doi:10.3969/j.issn.1672-5565.2014.03.06

基于 Bioperl 实现远程自动获取抗逆基因序列

张晓婧,潘伟民,曹兴芹*
(新疆师范大学生命科学学院,新疆 乌鲁木齐 830054)

摘 要:BioPerl 对于现今大量的生物数据来说,具有很好的获取和处理能力,并且 NCBI 提供了可以直接访问它下属的 Entrez 数据库(包括 PubMed)的编程接口,即 E-Utilities。为了从 NCBI 中自动获取大量抗逆基因序列,使得生物抗逆基因二次数据库得以搭建,该文基于 BioPerl 设计了远程自动获取大量抗逆基因序列的程序。此程序灵巧、精悍,Perl 语言强大的文本处理功能使程序能实现不同类型基因序列的远程获取。

关键词:BioPerl;E-Utilities;远程自动获取;抗逆基因
中图分类号:Q343.1 **文献标志码**:A **文章编号**:1672-5565(2014)-03-185-04

A BioPerl based remote automatic extraction of NCBI resistance gene sequences

ZHANG Xiaojing, PAN Weimin, CAO Xingqin*
(*School of Biology*, *Xinjiang Normal University*, *Urumchi Xinjiang* 830054, *China*)

Abstract: BioPerl has a good performance to acquire and process large amounts of biological data. NCBI provides BioPerl interface to directly access its subordinate Entrez databases (including PubMed), the E-Utilities. In order to get NCBI gene sequences automatically to build a secondly database for biological stress resistance genes, in this paper, We designed BioPerl based program to extract stress tolerance gene sequence program. This program is smart, lean, and with powerful text processing capability of Perl language, it can achieve different types of remote access to gene sequences for most of the researchers and provide a better way to gain a specific sequence.

Keywords: BioPerl; E-Utilities; Remote automatic download; Resistance gene

Perl 语言是所有编程语言中最擅长文字处理的语言[1],BioPerl 不仅具有 Perl 语言的所有优点,而且此模块中还包含大量获取分析生物数据的小模块,对于现在的研究者来说,不论是在生物领域还是计算机科学领域,都是一个非常棒的工具。

近几年来,关于抗逆基因的研究越来越受研究者们的关注,无论对于恶劣环境地区植物的生长,还是对于动植物抗病害来说,都是非常重要的。一般情况下,想要获取全面准确的抗逆基因数据,现行的方法只有手动去 NCBI 搜索或者手工记录(例如:DRASTIC INSIGHTS 网站中的抗逆基因数据都是通过手工记录搜集的[2]),一般搜索出的数据都是几万条,甚至几十万条,这其中大部分还不符合条件,需费时去筛选,最后才能下载,是一件耗时又耗力的工作;而对于需要搭建生物抗逆基因二次数据库的项目来说[3],这又是必须要实现的前提,只有保证数据源全面准确,才可称为有价值的二次数据库。目前也有类似解决这个问题的研究[4-6],但并没有达到很好的效果。

相比之下,利用 BioPerl 使用 NCBI 提供的 E-Utilities 编程接口,全面结合抗逆基因关键词来编写的程序,可以快速有效地远程自动获取抗逆基因,数据比较全面准确。在程序方法设计时,以 LEA 基因为例(其他抗逆基因类似,只需将对应的关键词换

收稿日期:2014-04-21;**修回日期**:2014-05-09.
基金项目:自治区自然科学基金(2010211022);新疆师范大学研究生科技创新基金(20131203)资助。
作者简介:张晓婧,女,硕士生,研究方向:生物信息学;E-mail:313741033@qq.com.
*通信作者:曹兴芹,女,副教授,研究方向:生物信息计算;E-mail:caoxin1900@163.com.

掉即可)。

1 程序设计

1.1 程序运行环境

程序环境:Windows XP+ActivePerl5.16.1 Build+BioPerl 1.6.1,以上的安装配置均参照 BioPerl 网站中 Installing BioPerl on Windows 文件[7]。

1.2 关键词的筛选

关键词即本程序的检索条件,为了获得更全面、更准确的序列,一定要筛选出最合适、最全面的关键词。首先得到的关键词,是从抗逆基因的定义及特征中总结出的,共 10 个,2012 年白琳的硕士学位论文《植物抗逆基因资源平台的构建与分析》中也提到了这几个关键词[6],证明此处关键词的准确性;其次,在大量的抗逆基因文献中[8-9],提取出了较完整的抗逆基因种类,从中可以总结出所有抗逆基因种类直接描述的关键词,共 29 个,可靠性可以得到保证;最后是关于 LEA 基因的关键词,现今对 LEA 基因的研究越来越多,相关文献也涌现出许多,在这些文献中将 LEA 基因家族进行分类[10-11],其中有两个 LEA 基因族,文献中有提到过它们的别名,即 LEA2 族也被称作 dehydrin(脱水素),LEA4 族也被称作 seed maturation protein,由此又得到两个关键词,且这两个关键词在 LEA 基因序列中也得到验证,如 AF031248.1、AY044271.1 等。具体关键词列表(见表 1)。

表 1 关键词列表

Table 1 Keywords list

根据抗逆基因定义及特征得到的关键词	直接描述的关键词	根据 LEA 基因族分类中得到的关键词
resistance、resistant; threatened、threaten; inducible; tolerant、tolerance; stressed、stress; defense	bZIP;LEA、late embryogenesis abundant proteins; MYC;AFPs、antifreeze;APX;mechanosensitive channel、ion channels; osmotic;AQP、aquaporin; CDP;DREB、dehydration responsive element;MAP、 Mitogen-activated protein kinase;MYB、myeloblastosis; RLKs、Receptor-like protein kinases、LRR、 Leucine-rich repeat;chaperone;pro;Phospholipase; Phospholipase C、PLC;betaine	dehydrin、 seed maturation protein

1.3 程序方法设计

LEA 蛋白基因(late embryogenesis abundant proteins, LEA)是一类具有重要抗逆功能的抗逆基因,特别是在抗干旱、高盐、高温等环境胁迫方面显示出强大的保护功能[10-11]。程序以 LEA 基因为例,源代码见图 1。

```
use Bio::DB::EUtilities;
use Bio::SeqIO;

my $eutil = Bio::DB::EUtilities->new(-eutil      => 'esearch',
                                     -email      => 'mymail@foo.bar',
                                     -db         => 'nucleotide',
                                     -term       => 'LEA[ALL] OR late embryogenesis abundant proteins[ALL] OR dehydrin[ALL]
                                     -usehistory => 'y'
                                     );

my $ct    = $eutil->get_count;
print "Count = ",$eutil->get_count,"\n";

my $hist  = $eutil->next_History || die "No history data returned";

$eutil->set_parameters(-eutil   => 'efetch',
                       -rettype => 'gb',
                       -history => $hist);

$eutil->get_Response(-file => 'lea-keyword.gb');
```

图 1 程序代码

Fig.1 The code of program

程序具体实现过程如下:第一步,调用 esearch 服务端程序,此服务端程序的作用是可以根据给定条件来查询序列[12],这里用“LEA[ALL] OR late embryogenesis abundant proteins[ALL] OR dehydrin[ALL] OR seed maturation protein[ALL] AND 0:3000[SLEN]”作为关键词条件,“0:3000[SLEN]”这个条件将检索范围缩小到长度为 0~3 000 bp 的序列,由于 esearch 只能进行检索序列的工作,而不具有下载功能,所以在程序后半部分需要用到另一个服务端程序 efetch。第二步,在用 efetch 程序之

前,需要一个中间变量,储存上一步的检索数据,为下一步提供下载的原始记录,这里先将 esearch 程序中"-usehistory"参数设为"y"[13],保存历史浏览记录(注意这里只是缓存数据,并没有直接下载到本地),将历史浏览数据赋值给中间变量 $hist,为下步做好准备;第三步,用到 efetch 服务端程序,将历史数据以'genbank'格式下载到本地。

2　结果与讨论

2.1　程序结果

本文设计的程序为实现远程自动获取大量基因序列提供了一种较好的解决办法。程序在 Windows XP 平台下经测试运行稳定,跨平台移植性好。该程序从 NCBI 中获取 LEA 基因序列共 47 061 条(3 000 bps以内的序列),截止 2013 年 11 月 12 日。

此程序将远程获取的序列数据存储到'lea-keyword.gb'文件中,在这里可以注意到本文程序可以大量自动下载到'genbank'格式文件,而以往在 BioPerl 中用到 E-Utilities 这个接口时,只能大量下载到'fasta'或'xml'格式的文件,要下载'genbank'格式文件只能是小量下载(即给定 gi 号来下载),笔者也曾试过用下'fasta'格式文件的方法来下载'genbank'格式文件,虽然是下载成功了,可是'genbank'文件中的结构已经完全不同了(见图 2),内容虽然完整,可是格式完全变了,作为数据库的源数据是不可能的,fasta 格式中又不包含特征表的内容,而 xml 格式更不适合,白琳的硕士学位论文《植物抗逆基因资源平台的构建与分析》中[6],下载到的便是 xml 格式文件,她之后还需要从中提取出 gi 号,再根据这个下载'genbank'格式文件,程序变得很复杂。所以本文的程序至少有两点好处:一是打破以往只能小量下载'genbank'文件的限制,二是不需要中间转换程序,便捷、灵巧。

```
Seq-entry ::= set {
  class nuc-prot ,
  descr {
    source {
      genome genomic ,
      org {
        taxname "Arabidopsis thaliana" ,
        common "thale cress" ,
        db {
          {
            db "taxon" ,
            tag
              id 3702 } } ,
        orgname {
          name
            binomial {
              genus "Arabidopsis" ,
              species "thaliana" } ,
          mod {
            {
              subtype ecotype ,
              subname "Columbia" } } ,
          lineage "Eukaryota; Viridiplantae; Streptophyta; Embryophyta;
Tracheophyta; Spermatophyta; Magnoliophyta; eudicotyledons; core
eudicotyledons; rosids; malvids; Brassicales; Brassicaceae; Camelineae;
Arabidopsis" ,
          gcode 1 ,
          mgcode 1 ,
          div "PLN" } } ,
```

图 2　Genbank 文件

Fig.2　Genbank document

本程序中用到的关键词条件,是目前查询条件中较全面、准确的条件,不仅涵盖了表面意思中得到的关键词,还考虑到基因本身功能及基因分类之后的别名,使得自动获取的抗逆基因更加全面、准确,为生物二次数据库的构建打好了坚实的基础。

2.2　讨　论

2009 年 5 月,NCBI 创建了 Eutilities 编程接口之后,BioPerl 便添加了 Bio::DB::Eutilities 对象包,此对象包可以使 Perl 调用 Eutilities 所包含的所有功能,可远程操作 NCBI 中的序列数据,为生物信息学的发展带来了福音。Eutilities 包括 8 个服务端程序:efetch、esearch、einfo、egquery、esummary、elink、espell、epost[13],这些服务端程序不仅可以检索 NCBI 数据库,从中远程下载序列,还可以向数据库提交序列,返回序列中所有相关数据库信息等,并且还可以检索 PubMed 数据库,大家都知道 PubMed 数据库中的信息量非常庞大,且具有很高的利用价值,但却不容易提取出来,Eutilities 为它提供了可能性。同时,Perl 语言是最强大的文本处理程序语言,有这两者的结合,相信不久的将来 PubMed 中的数据也能被广泛的利用。

利用 BioPerl 可以处理大部分序列分析的工作,例如它可以读取大量的序列格式,Fasta、Genbank、EMBL、PIR、GCG 等,可以远程获取序列数据,不仅

免去了格式转化的麻烦,还解决了手工获取序列时费时费劲的不便。且 Bio:: SeqIO 对象包不仅能读取多种格式,包括 Fasta、EMBL、GenBank、PIR、Swissprot、GCG、SCF、phd/phred、Ace、fastq、exp、chado 以及 raw(plainsequence) 等,还可以阅读一个大文件(其中包含许多序列信息),将其中每条序列信息读出,对于分析生物数据来说解决了许多费时费力的工作。

3 结束语

虽然本文程序解决了远程自动获取大量序列的问题,但是由于使用关键词来作为检索条件,还是属于模糊查询范畴,在下载的结果中,或多或少会存在一些不太正确的序列,这并不是关键词的问题,因为关键词只是一类序列的简单描述,并没有从深层面去考虑序列的特征,例如我们也许可以从序列的特征表出发,应该可以做到精确查询。还有一点就是,本文程序查询的数据库比较单一,如果以后能将 PubMed 文献数据库加以利用[14],一定能比现在达到更好的效果。

BioPerl 一直以来都是生物信息学家的首选工具,它不仅具有上文所说的序列格式转化功能、远程下载功能,并且 BioPerl 还能识别限制性酶切位点,可分析 blast 的结果,可操作系统发育树等等,大部分在生物中要需要分析的功能,基本上都包含了。再加上 Bio::DB::Eutilities 对象包,对于 BioPerl 来说真是如虎添翼,这些在很大程度上,推动了生物信息学这门新兴交叉学科的发展,并为这门学科做出了很卓越的贡献,是人们在研究这块领域时,首先选择的工具。

参考文献(References)

[1] TOM P.Perl 语言入门(第六版)[M].盛春译.江苏:东南大学出版社,2012:130-179.
TOM P. Introduction to the Perl language[M]. SHENG Chun.Jiangsu: Southeast university press,2012:130-179.

[2] Gary L. Drastic insights[EB/OL]. http://www.drastic.org.uk/,2014-6-16.

[3] 邢仲璋,林王源,林毅申.基于 bioperl 的生物二次数据库建立及应用[J].计算机系统应用,2004,11(11):58-60.
XING Zhongzhang, LIN Wangyuan, LIN Yishen. Based on the bioperl biological secondary database establishment and the application [J]. The Computer System Application, 2004,11(11):58-60.

[4] 向福,余龙江,栗茂腾.用 bioperl 实现种子植物 18srRNA 基因序列的大规模获取[J].华中农业大学学报,2005,24(4):330-333.
XIANG Fu, YU Longjiang, JIA Maoteng. By bioperl implementation of seed plant large-scale access srrna gene sequences of 18[J]. Journal of Huazhong Agricultural University, 2005,24(4):330-333.

[5] 向福,余龙江,陈悟.基于 Bioperl 的基因序列获取的程序设计与实现[J].生物技术,2004,14(6):64-66.
XIANG Fu, YU Longjiang, CHEN Wu. Based on the bioperl gene sequence for program design and implementation[J]. Biotechnology, 2004,14(6):64-66.

[6] 白琳.植物抗逆基因资源平台的构建与分析[D].浙江:浙江大学生命科学学院,2012:7-9.
BAI Lin. Plants to genetic resources platform construction and analysis[D].Zhejiang: Zhejiang University College of Life Science,2012:7-9.

[7] BioPerl. Installation[EB/OL]. http://www.bioperl.org/wiki/Installing_BioPerl,2014-4-19.

[8] 高银.植物抗逆机制与基因工程研究进展[J].内蒙古农业科技,2007,6(5):75-78.
GAO Yin. Plants to mechanisms and gene engineering are reviewed[J]. Inner Mongolia Agricultural Science and Technology,2007,6(5):75-78.

[9] 杨柳,张振乾,宋继金.植物抗逆基因研究进展[J].作物研究,2010,4(1):126-129.
YANG Liu, ZHANG Zhenqian, SONG Jijin. Plants gene research progress[J]. Crop Research,2010,4(1):126-129.

[10] 李乐,许红亮,杨兴露.大豆 LEA 基因家族全基因组鉴定、分类和表达[J].中国农业科学,2011,5(5):3945-3954.
LI Le, XU Hongliang, YANG Xinglu. Soybean LEA gene families genome-wide identification, classification and expression[J]. Scientia Agricultural Sinica,2011,5(5):3945-3954.

[11] 白永琴,杨青川.LEA 蛋白研究进展[J].生物技术通报,2009,9(9):1-5.
BAI Yongqin, YANG Qingchuan. LEA proteins is reviewed [J]. Biological Technical Bulletin,2009,9(9):1-5.

[12] 夏武青,葛芬,宋霞.基于 NCBI 开放编程接口的局域网 PubMed 检索平台设计与实现[J].中华医学图书情报杂志,2012,21(8):66-69.
XIA Wuqing, GE Fen, SONG Xia. Local area network (LAN) based on NCBI open programming interfaces PubMed retrieval platform design and implementation[J]. The Chinese Medicine Books Intelligence Magazine, 2012,21(8):66-69.

[13] 许丹,朱斐.从 PubMed 数据库中挖掘生物医学中的十大热点话题[J].计算机与现代化,2013,1(1):192-199.
XU Dan, ZHU Fei. In the biomedical PubMed database mining top ten hot topic[J]. Computer and Modern, 2013,1(1):192-199.

[14] PHILIPPE T, JOHANNES S, ALEXANDER V. GeneView[J]. Nucleic Acids Research,2012,6(6):585-591.