs("D:/data/sequences.nex")
```

as.fasta()

描述： 将读取的 phylip 或 nexus 转换为 fasta 格式
参数： as.fasta(x)
参数说明： 经过 read.phy()或 read.nxs() 读取进来的对象
详情说明： 模拟酶切，需要使用 fasta 类型的数据，输入的数据为 phylip 格式或 nexus 格式时，需要先转换为 fasta 文件。在 seqRFLP 中，无论是读取的 phylip 还是 nexus 文件，都可以用 as.fasta()转换为 fasta 类型的数据，供数据分析中使用。
实例

```
seqphy <- read.phy("D:/data/example.phy")
seqfas <- as.fasta(seqphy)
seqfas
```

gnames.fas()

描述： 提取各序列的名称
参数： gnames.fas(x = NULL)

参数说明：fasta 类型的数据，可以用 read.fasta 函数读取，或 read.phy,read.nxs 读取后，经过 as.fasta 转换后的 fasta 类型的数据。

实例

```
data(fil.fas)
gnames.fas(file.fas)
```

rename.fas()

描述：为 fasta 的对象内的序列重命名。

参数：rename.fas(x, names = NULL)

参数说明：x 必须是 fasta 类型的数据，可以用 read.fasta 函数读取，或 read.phy,read.nxs 读取后，经过 as.fasta 转换后的 fasta 类型的数据。

names 是要更改的序列名称，必须与 fasta 数据内的序列的条数一致。

实例

例如：

已经读取的序列的各名称

```
data(fil.fas)
gnames.fas(fil.fas)
```

一致 fil.fas 里面有 19 条序列，更改序列名称

```
rename.fas(fil.fas, name = paste("Sequence", as.character(1:19),
sep = ""))
```

frag.dat()

描述： 预测 RFLP 酶切结果

参数：frag.dat(fil, enznames, enzdata)

参数说明

fil 必须是 fasta 类型的数据，可以用 read.fasta 函数读取，或 read.phy,read.nxs 读取后，经过 as.fasta 转换后的 fasta 类型的数据。

enznames 内切酶的名称，必须是字符串。

enzdata 内切酶数据表，包含酶切位点信息，详情查看 ?enzdata

输出结果:酶切片段长度的数据表

第一列为序列的名称

enznames 为所用内切酶的名称

recogSite 内切酶所识别的位点的碱基序列

cutting_Site 内切酶

fragment_Length 该序列的内切酶酶切各片段长度

T5 5′末端片段长度

T3 3′末端片段长度

实例

```
data(enzdata)
data(fil.phy)
fas <- ConvFas(fil = fil.phy, type = "phy")
frag.dat(fil = fas, enznames = "EcoRI", enzdata = enzdata)
```

```
          enznames recogSite cutting_Site fragment_Length  T5  T3
>006_ITS1F    EcoRI  G'AATT_C         345        344,329 344 329
>006_ITS4_r   EcoRI  G'AATT_C         345        344,304 344 304
>003_ITS1F    EcoRI  G'AATT_C         326        325,338 325 338
>003_ITS4_r   EcoRI  G'AATT_C         326        325,311 325 311
>004_ITS1F    EcoRI  G'AATT_C         340        339,330 339 330
>004_ITS4_r   EcoRI  G'AATT_C         340        339,306 339 306
>002_ITS1F    EcoRI  G'AATT_C           0           561 561 561
>002_ITS4_r   EcoRI  G'AATT_C           0           533 533 533
>001_ITS1F    EcoRI  G'AATT_C         306        305,334 305 334
```

plotenz()

描述： 以电泳图的结果展示模拟酶切结果

参数： `plotenz(sequences, enznames, enzdata, side = TRUE, type = c("RFLP", "TRFLP"), Terminal = c("T5", "T3"))`

参数说明：

`sequences`, `fasta` 类型的数据，可以用 `read.fasta` 函数读取，或 `read.phy`,`read.nxs` 读取后，经过 `as.fasta` 转换后的 `fasta` 类型的数据。

`enznames`, 内切酶的名称，必须是字符串。

`enzdata`, 内切酶数据表，包含酶切位点信息，详情查看 `?enzdata`

`side = TRUE`, 是否将酶放入同一泳道

`type = c("RFLP", "TRFLP")`： RFLP 抑或 TRFLP

`Terminal = c("T5", "T3")` ：在选择 TRFLP 后，是选择标记 5’抑或 3’端的引物

实例

```
# 导入内切酶数据
data(enzdata)
# 导入内置的 phylip 数据
data(fil.phy)
#转换成 fasta 格式
fas <- ConvFas(fil = fil.phy, type = "phy")
```

RFLP 选择的酶名称

```
enznames <- c("EcoRI", "Acc65I", "HinfI")
plotenz(sequences = fas, enznames = enznames, enzdata = enzdata,
side = TRUE, type = "RFLP")
```

模拟电泳图

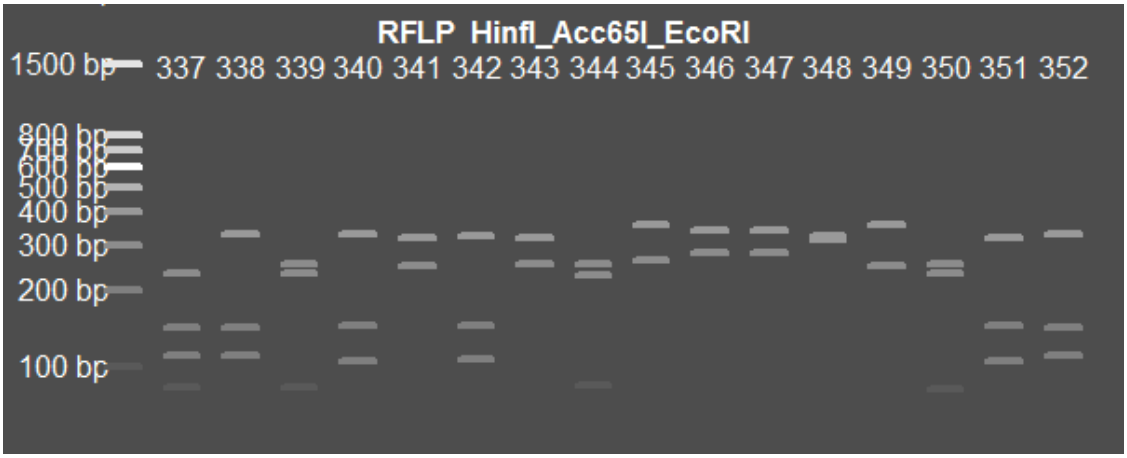

图 3 每一个电泳槽表示经过 RFLP, TRFLP后一个条带的电泳结果。

clipprobe()

描述： 在序列中寻找可以被 PCR 扩增出的序列
参数： `clipprobe(fil, forProbe, bacProbe, tol = 3, clipped.only = TRUE)`
参数说明：

`fil,fil` ： `fasta` 类型的数据，可以用 `read.fasta` 函数读取，`read.phy,read.nxs` 读取后，经过 `as.fasta` 转换后的 `fasta` 类型的数据。
`forProbe`：5′端引物序列，需要以字符串的形式给出
`bacProbe`：3′端引物序列，需要以字符串的形式给出
`tol` 允许错配的碱基数
`clipped.only`： 只显示可以被酶切的序列

实例

筛选 PCR 序列

```
## 第一步 设定前后引物
forProbe = ITS1F = 'CTTGGTCATTTAGAGGAAGTAA' # forward primer should
be from the 5' to 3' end.
bacProbe = ITS4 = 'GCATATCAATAAGCGGAGGA'    # backward primer

### 第二步 数据读取
directory <- system.file("extdata", package = "seqRFLP")
```

```
path <- file.path(directory, "seqs.fasta")
fas <- read.fasta(path)
```

第三步 寻找到可被 PCR 扩增出的序列

```
clipped <- clipprobe(fas, forProbe, bacProbe, tol = 2, clipped.only
= TRUE)
```

第四步 结果检查

```
## 通过获得序列名称数量，给出可被 PCR 扩增出序列数
length(gnames.fas(clipped))
## 原始输入序列中有 393 条
length(gnames.fas(fas))
## 输出不能被酶切的序列名称
setdiff(gnames.fas(fas), gnames.fas(clipped))
```

5 分析实例:

下载 example1.fasta 文件

下载网址: http://www.biodiv.ibcas.ac.cn/

另存到如 D:/data/example1.fasta

```
#读取 example1.fasta
test <- read.fasta("D:/data/example1.fasta")
#获取各序列名称
gnames.fas(test)
#更改第四条序列的名称
nams <- gnames.fas(test)
nams[4] <- "sequences4"
test2 <- rnames.fas(test, nams)
#查看各序列被 HinfI 酶切结果
data(enzdata)
frag.dat(test2, "HinfI", enzdata = enzdata)
#查看各序列的被 HinfI 的酶切结果
plotenz(test2, "HinfI", enzdata = enzdata,side = TRUE, "RFLP")
```

致 谢

感谢马克平研究员、梁宇博士的指导，感谢博士研究生孙秀峰及龙恩熙对软件的测试，感谢软件包的编写有过帮助的研究组各同学和老师。

参考文献

Saiki RK, Scharf S, Faloona F, Mullis KB, Erlich HA, Arnheim N (1985). Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. Science 230 (4723):1350-4

Roberts, R.J., Vincze, T., Posfai, J., Macelis, D. (2010) REBASE–a database for DNA restriction and modification: enzymes, genes and genomes. Nucl. Acids Res. 38: D234-D236. http://rebase.neb.com