doi:10.3969/j.issn.1672-5565.2013.03.05

ChooseMaterials.pl,控制变量挑选实验材料的 perl 脚本

李旭凯,郭 凯,彭良才,王令强*
(华中农业大学作物遗传改良国家重点实验室,华中农业大学植物科学技术学院;华中农业大学生物质与生物能源中心,湖北 武汉 430070)

摘 要:用 Perl 语言编写了一个脚本—ChooseMaterials.pl,通过计算两个材料特定某几列数据之间增加或减少的百分数水平,达到控制变量突出关键因子的作用,方便生物学工作者选取材料,做进一步的实验或分析。

关键词:ChooseMaterials.pl;Perl 脚本;挑选实验材料;控制变量
中图分类号:Q811.4; R857.3 **文献标识码**:A **文章编号**:1672-5565(2013)-03-186-06

ChooseMaterials.pl, a perl script for picking out the samples by controlling variable

LI Xu-kai, GUO Kai, PENG Liang-cai, WANG Ling-qiang*
(*National Key Laboratory of Crop Genetic Improvement, College of Plant Sciences and technology, Biomass and Bioenergy Research Center, Huazhong Agricultural University, Wuhan* 430070 ,*China*)

Abstract: A perl script ChooseMaterials.pl was written for picking the choose samples by calculate the percentage of the increased or decreased level at pair: subtraction of two samples divided by means of two values at pair. It achieves controlling variable and outstanding key factors effects, and convenient for picking the choose samples by investigators to further experimental or analysis.

Keywords: ChooseMaterials.pl; Perl Script; Picking Damples; Vontrol Variable

随着生物数据的海量增长,如测量或者实验获得的性状、含量、指标等数字数据,快速高效的方法将在生物学领域扮演越来越重要的角色。由于在日常实验数据分析中,研究单个因子或者某几个多因子的作用,控制变量是一种很好的方法。但是从众多数据中查找出研究人员预期的材料对是一件费时费力的事情。通过计算机软件,简单快捷地挑选出实验材料,可以缩减人为查找挑取材料是时间和精力,也可以避免人为计算带来的误差和遗漏。

1 软件安装

1.1 Perl 在 windows 操作系统下的安装

Perl 是"实用报表提取语言"(Practical Extraction and Report Language)的缩写,由拉里·沃尔(Larry Wall)最初设计编写的。此外还有其他不同的全名展开方式,如 Pathologically Eclectic Rubbish Lister。Perl 是个溯写字(backronym),而不仅仅是缩写词,是 perl 这个词本身先被使用,后来才给出展开的词作诠释。Perl 是可跨平台运行的,可用于 UNIX、Linux、MAC 和 Windows 等环境下编程,并由 CPAN 不断更新和维护[1]。Perl 借取了 C、sed、awk、shell 脚本以及很多其他编程语言的特性,其中最重要的特性是他内部集成了正则表达式的功能。

ChooseMaterials.pl 是 Perl 脚本,要求在电脑上安装有编译并运行的 perl 解释器。大部分类 UNIX 系统(包括 Linux 和 Mac OS X),perl 是随系统安装的,可在命令行终端输入命令 perl -v,查看版本,

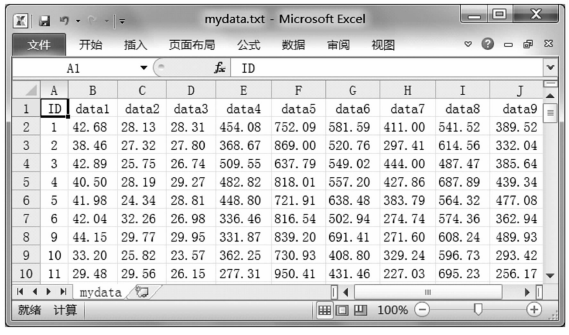
收稿日期:2013-02-20;**修回日期**:2013-04-18.
资助项目:国家自然科学基金(30900890,31171524)、全国博士后基金(20070420917)、湖北省自然科学基金(2009CDB324)和中央高校基本科研业务费专项资金(2011PY047)。
作者简介:李旭凯,男,山西省太原市人,博士生,主要研究方向:生物信息学;E-mail:specterae@163.com.
*通讯作者:王令强,男,浙江省奉化市人,硕导,副教授,研究方向:能源作物分子生物学和遗传育种;E-mail:lqwang@mail.hzau.edu.cn.

Windows 有两种版本可用:Strawberry Perl 与 ActivePerl。在 Windows 下刚开始使用 Perl 的人,大部分会采用 ActivePerl。但对于熟悉 Linux 环境下 Perl 编程的人来说,用 Strawberry Perl 会更加习惯。Strawberry Perl 可视为是 windows 下的 the core Windows distribution of Perl 的一个版本[2],它尽可能地在 Windows 平台上保持了 Perl 在 Unix 上的特性,从而也最大程度地保证了可移植性。因此,CPAN 上的包[3],在 Strawberry Perl 下容易编译通过。Strawberry Perl 目前最新版本为 Strawberry Perl 5. 16. 2. 1,(下载地址为 http://strawberry-perl. googlecode. com/files/strawberry-perl-5. 16. 2. 1-32bit. msi)。双击安装文件并按照其安装步骤即可完成 Perl 在 windows 操作系统中的安装,一般安装在 C:\根目录下,包含 c、cpan、licenses、perl 和 win32 五个文件夹。

1.2　数据的准备

测量或者实验获得的数据(如:性状、含量、指标等)都以数字的形式存储。

以上这类数据表可以转存在 Excel 表格中,第一行为数据表的表头,从第二行开始是材料编号和数据,A 列为材料的编号,其他列是对应的数字数据(见图 1)。ChooseMaterials. pl 定义在从 B 列开始计数为 1 列、C 列为 2 列、D 列为 3 列,以此类推。要求读者分别确定相似度参数和差异度参数,以及这两个参数分别对应的列数。数据结构见表 1。



	A	B	C	D	E	F	G	H	I	J
1	ID	data1	data2	data3	data4	data5	data6	data7	data8	data9
2	1	42.68	28.13	28.31	454.08	752.09	581.59	411.00	541.52	389.52
3	2	38.46	27.32	27.80	368.67	869.00	520.76	297.41	614.56	332.04
4	3	42.89	25.75	26.74	509.55	637.79	549.02	444.00	487.47	385.64
5	4	40.50	28.19	29.27	482.82	818.01	557.20	427.86	687.89	439.34
6	5	41.98	24.34	28.81	448.80	721.91	638.48	383.79	564.32	477.08
7	6	42.04	32.26	26.98	336.46	816.54	502.94	274.74	574.36	362.94
8	9	44.15	29.77	29.95	331.87	839.20	691.41	271.60	608.24	489.93
9	10	33.20	25.82	23.57	362.25	730.93	408.80	329.24	596.73	293.42
10	11	29.48	29.56	26.15	277.31	950.41	431.46	227.03	695.23	256.17

图 1　Excel 表中的数据形式

Fig. 1　Data form in the Excel table

表 1　输入的的数据结构

Table 1　The data structures of input should as follows (RowNColumnN is your data)

ID	data1	data2	data3	…	dataN
Id1	Row1Column1	Row1Column2	Row1Column3	…	Row1ColumnN
Id2	Row2Column1	Row2Column2	Row2Column3	…	Row2ColumnN
Id3	Row3Column1	Row3Column2	Row3Column3	…	Row3ColumnN
…	…	…	…	…	…
IdN	RowNColumn1	RowNColumn2	RowNColumn3	…	RowNColumnN

将此 Excel 表另存为“文本文件(数据用制表符分割)(*. txt)”类型的记事本文件。将这个数据文件与 ChooseMaterials. pl 存放在同一个目录“C:\strawberry\perl\bin”下。

1.3　ChooseMaterials. pl 的运行

将此文件拷贝到 C:\strawberry\perl\bin\目录中,然后在“开始”菜单的“所有程序”的“附件”中打开“命令提示符”,键入“cd C:\strawberry\perl\bin”后回车(即键入“Enter”键),即将当前目录转到含 perl 命令执行程序的目录下。在“C:\strawberry\perl\bin>”的提示符下键入“perl ChooseMaterials. pl data. txt Result. txt”,回车即可运行。其中“data. txt”是要挑选的数据材料文件名,“Result. txt”是输出结果的文件名,键入“perl ChooseMaterials. pl -h”可获得帮助信息。

运行“perl ChooseMaterials. pl data. txt Result. txt”后,程序会提示“Please input the similar parameter and follow the number of column that you choose (Separated by Spaces between numbers):”即第一个数字输入相似度参数,随后的数字输入对应的列数(数字用空格隔开),然后回车,如果没有输入程序默认输入“51”。程序接着提示“Please input the differences parameter and follow the number of column that you choose (Separated by Spaces between numbers):”即第一个数字输入差异度参数,随后的数字输入对应的列数(数字用空格隔开),然后回车,如果没有输入程序默认输入“202”,(见图 2)。示例中表示要从数

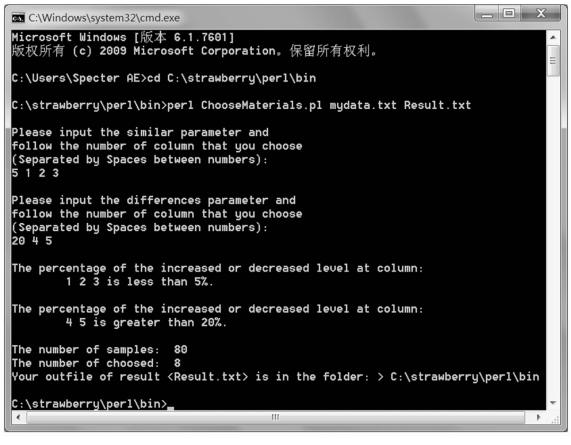


图 2　ChooseMaterials. pl 的运行界面示意图

Fig. 2　Running window of ChooseMaterials. pl

据“mydata. txt”中挑选出 data1、2、3 列相似度小于 5%,data4、5 列差异度大于 20% 的材料对,结果输出到“Result. txt”文件。

程序结束后,可以在“C:\strawberry\perl\bin”目录下找到输出的结果文件。结果文件中第一行为数据表的表头,与原始数据的表头相比多出了 Sample1、Sample2 和 Correlation 三列,分别对应的是挑选出来的材料编号和它们的相关系数。从第二行开始,每三行为一个组,每组的第一行是增加或减少的百分数水平、材料编号和相关系数,后面的两行是这两个材料的原始数据。从示例结果看出 data1、2、3 增加或减少的百分数水平都小于 5%,data4、5 的百分数水平都大于 20%,完全符合参数要求。结果同时给出了 data6 至 data9 数据的增加或减少的百分数水平作为参考(见图 3。)

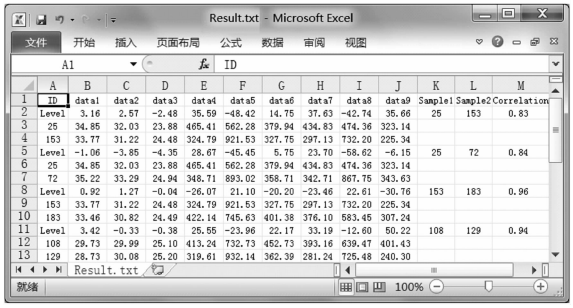


图 3 ChooseMaterials. pl 的运行结果

Fig. 3 The result of ChooseMaterials. pl

2 ChooseMaterials. pl 的原理和特点

ChooseMaterials. pl 是通过计算两个材料特定某几列数据之间增加或减少的百分数水平(两个数的差除以这两个数的平均值乘以一百)[4],达到控制变量突出关键因子的作用。正数表示第一行材料数据大于第二行的百分数,负数表示第一行材料数据小于第二行的百分数,并给出这两对样本的相关系数[5],以便生物学工作者选取材料,做进一步的实验或分析。ChooseMaterials. pl 要求其输入的文档为文本文档,其输出文档也为文本文档。在输入的数据文档中,第一行为数据表的表头,从第二行开始是材料编号和数据,A 列为材料的编号,其他列是对应的数字数据。编号与数据、数据与数据之间要用空格或 Tab 隔开。

运行 ChooseMaterials. pl 操作简单、快捷,可以减少人为查找挑取材料是时间和精力,也可以避免人为计算带来的误差和遗漏。

3 ChooseMaterials. pl 的源代码

ChooseMaterials. pl 的源代码如下,读者可以将其拷贝到记事本中,任意保存为一个文件名(如:ChooseMaterials. txt),然后重命名为 ChooseMaterials. pl,之后就可以按照 1.3 的运行方法使用。

代码如下:

```
#! /usr/bin/perl
die"Incorrect number of command line arguments. \nUsage: perl ChooseMaterials. pl  <infile. txt>
     <outfile. txt> \n\nTo show brief help usage, do "perl ChooseMaterials. pl -h\"\n" unless$ARGV[1];
my$choosed =0;

print"\nPlease input the similar parameter and \nfollow the number of column that you choose\n(Separated
     by Spaces between numbers):\n";
my @col1 =split(/\s+/, <STDIN>);
$col1[0] = ~ /\d+/? (@col1):(@col1 =(5,1,2,3));
my$similar =shift(@col1); ###相似度参数
print"\nPlease input the differences parameter and \nfollow the number of column that you
     choose\n(Separated by Spaces between numbers):\n";
my @col2 =split(/\s+/, <STDIN>);
$col2[0] = ~ /\d+/? (@col2):(@col2 =(20,4,5));
my$differences =shift(@col2); ###差异度参数
print"\nThe percentage of the increased or decreased level at column:\n\t@col1 is less than
     $similar\%. \n\nThe percentage of the increased or decreased level at column:\n\t@col2 is greater
     than$differences\%. \n";
```

```
open(DATA,$ARGV[0])||die"Couldn't open infile:$!";
my @data = <DATA>; ###将所有数据储存在数组@data 中
close DATA;
chomp @data;
open(OUTFILE,">$ARGV[1]")||die"Couldn't open outfile:$!";
print OUTFILE
join("\t",$data[0],'Sample1','Sample2','Correlation'),"\n";

foreach my$i (1..$#data){ ###循环不重复的取出将某两行数据
    foreach my$j($i+1..$#data){
      my @row1 = split(/\s+/,$data[$i]);
      my @row2 = split(/\s+/,$data[$j]);
      foreach my$k (1..$#row1){
        $Col[$k] = ($row1[$k]-$row2[$k])/($row1[$k]+$row2[$k])*2*100;
      }
      foreach$col1(@col1){
        $MAX[$col1] = abs($Col[$col1]);
      }
      foreach$col2(@col2){
        $MIN[$col2] = abs($Col[$col2]);
      }
      my$max = max(@MAX); ###@col1 中的最大值
      my$min = min(@MIN); ###@col2 中的最小值
      if($max <$similar and$min >$differences){ ###满足设置的参数
        $choosed++;
        $Col[0] = 'Level';
        my$name1 = shift @row1;
        my$name2 = shift @row2;
        my$correlation = pearson(\@row1,\@row2); ###计算这两组数据的相关系数
        print OUTFILE
    join("\t",@Col),"\t$name1\t$name2\t$correlation\n",join("\t",$name1,@row1),"\n",join("\t",$
name2,
    @row2),"\n";
      }
    }
}
close OUTFILE;
print"\nThe number of samples:\t$#data\nThe number of choosed:\t$choosed\nYour outfile of result
    《$ARGV[1]》 is in the folder: > ";
system"cd";

###该子程序用于输出数组中的最大值###
sub max{
    my$max = shift;
   $max =$_>$max?$_:$max for @_;
    return$max;
```

```
}
###该子程序用于输出数组中的最小值###
sub min{
    my$min = 100000;
    for( @ _) {
      if($_ = ~ /\d +/ and$_ <$min) {
        $min =$_;
    }
  }
  return$min;
}
###该子程序用于计算两行数据的相关系数###
sub pearson{
    my  ($ref_a,$ref_b) = @ _;
    my  @ x = @ {$ref_a};
    my  @ y = @ {$ref_b};
    if($#x = =$#y) {
      my$N =$#x;
      my$sum_sq_x = 0;
      my$sum_sq_y = 0;
      my$sum_coproduct = 0;
      my$mean_x =$x[1];
      my$mean_y =$y[1];
      for (my$i = 2;$i < =$N;$i + + ) {
        my$sweep = ($i-1.0)/$i;
        my$delta_x =$x[$i]$mean_x;
        my$delta_y =$y[$i]$mean_y;
        $sum_sq_x + =$delta_x *$delta_x *$sweep;
        $sum_sq_y + =$delta_y *$delta_y *$sweep;
        $sum_coproduct + =$delta_x *$delta_y *$sweep;
        $mean_x + =$delta_x/$i;
        $mean_y + =$delta_y/$i;
      }
      my$pop_sd_x = sqrt($sum_sq_x);
      my$pop_sd_y = sqrt($sum_sq_y);
      my$cov_x_y =$sum_coproduct;
      my$correlation =$cov_x_y/($pop_sd_x *$pop_sd_y);
      return$correlation;
    }
}

### Print help ###
my$PrintHelp = qq(
USAGE:
  perl ChooseMaterials.pl  < infile.txt >  < outfile.txt >
```

AUTHOR:

Xukai Li (specterae\@163. com) 2013/01

DESCRIPTION:

This script was written for picking the choosed samples by

calculate the percentage of the increased or decreased level at pair:

subtraction of two samples divided by means of two values at pair.

Written by Xukai Li and test under the perl enviroment 5. 12. 3.

DATA STRUCTURES:

The data structures of input should as follows (RowNColumnN is your data):

ID	data1	data2	data3	...	dataN
Id1	Row1Column1	Row1Column2	Row1Column3	...	Row1ColumnN
Id2	Row2Column1	Row2Column2	Row2Column3	...	Row2ColumnN
Id3	Row3Column1	Row3Column2	Row3Column3	..	. Row3ColumnN
...	...	...	...	...	...
IdN	RowNColumn1	RowNColumn2	RowNColumn3	...	RowNColumnN

EXAMPLE:

perl ChooseMaterials. pl Mydata. txt Result. txt

\n);

die($PrintHelp) if($ARGV[0] = ~/-[hH] +/);

4　讨论

在准备数据表时,一般通过 Excel 电子表格将其存为文本文档,也可以直接在文本文档中编辑数据。但文件中的第一行是表头说明,第二行开始为材料名和数据,在材料名和数据、数据与数据之间要用空格或 Tab 分开。

ChooseMaterials. pl,程序并不能完全替代人的工作,它只是缩小了挑选材料的范围,读者还需要从程序运行的结果中再次人工确定需要挑选的材料。

这个程序的不足是设置的参数只有两个,即相似度参数和差异度参数,而不能对每个列分别设置单独的参数。

参考文献(References)

[1] Randal L. Schwartz, Tom Phoenix and brian d foy. Learning Perl. 5th Edition [M]. Sebastopol: O'Reilly Media, 2008. 1-17.

[2] Adam Kennedy. Strawberry Perl for Windows [EB/OL], http://strawberryperl. com, 2013-03-12.

[3] The Comprehensive Perl Archive Network (CPAN) [EB/OL], http://www. cpan. org, 2012-10-09.

[4] Ning Xu, Wei Zhang, Shuanfeng Ren, Fei Liu, Chunqiao Zhao, Haofeng Liao, Zhengdan Xu, Jiangfeng Huang, Qing Li, Yuanyuan Tu, Bin Yu, Yanting Wang, Jianxiong Jiang, Jingping Qin, Liangcai Peng. Hemicelluloses negatively affect lignocellulose crystallinity for high biomass digestibility under NaOH and H2SO4 pretreatments in Miscanthus [J]. Biotechnology for Biofuels, 2012, 5:58-70.

[5] Eric Weisstein. Correlation Coefficient of Wolfram [EB/OL], http://mathworld. wolfram. com/CorrelationCoefficient. html, 2013-04-23.