Hans Journal of Computational Biology 计算生物学, 2016, 6(2), 27-32
Published Online June 2016 in Hans. http://www.hanspub.org/journal/hjcb
http://dx.doi.org/10.12677/hjcb.2016.62004

Hans汉斯

Realize Accurately the Download of *Populus trichocarpa* Protein Sequences Based on BioPerl

Pengfang Xie*, Jiarong Huang#

College of Forestry, Henan Agricultural University, Zhengzhou Henan
Email: 564345631@qq.com, #huangjiarong137@163.com

Received: Jun. 2nd, 2016; accepted: Jun. 16th, 2016; published: Jun. 21st, 2016

Abstract

Based on *Populus trichocarpa* (*Populus trichocarpa*) and Protein as keywords, a program was designed to download the protein sequences *of Populus trichocarpa* family based on the BioPerl. This program laid a solid foundation for building secondary database of *Populus trichocarpa* protein, and it also provided a quick and accurate method for accurately acquiring the protein sequences of *Populus trichocarpa* for researchers who studied the protein of *Populus trichocarpa*.

Keywords

***Populus Trichocarpa*, Protein, Protein Sequences, Bioperl**

基于BioPerl实现精确下载毛果杨蛋白质序列

谢鹏芳*，黄家荣#

河南农业大学林学院，河南 郑州
Email: 564345631@qq.com, #huangjiarong137@163.com

收稿日期：2016年6月2日；录用日期：2016年6月16日；发布日期：2016年6月21日

*第一作者。
#通讯作者。

文章引用: 谢鹏芳, 黄家荣. 基于 BioPerl 实现精确下载毛果杨蛋白质序列[J]. 计算生物学, 2016, 6(2): 27-32.
http://dx.doi.org/10.12677/hjcb.2016.62004

摘　要

本文以毛果杨(*Populus trichocarpa*)和蛋白质名称为关键词，基于BioPerl设计了下载毛果杨家族蛋白质序列的程序。此程序为构建毛果杨蛋白质二次数据库奠定了基础，也为以毛果杨蛋白质为研究对象的研究人员提供了一个方便快捷精确获取毛果杨蛋白质序列的手段。

关键词

毛果杨，蛋白质，蛋白质序列，BioPerl

1. 引言

蛋白质是一切生命的物质基础，是生理功能的执行者，是生命现象的直接体现者，对蛋白质结构和功能的研究将直接阐明生命在生理或病理条件下的变化机制。随着人类基因组计划的完成，生命科学研究已进入了后基因组时代。在这个时代，拟南芥(*Arabidopsis thaliana*)、水稻(*Oryza sativa*)和毛果杨(*Populus trichocarpa*)等植物全基因组序列的测定以及基因组学的深入研究已经完成，而植物蛋白质组学研究已成为后基因组时代的热点之一，因此也产生了海量的生物数据[1]-[6]。面对这些数据，以往都是通过数据库手动搜索，还要对多余信息去除后才能下载到所需的蛋白质序列，在信息量大、资源繁杂的数据库面前，研究者在获取自己所需要的蛋白质序列时往往要耗费大量的时间[7] [8]。在基于 BioPerl 的生物信息数据下载研究方面，相对草本、农作物等其他植物，木本植物的下载研究还未见报道；相对基因序列，蛋白质序列的下载研究极少见报道[9]-[12]。针对这一问题，为让林学研究者利用程序获得自己所需的木本植物蛋白质序列，本文基于 BioPerl 设计了精确下载毛果杨蛋白质序列的程序，并以毛果杨的组蛋白去乙酰化酶 HDAC [13]和铵转运蛋白 AMT [9]为例进行说明。

由于杨树基因组相对较小、且具有周期短、生长快、遗传转化容易等特点，近年来已成为木本植物中的模式物种，由美国橡树岭国家实验室和能源部联合基因组研究所领导的一个联合研究组已于 2006 年完成属于白杨派的毛果杨(*Populus trichocarpa*)的基因组草图。组蛋白去乙酰化酶 HDAC 和铵转运蛋白 AMT 的研究主要集中在拟南芥、水稻等草本植物中，而对木本植物中的研究相对较少。对组蛋白去乙酰化酶 HDAC 和铵转运蛋白 AMT 的研究大多是基于杨树基因数据库(*Populus trichocarpa*)完成的，杨树基因数据库以更新到 3.0 版本(*Populus trichocarpa* v3.0)。

2. 程序方法设计

2.1. 程序运行环境

程序环境：Windows7+ActivePerl 5.20.2 Build+BioPerl 1.6.9，安装配置参照 BioPerl 网站中 Installing BioPerl on Windows 文件[14]，文件中提到了三种安装方法：1) GUI 图形界面安装；2) PPM 命令安装；3) 利用 CPAN 或者手动安装。

第一种方法操作简单，但是打开界面速度慢，容易死机，而且所给的安装网页也都不是最新版本；第二种 PPM 命令行安装，首先要确保有 PPM-Repositories 模块，这里安装也都不是最新版本；所以我们选择第三种安装方式，利用 CPAN 安装，既可以获得最新版本，又可以选择安装的依赖关系。

在安装 bioperl 前，首先要确定 Perl 是否安装成功。检验安装成功后，就可以安装 bioperl 了：1) 打开命令窗口，输入 cpan，安装 MinGW 包；2) 输入 d /bioperl/，出现可安装的版本和模块；3) 输入 force

install CJFIELDS/BioPerl-1.6.924.gz，等待系统下载安装，安装完成后可输入 perldoc Bio::SeqIO 检验 BioPerl 是否安装成功；4) Bioperl-DB、bioperl-network、bioperl-run 模块安装与 bioperl 安装一样，一般我们所需要安装的就是这 4 个模块，还可以根据自己的需要进行其他模块的安装，方法是一样的。

2.2. 程序设计

程序设计流程图如图 1。程序先根据我们给定的关键词进入 GenBank 数据库，读取数据库中的序列条目(genbank entry)，每次读取一个序列(next_seq)，取得序列成功后程序指向序列的 FEATURES 部分(get_SeqFeatures)，并对其进行判断解析以确定是否满足条件。执行完条件判断后，满足条件的就将此序列下载下来，并打印其 display_id，不满足条件的就读取下一条序列，如此反复循环，最终将所有符合条件的毛果杨 HDAC 蛋白家族序列下载下来。图中虚线框部分可以替换，以适应不同需求的蛋白质序列的获取。

2.3. 毛果杨 HDAC 家族蛋白质序列的获取

第一步是根据毛果杨和蛋白质名称这两个关键词来检索毛果杨蛋白质序列条目。

第二步是利用毛果杨蛋白质产物中“family”来匹配上一步检索的序列是否符合条件，如果成功匹配就将其下载下来，并以其“accession 号”来命名序列文件，匹配不成功就再次读取序列条目进入循环，直到匹配成功，最后完成精确下载毛果杨蛋白质家族序列的任务。如图 2 所示，是程序下载毛果杨 HDAC 家族蛋白质的核心代码。

3. 程序运行结果与分析

3.1. 程序运行结果

打开 cmd，将此程序在命令符中运行，执行 perl 脚本，等待一会儿后，毛果杨 HDAC 家族蛋白质数据便成功下载到本地文件中，以其“accession”编号命名序列文件。将程序中关键词 HDAC 换成 AMT，程序就可以实现下载毛果杨 AMT 家族蛋白质序列。此程序实现了精确远程下载毛果杨蛋白质序列。如图 3、图 4 所示，是程序运行结果。

程序将下载到的蛋白质序列保存到所运行代码目录下，并以蛋白质“accession”编号命名保存每条蛋白质序列信息，文件格式保存为 genbank。用记事本打开.gbk 文件，可以获取蛋白质的一些基本信息，比如：序列名称(LOCUS)，序列简单说明(DEFINITION)，序列编号(ACCESSION)，序列版本号(VERSION)，与序列相关的关键词(KEYWORDS)，序列来源的物种名(SOURCE)，文献(REFERENCE)，特性表(FEATURES)，碱基组成(BASE COUNT)及碱基排列顺序(ORIGIN)等。

3.2. 程序运行结果分析

在 NCBI 中输入毛果杨和 HDAC 这两个关键词检索蛋白质，最后检索出来 45 条结果，输入毛果杨和 AMT 这两个关键词检索出 14 条结果，这些数据量相对较少，可以一个一个进行筛选，去除不需要的假定蛋白质，找到所需的家族蛋白质。但要是数据量大的情况下，在一个一个的筛选将要耗费大量的时间和精力，而且准确率也不能保证。

如果将关键词中的 *Populus trichocarpa* 写成 populus，在 NCBI 中检索出来的将会分别有 196 条和 102 条结果，其中包含很多假定蛋白和预测蛋白，甚至有其他物种的预测蛋白，这显然不是研究者所想要的。而将程序关键词中的 *Populus trichocarpa* 换成 populus，最后下载保存得到的结果与之前是一样的，这表明程序循环语句中的标签匹配是可以达到精确下载蛋白质序列的，这与研究的目的相符。

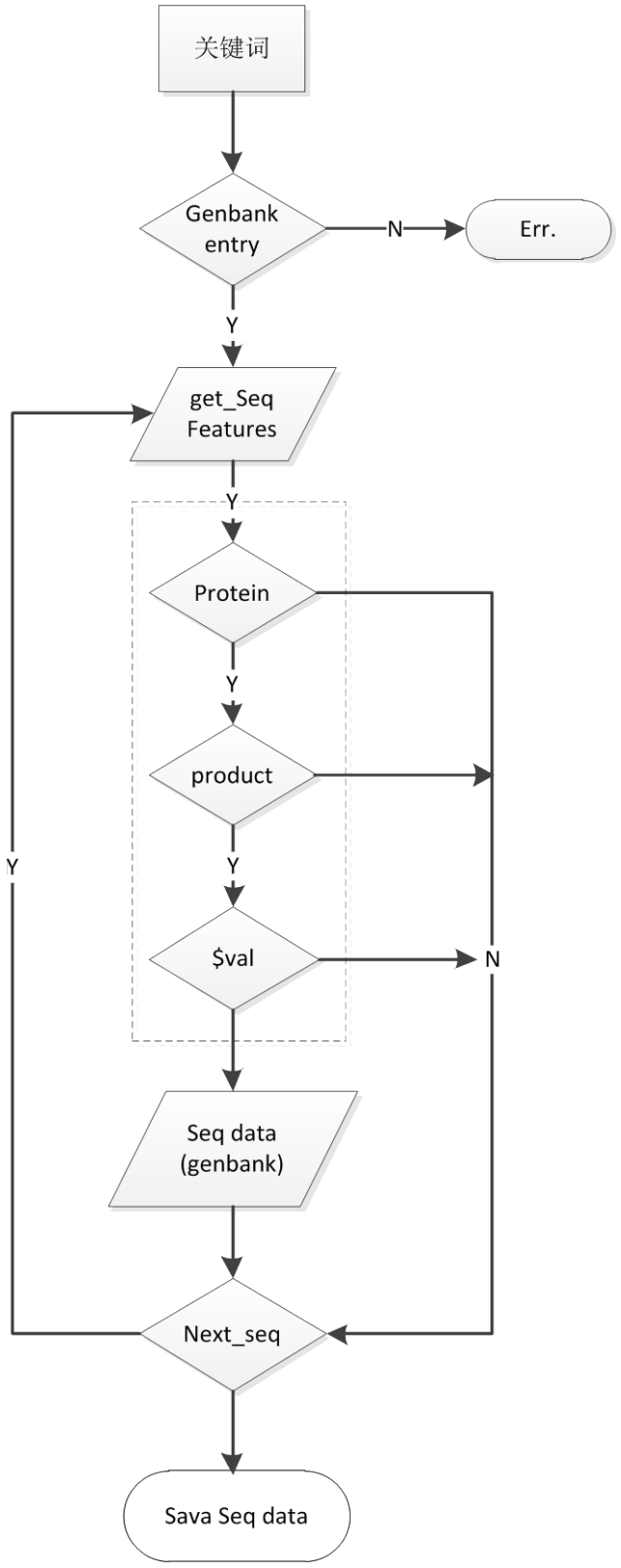

Figure 1. Flow sheet of program
图 1. 程序流程图

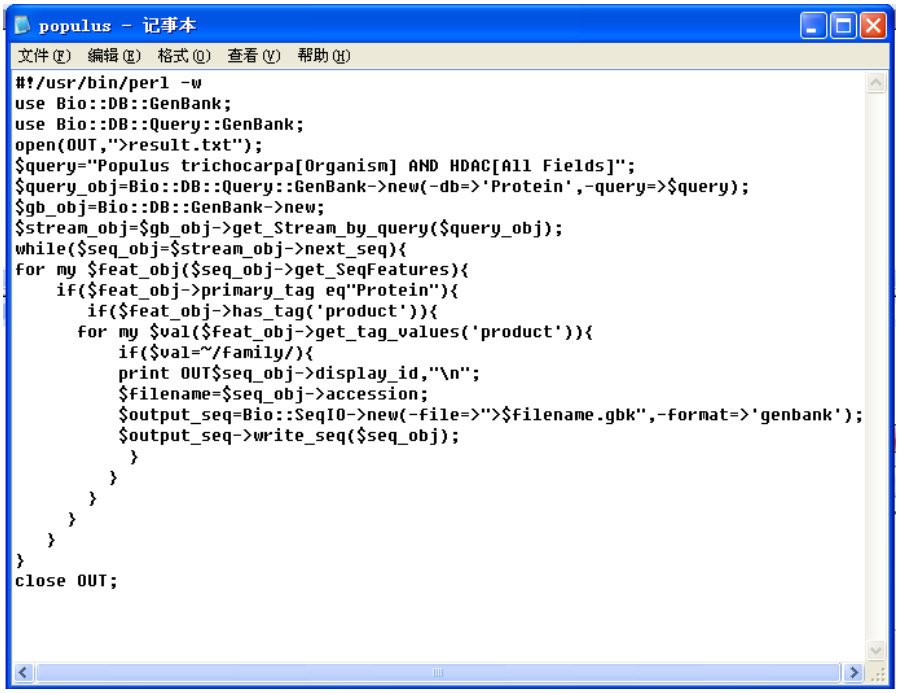

Figure 2. The core code of program
图 2. 程序核心代码

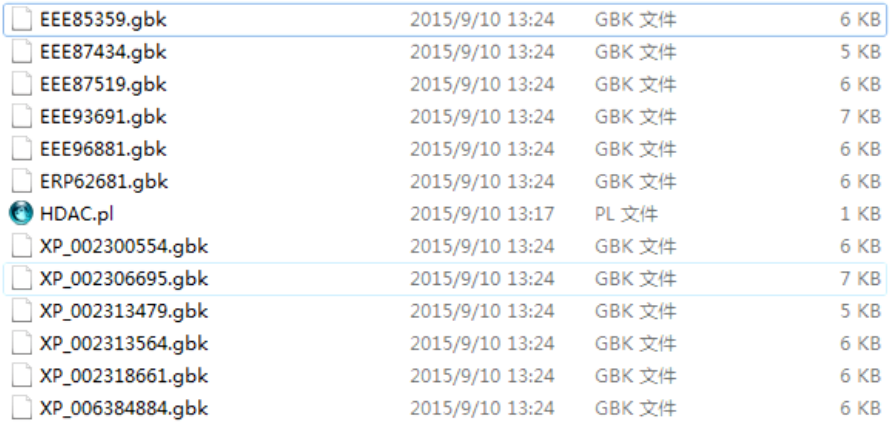

Figure 3. The running result of program HDAC.pl
图 3. HDAC.pl 脚本程序运行结果

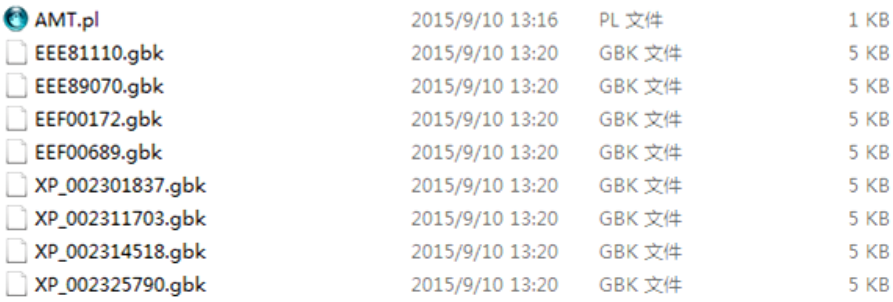

Figure 4. The running result of program AMT.pl
图 4. AMT.pl 脚本程序运行结果

蛋白质序列下载可以根据研究需求的不同保存为不同格式的文件。若只需要提取蛋白质序列进行序列分析功能研究，可以保存为 FASTA 格式，这是一种基于文本用于表示核苷酸序列或氨基酸序列的格式。在这种格式中碱基对或氨基酸用单个字母来编码，且允许在序列前添加序列名及注释。若需蛋白质的详细信息，GENBANK 格式则是首选，序列文件的基本单位是序列条目，包括核甘酸碱基排列顺序和注释两部分。也可保存为 EMBL、PIR 等格式。

4. 结论与讨论

程序运用了两个例子进行了验证，又分别与在 NCBI 中检索出来的结果进行了分析对比，相比较之下 Perl 程序的蛋白质序列智能精确下载远远优于在 NCBI 中检索下载，这与许多之前研究的基因序列批量下载，或者说是基因序列精确下载是大不相同的。

1) 本文实现了从 NCBI 中精确下载毛果杨蛋白质序列。程序设计中，在程序流程图中画虚线框部分可以根据需求进行更换，这提高了程序的应用性。

2) 程序改变了以往通过数据库手动搜索，还要对多余信息去除后才能下载到所需的蛋白质序列的格局，大大提高了研究者的效率。

3) 程序也为构建毛果杨蛋白质二次数据库奠定了基础，但想要为以毛果杨蛋白质为研究对象的研究人员提供了一个方便快捷的研究平台还需要进一步的研究探索。

4) 本研究中程序关键词的更改只能在程序脚本中实现，这为不同蛋白质的精确下载带来了不便，因此下一步的研究就是程序图形界面，以方便今后不同蛋白质的精确下载。

项目基金

河南省科技攻关资助项目(项目编号：0624050007)。

参考文献 (References)

[1] 罗静初. 分子生物信息数据库[A]. *The 2nd Cross-Straits Symposium on Biology-Inspired Theoretical Problems—Proceedings of CCAST (World Laboratory) Workshop*. 北京：中国高等科学技术中心, 2000: 41.

[2] 吴大强, 蔡诚, 魏国, 等. 毛果杨全基因组 NBS 类型抗病基因分析[J]. 林业科学, 2009, 45(2): 152-157.

[3] Barker, W.C., Garavelli, J.S., Huang, H., *et al*. (2000) The Protein Information Resource (PIR). *Nucleic Acids Research*, **28**, 41-44.

[4] 万跃华, 何立民. 网上生物信息学数据库资源[J]. 情报学报, 2002, 21(4): 497-512.

[5] Stoesser, G., Baker, W., van den Broek, A., *et al*. (2003) The EMBL Nucleotide Sequence Database: Major New Developments. *Nucleic Acids Research*, **31**, 17-22. http://dx.doi.org/10.1093/nar/gkg021

[6] 喻娟娟, 戴绍军. 植物蛋白质组学研究若干重要进展[J]. 植物学报, 2009, 44(4): 410-425.

[7] 李淑娟, 张超, 等. 毛果杨 HDAC 基因家族序列及其表达分析[J]. 西北农林科技大学学报(自然科学版), 2015, 43(3): 63-76.

[8] 安飞飞, 李庚虎, 陈霆, 等. 植物耐寒生理及蛋白质组学研究进展[J]. 中国农学通报, 2015, 3(14): 96-101.

[9] 李磊, 罗杰, 李红, 罗志斌. 毛果杨全基因组铵转运蛋白家族成员及其序列分析[J]. 西北农林科技大学学报(自然科学版), 2011(2): 133-142.

[10] 向福, 余龙江, 栗茂腾, 等. 用 bioperl 实现种子植物 18S rRNA 基因序列的大规模获取[J]. 华中农业大学学报, 2005, 24(4): 330-333.

[11] 向福, 陈悟, 余龙江. 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.

[12] 张晓婧, 曹兴芹, 潘伟民. 基于 BioPerl 实现从 NCBI 中精确下载 LEA 基因序列[J]. 计算生物学, 2014, 4(2): 13-19.

[13] 周猛, 童春发, 施季森. 充分利用 Bioperl 加速生物信息学的研究[J]. 生物信息学, 2008, 6(1): 43-45.

[14] BioPerl. Installation. http://bioperl.org/INSTALL.html