基于 Bioperl 的生物二次数据库建立及应用①

Establishment and Utilization of a Bioperl－based Bioinformatical Secondary Database

邢仲璟　林丕源　林毅申　（广州华南农业大学计算机科学系 510642）

摘要：在生物信息学中，建立生物二次数据库可以针对特定物种进行更深入的研究。本文以人类 EST 基因为例，探讨了利用生物信息学专业软件包 Bioperl 建立一个生物二次数据库，并应用以构建一个生物信息研究平台。

关键词：生物信息学　生物二次数据库　Bioperl

1　生物二次数据库

一般而言，生物信息数据库可以分为一级数据库和二级数据库。一级数据库的数据都直接来源于实验获得的原始数据，只经过简单的归类整理和注释。国际上著名的一级核酸数据库有 Genbank 数据库、EMBL 核酸库和 DDBJ 库等；蛋白质序列数据库有 SWISS－PROT、PIR 等；蛋白质结构库则有 PDB 等。这些数据库都是建立二次数据库的主要数据来源。二级数据库是在一级数据库、实验数据和理论分析的基础上针对特定目标衍生而来的，是对生物学知识和信息的进一步整理。

2　生物信息专业软件包 Bioperl

Bioperl 功能全面，且源代码全部开放，是生物信息学研究的有效工具。作为基于 Perl 语言的软件包，Bioperl 软件包具有所有 Perl 的优点。Perl 是一种功能非常全面的开发语言，它小巧、灵活，适用于模块化的编程，更重要的是 Perl 有强大的正则表达式模式匹配功能，生物信息学大部分的工作恰是在数据中进行模式搜索匹配；另一方面，Perl 的容错性非常适用于通常不完全的生物序列数据；而 Perl 的元件导向性则促进了程序的模块化处理。Bioperl 综合了以上优点，形成了功能极其全面的专业软件包，非常适用于进行生物信息各方面的处理。

Bioperl 软件包一共分为以下几个功能模块[3]：

(1) Bioperl 核心：Bioperl 基本程序集合，提供多种操作基类和数据基本接口；

(2) Bioperl－Ext：Bioperl 扩展，里面包含了大量生物信息学常用算法的源代码；

(3) Bioperl－DB：二次数据库结构及数据操作接口；

(4) Bioperl－Run：外接程序接口，可以很方便地调用各种第三方程序。

本文将介绍使用 Bioperl－DB 模块构建二次数据库的基本结构和基本接口，以及使用 Bioperl 核心模块构建生物信息系统的基本环境，并将二者组成一个完整的生物二次数据库及生物信息研究平台。

3　生物二次数据库分析及建立

3.1　生物二次数据库基本结构

Bioperl－DB 模块提供了生物二次数据库的基本结构定义（定义表结构的 SQL 语句请查看 sql/basicseqdb－mysql.sql 文件），使用的数据库后台是 Mysql。在 Bioperl－DB 所定义的结构中，由于是为兼容多种格式的数据而制订，而实际使用中只需要其中的几个表，因此存在一定的冗余。图 1 是 Bioperl－DB 中所定义关键部分的 E－R 图。

图 1 中基本表 biodatabase 存放的是数据库的名称（在此二次数据库结构中，可以存放多种不同类型数据库中的数据，比如 DNA 序列数据、蛋白质序列数据等，因此，使用此表标明存放的不同种类的数据）。表 bioentry 则是二次数据库序列目录，存放了序列的 id 号、版本号等，它们与表 bioentry_keyword、表 bioentry_reference 组成了二次数据库的核心。表 biosequence 存放的是具体序列的数据，表 comment 与表 reference 分别存放序列的注释和序列信息的引用文献资料。

以上的表结构显示，使用这个生物二次数据库的数据结构，可以存放多种类型的数据，同时也可以存放对应的文献资料数据，基本能够满足日常生物信息研究的需要。

3.2　数据转换接口

在生物二次数据库的设计中，另一个关键问题是接口，即二次数据库与数据源的连接及转换模块。为了使用户能

① 本文得到华南农业大学新学科扶持基金资助

在二次数据库中使用更多不同的数据源，必须设计一个连接二次数据库与各种不同生物信息数据源的接口。Bioperl 提供了一套完整的数据接口，图 2 是 Bioperl 数据接口的工作结构模型。

该模型按功能被划分为三部分：数据提供层、数据接口层和标准数据层。其中，由数据接口层转换数据提供层到标准数据层，即需要建立的二次数据库中。

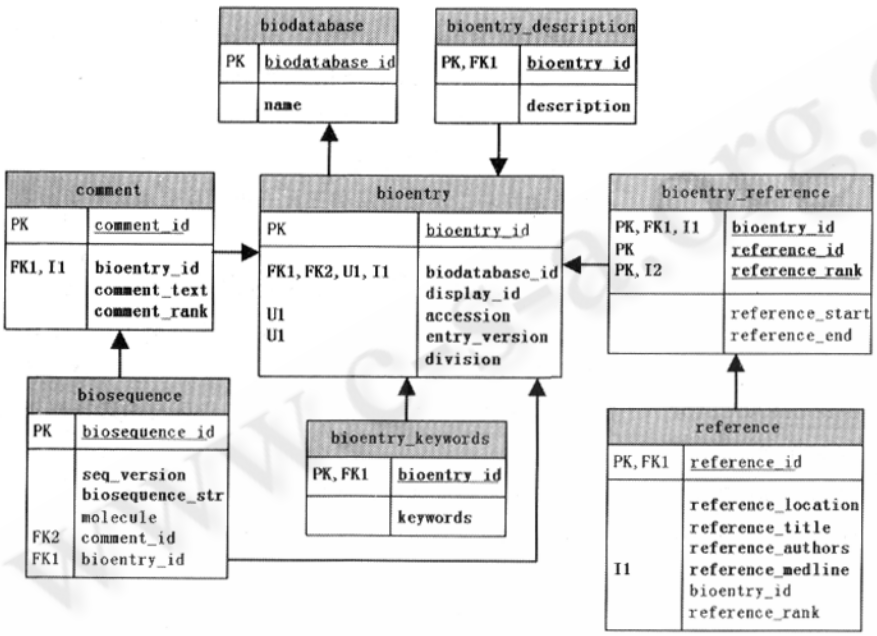

图 1　二次数据库关键部分 E－R 图

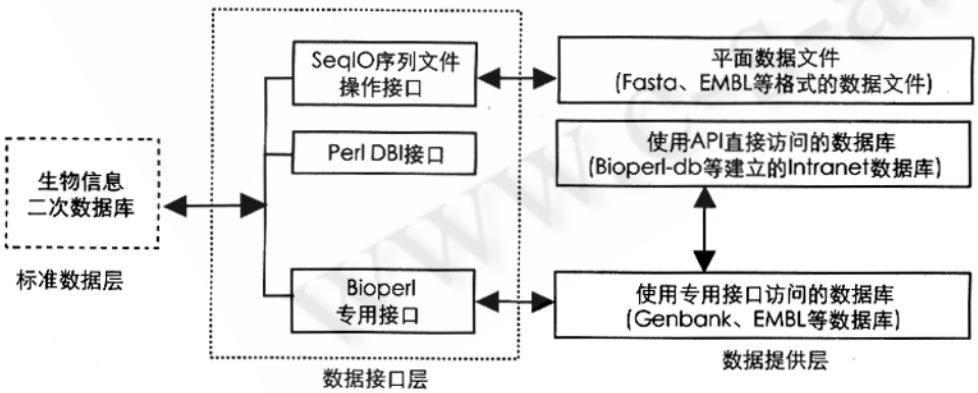

图 2　Bioperl 数据接口结构模型

数据提供层提供的序列数据，一般以数据文件或数据库基本表方式存在。数据提供层的数据根据访问方式不同，可以分为平面数据文件、由 API 直接访问的数据库和由专用接口访问的数据库。平面数据文件是由形如 EMBL 所提供的数据格式文件等所组成；以 API 直接访问的数据库一般是位于 Intranet 上，能够直接进行读写的数据库；专用接口访问的数据库则指 Genbank 等位于 Internet 上公开的大型数据库。针对数据提供层的不同数据源，数据接口层通过应用 Bioperl 软件包提供的各种接口模块对数据文件或数据库进行操作。对于平面数据文件，Bioperl 软件包使用了 SeqIO 序列文件操作接口；对于使用 API 直接访问的数据库，则使用 DBI 数据接口；对于专用接口访问的数据库，使用专用数据接口对数据进行处理。这些处理对使用者而言是透明的，但保证了用户直接使用的二次数据库能通过多种渠道获得尽可能多的相关数据源。

3.3　数据的获取

要建立二次数据库，首先必须准备相关数据源，即获取相应的生物数据。主要的途径包括：

(1) 获取大批量数据。一次数据库(如 Genbank、EMBL 等)是提供生物数据的主要数据源，基本上建立生物二次数据库中所需的数据都从一次数据库中获取。在一次数据库中，会定期把序列数据打包，通过 Ftp 进行发布。因此，可以通过 Ftp 获取发布的各种生物序列(基因、蛋白质等)数据。例如，可以通过 EMBL 的 FTP 服务器：ftp://ftp.ebi.ac.uk/ 和 Genbank 的 FTP 服务器：ftp://ftp.ncbi.nih.gov/genbank/ 获取数据。

(2) 获取小批量或特定数据。可以通过一次数据库提供的专用接口。所谓的专用接口，是一次数据库提供 CGI 程序，用户通过发送参数到此程序，程序会根据要求返回所查询的数据，这样就可以对一次数据库的数据进行查询和获取。比如通过 Genbank 的接口：http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?，使用此接口，可以附加多种参数，比如指定序列号或者模糊搜索物种名称等，可以直接检索序列数据并下载到本地。

(3) 使用者的研究数据。通过普通的数据导入接口或手工输入界面，可以允许用户将自己研究的相关数据放置在

二次数据库中,以供处理平台界面调用。

3.4 导入数据到二次数据库

笔者以人类 EST 序列为例,建立了一个人类 EST 序列的二次数据库。首先,从 ftp://ftp.ebi.ac.uk/下载 EMBL 发布的人类 EST 序列的数据包,共有 55 个文件,下载到本地解压缩后通过以下 Perl 程序导入二次数据库中:

```
#! /usr/local/bin
use Getopt::Long;
use Bio::DB::SQL::DBAdaptor;
use Bio::SeqIO;
my $format = 'embl';
$dbadaptor = Bio::DB::SQL::DBAdaptor->
new(
        -host => 'databasehost',
        -dbname =>'databasename',
        -user => 'user',
        -pass => 'password');
$dbid = $dbadaptor->get_BioDatabaseAdaptor
->fetch_by_name_store_if_needed($dbname);
my $seqadp = $dbadaptor->get_SeqAdaptor;
my $seqio = Bio::SeqIO->new(-file => $file,
-format => $format);
while( $seq = $seqio->next_seq ) {
      print "Sequences ID ". $seq->id."\n";
      $seqadp->store($dbid,$seq);
    }
```

导入基本数据后,余下的工作就是随时将发布在三大数据库中相关的新数据导入到二次数据库中,以下程序,是在 Genbank 中搜索相关序列数据,并把搜索的结果存入二次数据库中:

```
#! /usr/local/bin
#从 GenBank 检索关键字为"human"的序列,并以
GenBank 格式存到 new_file.gb 文件中
use Bio::DB::GenBank;
use Bio::Perl;
use Getopt::Long;
use Bio::DB::SQL::DBAdaptor;
use Bio::SeqIO;

$gb = new Bio::DB::GenBank;
$seq_object = $gb->get_Stream_by_query
("human");
$dbadaptor = Bio::DB::SQL::DBAdaptor->
new(      -host => 'databasehost',
          -dbname =>'databasename',
          -user => 'user',
          -pass => 'password');
$dbid = $dbadaptor->get_BioDatabaseAda-
ptor->fetch_by_name_store_if_needed($dbname);
my $seqadp = $dbadaptor->get_SeqAdap-
tor;
while( $seq = $$seq_object->next_seq ) {
  print "Sequences ID ". $seq->id."\n";
  $seqadp->store($dbid,$seq);}
```

构造好搜索和过滤条件后,可将程序写成一个"机器人",自动同步二次数据库和三大一级数据库中的相关数据。

4 构建基于生物二次数据库的信息平台

在上述已建立好的生物二次数据库基础上,利用 Bioperl 软件包在生物信息的处理方面的强大功能,构建相应的生物信息处理平台显得相当轻松。用户可自由选择以下相关功能:

(1) 从本地或远程数据库中访问序列数据;
(2) 转换数据库或文件记录的格式;
(3) 处理单个序列;
(4) 搜索相似序列;
(5) 创建与处理序列的排列;
(6) 从基因组 DNA 中搜索基因与其他结构。

这些功能以模块与函数的形式实现,便于用户调用以及系统的进一步开发。为特定的物种研究提供了极大的帮助。

5 结束语

建立生物二次数据库,可以针对生物信息学领域内的某一物种进行更深入的研究。专业软件包 Bioperl 中集成的基本数据接口、二次数据库的基本结构以及大量的功能扩展模块,对建立生物二次数据库的工作提供了有效的工具以及灵活的开发环境,此外,也为开发各种生物信息的研究工具及平台提供了便利的手段,对开发工作产生极大的影响。

参考文献

1 北京大学生物信息中心,http://www.cbi.pku.edu.cn/chinese/.
2 张成岗、贺福初,生物信息学方法与实践[M],科学出版社,2002.9。
3 Bioperl 网站.http://www.bioperl.org/
4 Perl Module.http://www.cpan.org/