真核生物内含子数据库构建*

何 淼, 李继东, 张尚宏
(中山大学生命科学学院, 广东 广州 510275)

摘 要: 通过下载国际一级序列数据库 GenBank 的序列数据, 开发研制了真核生物基因内含子数据库 EID, 该数据库由一个主数据库和多个子数据库组成, 子数据库包括细胞核内含子数据库、细胞器内含子数据库, 以及植物、无脊椎动物、灵长类、啮齿类、其他哺乳类和其他脊椎动物等不同类型生物内含子数据库。文中详细介绍了数据的获取和筛选方法, 主数据库中 dEID, pEID, hEID 的文件格式和结构, EID 数据库构建程序, 以及 EID 的查询方式, 提供的分析工具等。EID 将建设成为研究内含子功能和进化起源的生物信息学平台。

关键词: 真核生物, 内含子, 数据库
中图分类号: Q754 文献标识码: A 文章编号: 0529-6579 (2004) S1-0050-06

研究表明, 在微生物中, 非编码序列的区域只占整个基因组序列的 10% ~ 20%。随着生物的进化, 非编码区越来越多, 在高等生物和人的基因组中, 非编码序列已占到基因组序列的绝大部分。表明这些非编码序列必定具有重要的生物功能, 普遍的认识是它们与基因的表达调控有关。

内含子作为非编码序列的重要成分之一, 通常真核生物基因内含子表现自我剪接功能。近年来, 对内含子的进一步研究发现其具有多样性功能的特征。主要表现为以下几个方面, 内含子对基因表达具有增强作用, 即充当增强子, 调控基因表达的重要方式是通过由环状蛋白组成的转录因子作用 DNA 完成[1,2], 内含子可充当启动子的角色[3]。内含子可以充当介导因子[4]; 一个由 RNA 介导的基因调控例子就是线虫的 LINE4 基因, 它调控 LINE14 基因转录后表达, 而它源于一个没有开放读码框的基因中的内含子[5]。第二类则为小核 RNA (snRNA)[6]。内含子能够参与 RNA 编辑[7,8]。内含子对基因表达抑制的作用。内含子不仅对基因表达起增强作用, 在一些基因的表达过程中也起抑制作用。Bornstien 等[9] 在研究人胶原蛋白 α (Ⅰ) 基因的表达调控时, 发现第一个内含子内有一段 247 bp 反向重复序列 A274 可阻碍基因的转录。内含子编码 ORF (开放读码框)。许多小核仁 RNA (snoRNA) 和小核 RNA (snRNA) 都被编码在内含子内[10,11], 而且还有一些别的基因和假基因由内含子来编码[12]。内含子中含有基因的现象非常有趣, 因为它揭示了关于 RNA 加工的信息, 同时也表现出内含子的功能重要性。内含子的具有选择剪接功能。内含子的选择剪接已经被证明在生物发育的不同时期和不同组织里起着很重要的作用。例如, 人的基因组中 30%以上的基因都是经过选择剪接而成的。通过选择剪接, 一个基因可以不同的剪接翻译成几个不同功能但相关的蛋白, 从而提高了遗传的复杂性。极端的时候, 不同的蛋白会有着截然不同甚至相反的功能[13]。内含子在选择剪接过程中可以起标点符号作用来分开外显子, 或提供支持选择剪接机制的作用[14]。

正确掌握内含子的位置信息, 对于研究内含子的功能和进化起源有着至关重要的作用。任何此类数据的错误, 哪怕是很少, 都会对由同源蛋白推导出内含子位置信息作图、研究内含子移动等其它分析研究产生很大的影响。事实上, GenBank 本身含有很多错误; 第一类错误主要是书写错误, 比如有些内含子—外显子边界与实际的边界有偏离; 第二类错误主要是由于基因预测软件的预测误差。GenBank 中很多基因是没有被实验所验证的, 因而这些数据是不可靠的。更麻烦的是, 有些软件需要运用这些数据 (当然也包括实验验证过的数据) 来提高识别外显子和内含子的准确度。本项目研究霜目的就是构建一个富含准确真核生物内含子信息的数据库, 为进一步研究工作的开展, 例如进化和功能研究, 奠定良好的基础。

* 收稿日期: 2003-06-25
基金项目: 国家自然科学基金资助项目 (30270752); 广东省自然科学基金资助项目 (031616)
作者简介: 何淼 (1963 年生), 男, 博士, 副教授; 通讯联系人: 张尚宏; E-mail: lsbrc11@zeu.edu.cn

1 真核生物内含子数据库的构建

1.1 数据的获取

1.1.1 *数据的来源和筛选* 真核生物内含子数据库 (EID) 的数据来自于从 GenBank 下载并提取的真核生物序列信息。GenBank 每两个月发布一次新版本，本数据库的数据来源于 GenBank Release 125 版本。各生物类别的下载的原始数据分别为，无脊椎动物：Gbinv1. seq, Gbinv2. seq, Gbinv3. seq, Gbinv4. seq；脊椎动物：Gbvrt. seq；哺乳动物：Gbmam. seq；灵长类：Gbpril1. seq, Gbpril2. seq, Gbpril3. seq, Gbpril4. seq, Gbpril5. seq, Gbpril6. seq, Gbpril7. seq, Gbpril8. seq, Gbpril9. seq, Gbpril10. seq, Gbpril11. seq, Gbpril12. seq, Gbpril13. seq；啮齿类：Gbrod1. seq, Gbrod2. seq；植物：Gbpln1. seq, Gbpln2. seq, Gbpln3. seq, Gbpln4. seq。下载的文件共 4 GB。

由 GenBank 提取基因序列数据通常采用：方法 1，搜索 DEFINITION 行中有 exon 的序列；方法 2，找 FEATURE 中有“CDS … join”或“CDS … complement (join”。

根据国际上构建二三级数据库的经验，方法 2 提取的数据更多，因为有些序列虽然包含 exon 但没有“CDS …join”或“CDS …complement (join”的注释。事实上，即使采用方法 2 仍会遗漏一些数据，因为 GenBank 建立初期，入库的序列没有“CDS join”的特征栏，不过这类序列的数据少的几乎可以忽略不计。另外，该方法只通过查找 CDS 来获得序列，虽然有些基因位于 CDS 之外 (如 3′UTR，5′UTR)，但是由于它们出现在 GenBank entry 中的格式没有统一标准，暂时无法考虑。

用 Perl 解析下载的真核生物 GBFF 文件，查找 *. seq 文件特性表中包含的内容：

CDS　　join

CDS　　complement (join

如果查到此内容，还不能将此序列写入数据库。提取一个合格的序列记录还要符合以下条件：①记录中有关于该基因序列外显子的详细位置信息；②每个 CDS 特征栏都包括该编码区所编码的蛋白序列；③通过计算外显子碱基数得出的氨基酸数不能跟实际的有出入。

满足所有 3 个条件的才可以进行下一步处理，即提取相关特征信息，如内含子序列、内含子位置、内含子相位、外显子长度、蛋白长度、蛋白序列等。最后所有这些信息才能作为一个完整的记录 (entry) 添加到数据库中。

1.1.2 *记录信息的获取* 外显子的获得：将 CDS join 特征栏内 DNA 片段 (有些在 GenBank 一个 entry 内，有些出现别的 entry) 组装成一个基因序列。外显子用大写字母表示。

内含子的获得：通过读取位于同一个 entry 并且在同一条 DNA 链上内含子两侧的外显子来抽提其间的序列作为内含子。当然有些时候内含子两侧的外显子不在同一个 entry，或在同一个 entry 的不同链上，或来源于 mRNA 序列，这时则不能确定内含子序列。这种情况下，程序仍然可以试图去读外显子相邻的序列来推测内含子的剪切位点。如果内含子序列本身不能确定的话，则用“..”表示。内含子用小写字母表示。

内含子大小：如果内含子大小未知，将内含子的大小写为 u；如果一个序列的任何一个内含子长度未知，内含子总长度则为 -1。

蛋白序列：从 CDS 表中的 translation //后的序列提取，当然同时要参照外显子的长度。如果某个残基所代表的密码子被内含子分开或紧挨着内含子，将该残基用小写字母表示。

内含子位置的获得：根据重新装配的基因 (外显子和内含子的总和) 上相邻的外显子来确定。例如，特性表中有这么一段：

CDS join (1…100, 200…300)

这样程序就可以推断出这里的内含子位置是从 101…199。

内含子相位的获得：内含子相位 (intron phase) 指的是内含子在三联体密码子中的位置，相位 0、1、2 分别代表内含子在密码子第一个碱基之前、第一个碱基之后、第二个碱基之后。用内含子 5′端外显子的长度除以 3 所得余数即可表示内含子相位。

内含子剪切位点的获得：本数据库提供 4 个字母组成的字符串代表内含子的剪切位点。它包括两个外显子的之间的内含子 5′端和 3′端的两个碱基。有时，GBFF 记录本身只有 mRNA 序列，或者某一个内含子没有记录在 GenBank 中，就无法确定内含子的开始或结束部分。这种情况下，用“NN”代替其真实的剪切位点。还有一种情况就是相邻的外显子距离小于 4 个碱基，这是可能是人为的笔误造成的。用“EEEE”表示该内含子的剪切位点。

1.2 真核生物内含子数据库结构

1.2.1 *EID 结构* 真核生物内含子数据库 EID 的设计包括构建一个主数据库和多个子数据库。真核生物内含子数据库主数据库与附属子数据库的关系见图 1。

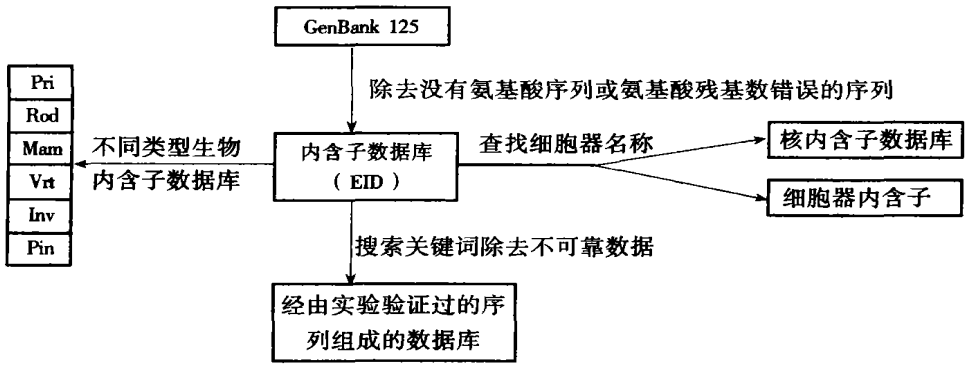

图 1 真核生物内含子数据库 (EID) 主数据库与各子数据库关系示意

Fig 1 Relationship between the main database and sub-databases of Eucaryote Intron (EID)

主数据库包括 3 个符合 FASTA 序列格式的文本文件库：dEID, pEID, hEID。文本文件库 dEID pEID hEID 分别代表：DNA-EID、Protein-EID、Header-EID, 其相应的 3 个文本数据库的结构见表 1。对于每一个序列的完整纪录 (entry), 数据库有相应的 DNA、蛋白序列和从 GenBank 提取出的头文件。此数据库采用文本数据库而非关系数据库, 主要是基于以下 3 点考虑:

表 1 真核生物内含子数据库(EID)主数据库结构

Tab 1 Structure of main database of Eucaryote Intron

dEID 字段	pEID 字段	hEID 字段
该记录简介	该记录简介	该记录简介
内含子相位	内含子相位	内含子相位
内含子大小	内含子大小	内含子大小
内含子总长度	内含子总长度	内含子总长度
外显子大小	外显子大小	外显子大小
外显子总长度	外显子总长度	外显子总长度
剪接位点模式	剪接位点模式	剪接位点模式
核甘酸序列	蛋白序列	GenBank Flatfile 头信息

(1) 数据库采用了符合 FASTA 格式文本文件, 它的应用性强, 在研究真核生物内含子过程中, 例如进化和起源研究中涉及许多分析软件的应用, 每种软件都有一定的文件输入和输出格式。例如: BLAST 的输入文件格式为 FASTA, 构建进化树软件 PHYLIP 的输出文件格式为 PHY, CLUSTALW 多序列比对软件的输出文件格式为 ALN, GDE(遗传数学应用环境系统) 的输出格式为 GDE 等, 它们都是文本文件, 相互转换非常方便。

(2) Perl 本身有较其它语言不能比拟的文本处理能力, 即使不用关系数据库, 用 PERL 写的查询速度也相当之快。

(3) 真核生物内含子数据库主要是用来研究某一个特有或一类基因家族的内含子, 用此数据库查询结果的信息量有限, 在浏览器端用此数据库来查询的用户数有限, 因而使用 PERL 写的 CGI 运行相对还是比较快(表 1)。

1.2.2 建库步骤

第 1 步: 新建 sequence.list 文件, 这个文件的内容是以上各 GenBank flat file 的路径, 每行只有一个文件路径, 如:

/EID/GenBank/gbinv1.seq
/EID/GenBank/gbinv2.seq
/EID/GenBank/gbmam.seq
/EID/GenBank/gbpln1.seq
/EID/GenBank/gbpri1.seq
/EID/GenBank/gbrod1.seq
/EID/GenBank/gbvrt.seq

第 2 步: 在 LINUX SHELL 下用 PERL 运行一个 parse.pl 程序, 将在 Feature 栏中含有

“CDS join
CDS complete men (join”

的序列提取出来。具体过程是: 首先读 seqfiles.list 文件, 将每个文件的每一行放到一个数组中去, 再根据 PERL 规则表达式寻找匹配, 其表达语句为:

```
$line= ~ /CDS join \ (| CDS comletement \ (join \ (/);
```

只要有匹配, 就将执行 scan-until-base 函数将从 BASE 到结束所有文本提取。

函数代码如下:

```
sub scan-until-base {
    while ( $line = <IN> ) {
      if ( $line= ~ /BASE/) {

    last;
      }
        else {
          $entry = $line;
      }
    }
}
```

运行完 parse.pl 会生成一个 first.EID 文本文件。

第 3 步：check_data.pl 程序，用来读取 first.EID，检查数据的合格性。除去没有氨基酸的序列，除去核酸碱基书不是氨基酸残基数 3 倍的序列。

除去没有氨基酸的序列通过如下代码实现：

```
if ( ( $line ! ~ /^\s * $/) &&( $line ! ~ /^
 $codon-start= 0;
 $translation="";
 $i--;
return 1;
}
/)) {   # # 没有包括 translation 的记录.
 # # 给起始密码子赋予一个虚值
 # # 给氨基酸序列赋予一个虚值
```

除去核酸碱基数不是氨基酸残基数 3 倍的序列。即：先计算外显子的总长度的，注意遇见 CDS complement 时总长度要减去单个外显子长度；再读氨基酸的残基数，检查外显子的总长度除以 3 是否等于氨基酸的残基数。此过程通过如下代码实现：

```
@protein = ("", split (//, $translation));
for ( $j= 0; $j< $ #introns; $j += 2) {
 $exonsize = $introns [ $j + 1] - $introns [ $j] +
1;
push (@exonsizes, $exonsize);
 $total-exon-size += $exonsize;
}
if ( $cdsline= ~ /CDScomplement \ (join/) {
@exonsizes = reverse ( @exonsizes);
}
 $exonsizes [ 0] -= $codon-start;
 $total-exon-size -= $codon-start;
 $diff = $total-exon-size /3 - $ #protein;
if ( $diff< -1 || $diff> 1) {
        return "";
}
```

运行 check-data.pl 成功后，会生成文件 second.EID。

第 4 步：makeEID.pl 程序，将每一个记录信息解析、计算和获取，并写入数据库。这是构建数据库的核心程序。程序首先读取 second.EID，获得每个序列的 reference 信息。然后分别将外显子序列、大小、位置、蛋白序列和内含子序列位置、相位、剪切位点等信息一一获得。

1.2.3 *数据库的文件格式说明* 主数据库中 dEID，pEID，hEID 的文件格式在 GenBank 文件格式的基础进行了修改。例如 dEID 的记录，第 1 行以“>”开始，这是 FASTA 文件的格式要求。接着是系统在构建数据库时给每一个记录赋予的 EID 索引号。再跟着是 Protein ID（蛋白质检索号和该记录简介。第 2 行是：内含子（相位，位置）；外显子（个数，大小，总长度）；{剪接位点模式：模式 1，模式 2 …}。最后一部分是序列本身，大写为外显子，小写为内含子。

1.2.4 *相关子数据库构建* 在 EID 中，一部分序列是通过实验验证过的，另一部分是通过一些基因预测软件得到的。而如果要使用本数据库的数据做细致的内含子研究，显然需要得到的序列只是已经被实验验证过的。

由于很多基因预测软件对内含子——外显子边界预测存在误差，GenBank 中很多基因数据是不可靠的。有 2 种过滤方法可以用来创建了一个含有最少未经实验验证过的内含子序列子数据库。

第 1 种方法是通过比较基因组序列和 mRNA 序列来过滤掉尚未被实验证明的序列。具体方法是解析那些 *.seq 文件，从中提取出 Locus 中包括 mRNA 的 entry，写入一个文件。用 BLAST2 将本数据库中基因序列的编码部分与该 mRNA 文件中的序列比对，如果有 95%的相似性的话，则说明这个基因的结构是通过正确可靠的方法而确定的。然后将这些 95%的数据写入数据库。

虽然应用这种过滤方法可以取得很纯的数据，但由于 GenBank 中的 mRNA 数据相对不多，而且两个数据库之间的比对非常耗内存，普通的计算机达不到要求，通常应用了第 2 种方法。

第 2 种方法是通过解析头分件查询关键词去除计算机软件预测内含子位置的序列。过滤过程包括查找 2 组关键词：第 1 次过滤要查找的关键词是查找预测软件相关的名词 Del-Predict1.pl，有 Evidence= experiment, Not-experiment, non-expe, riment, predict, FGENEH, grail, grail2, GeneID, hexexon, netplantgene, GeneView, genefinder, genscan, GeneMark, glimmer, netgene, GeneHacker, genquest, genequiz, GeneParser。第 2 次过滤要查找的关键词是基因测序用到的载体，有 BAC, cosmid, chromosome, PAC, htg。

首先看有没有 experiment，如果有，即使后面能查到其它关键词，也算是实验验证过的数据。如果没有，再通过 2 种方法查找不同类的关键词，如果查到，则要过滤掉。

注意：如果 GenBank 将来出现一些新的基因预测软件、载体名称和其它关键词，可在 filter1.pl 和 filter2.pl 稍加改动。

第 2 种方法好处是得到数量相当的序列，缺点是数据还是存在一定污染。

先后运行 filter1.pl 和 filter2.pl 后，就可以自动生成以上 3 个子数据库文件，即经过实验验证过的内含子序列：125.exp-keyw.dSID，125.exp-keyw.pSID，125.exp-keyw1.hSID。

另外，通过运行 extract-oraganell.pl 程序，查找线

粒体和叶绿体等细胞器, 数据库可生成了两类子数据库: 核内的内含子序列、细胞器内的内含子序列。

将 seqfiles. list 中的文件路径定制为各类生物的 flat file 所在位置, 然后分别执行 first_ parse. pl、check_translations. pl、finalmake. pl, 这样就生成了以生物类别区分的子数据库: 灵长类、啮齿类、其它哺乳动物、其它脊椎动物、无脊椎动物、植物真菌藻类。

2 EID 数据库的查询

2.1 WWW 查询

可通过 EID 索引号、GenBank 的 Accession number 和 locus 以及蛋白号来得到想要的序列信息。另外, 可选择子数据库进行查询。子数据库包括: 核内的内含子序列、细胞器、经过实验验证过的内含子序列、和各类不同生物的内含子序列。

查询方法跟一般关系数据库用 SQL 语句有很大不同, 采用 PERL 特有的哈希表临时存储相关的字段并查询。程序先查询 hEID 文件, 根据找到的匹配再查询 dEID 和 pEID 文件, 返回结果为文本, 查询程序的关键部分代码如下所示:

```
while( $entry=< HEADERS> ){
( $entry= ~ /$regular_ exp/) || next;
( $code) = ( $entry = ~ /> \s *( \d+_ \S+) /);
print OUTH $entry;
$all_ codes{ $code} =1;
}
while( $entry=< PROTS> ){
( $code) = ( $entry = ~ /> \s *( \d+_ \S+) /);
if( exists( $all_ codes{ $code}) ){
print OUTP $entry;
}
}
while( $entry=< DNAS> ){
( $code) = ( $entry= ~ /> \s *( \d+_ \S+) /);
if( exists( $all_ codes{ $code}) ){
print OUTD $entry;
}
}
```

2.2 利用 Perl 的正则表达式来查询

已经设计了一个功能更强大的查询工具 RE-search. pl, 即: 利用 Perl 自身的正则表达式。用户只需在 *nix 下运行命令"RE_ search. pl 数据库文件名前缀正则表达式"即可。

2.3 EID 数据库其它分析工具

EID 数据库设计提供以下分析工具可以用于相关的研究和分析。

extract_ species. pl: 可以从内含子进化数据库中提取出按物种定制的序列(包括物种名、dSID、pSID、hSID)。如运行: human extract_ species. pl 。

get_ exons. pl: 可以将任何一个 GenBank flat file 数据库中的外显子提取出来, 生成一个只含有外显子的数据库。

get_ introns. pl: 可以将任何一个 GenBank flat file 数据库中的内含子提取出来, 生成一个只含有内含子的数据库。

filter_ phase. pl: 可以从内含子进化数据库中提取出按内含子位相定制的序列。

filter_ splice. pl: 可以从内含子进化数据库中提取出按内含子剪位点切定制的序列。

2.4 BLAST 的 WWW 服务器

BLAST 是生物信息学研究中的十分流行的一个常用方法, 主要用于两个序列之间的比对。该软件是免费软件, 可以直接从 ftp. ncbi. nih. gov/blast/server/下载 wwwblast. linux. tar. gz 到 HTTP 服务器目录下, 然后将该文件解压安装。BLAST 已经被成功安装在 EID 数据库的服务栏目中, 并可以提供以下服务 (表 2)。

表 2 BLAST 服务和功能列表

Tab. 2 BLAST service list

BLAST 服务名称	功能
Regular BLAST	NCBI 标准 BLAST
PSI/PHI BLAST	反复精练数据库序列的 BLAST, 结果更精确
Mega BLAST	用于较长的查询序列的 BLAST
RPS BLAST	用来搜索保守结构域的 BLAST
BLAST 2 sequences	两个序列之间比对的 BLAST

以上 BLAST 5 种不同 BLAST 服务各有 2 种版本。一种无客户—服务器支持。另一种为客户—服务器版本。客户—服务器版本可以使用 NCBI 的 Entrez 的查询和分类学报告。

一次数据库搜索包括 4 种基本元素: BLAST 程序的名称, 数据库名称, 查询序列和大量的合适的参数, 很显然, 当以上元素发生变化时, 搜索的细节就会随之改变。

3 问题与讨论

3.1 存在的问题

EID 数据库存在一定的冗余度, 主要是由于许多基因在 GenBank 里就存在一个以上的拷贝。这主要是来自不同物种的直系同源基因或来自一基因家族和同一物种的旁系同源基因; 选择性剪切异构体; 同一基因被不同实验室测序并报道。

目前 EID 数据库只包括了真核生物的剪接体内

含子和少量细胞器内含子（叶绿体和线粒体内含子），并没有包含Ⅰ类，Ⅱ类和Ⅲ类内含子以及原细菌内含子和核tRNA前体内含子。即使是真核生物的序列，也没有3′UTR和5′UTR上序列。

关于内含子的信息收集和注释方面，除了有内含子位置、相位、剪切位点的信息外，EID没有提供与内含子相关的重复序列信息，内含子内的基因或RNA基因信息等。

另外，EID数据库中的每一个记录只包括了外显子和内含子序列，并不能代表一个完整的基因，因为缺少上游和下游的一些调控序列。

本数据库在http://gpgc.zsu.edu.cn/网站可以提供本地用户服务。

3.2 讨 论

进一步的工作首先需要着手去除数据库中的冗余度。计划采取全体序列两两比对的方法，剔除相似性达99%的序列。

收集和整理数据构建Ⅰ类、Ⅱ类、Ⅲ类和其它类内含子数据库，并将致力于它们与剪接体内含子进化关系和相互之间的关系的研究。

在内含子信息注释方面，将尽可能把与内含子相关的重复序列信息，内含子内的基因或RNA基因，保守区域等更多的信息挖掘出来。

将每一个记录跟国际上公共数据库GenBank连接起来，这样用户可以研究与序列相关的内容。

在内含子进化数据库和相关工具基础之上，开发更多的工具，如：内含子比对、内含子中的重复序列识别等，以建立一个完整的内含子进化、起源与功能研究的生物信息学平台。

参考文献：

[1] BERGET S M. An intron enhancer recognized by splicing factors activates polyadenylation [J]. Genes Dev, 1996, 10: 208－219.

[2] TAYLOR H. A homeobox gene regulatory element is composed of distinct conserved sub-elements that bind regulatory proteins [J]. J Soci Gyne Inve, 1996, 3: 110A.

[3] TEE M K. A promoter within intron 35 of the human C4A gene initiates adrenal-specific transcription of a 1 kb RNA: location of a cryptic CYP21 promoter element? [J]. Hum Mol Genet 1995, 4: 2109－2116.

[4] LEE R C. The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14 [J]. Cell, 1993, 75 (5): 843－854.

[5] HERBER T. A RNA editing, intron and evolution [J]. TIG, 1996, 12 (1): 6－9.

[6] HIGUCHI M. Intron-exon structure determines position and efficiency [J]. Cell, 1993, 75 (7): 1361－1370.

[7] ZUCKERKANDL E. Junk DNA and sectorial gene repression [J]. Gene, 1997, 205 (1－2): 323－343.

[8] SMITH T M. Complete genomic sequence and analysis of 117 kb of human DNA containing the gene BRCA1 [J]. Genome Res, 1996, 6: 1029－1049.

[9] KISS T. Exonucleolytic processing of small nucleolar RNAs from pre-mRNA introns [J]. Genes Dev, 1995, 9: 1411－1424.

[10] MILOL D. The oligodendrocyte-myelin glycoprotein of mouse: primary structure and gene structure [J]. Genomics, 1993, 17: 604－610.

[11] NADAL G B. Advances in Enzume Regulation, 1990, 31: 261－286.

[12] BREITBART R E. Intricate combinatorial patterns of exon splicing generated multiple regulated troponin T isoforms from a single gene [J]. Cell, 1985, 41: 67－82.

[13] HOWE K J. Intron self-complementarity enforces exon inclusion in a yeast pre-mRNA [J]. Proc Nat Acad Sci, 1997, 94 (23): 12467－12472.

[14] KURODA M. Whole genome sequencing of meticillin-resistant *Staphylococcus aureus* [J]. Lancet, 2001, 357: 1225－1240.

[15] KAN Z. UTR reconstruction and analysis using genomically aligned EST sequences [J]. ISMB, 2000, 8: 218－227.

[16] BROOKES A J. HGBASE: a unified human SNP database [J]. Trends Genet, 2001, 17 (4): 229.

[17] GERHARD M. Comparative genomics and evolution of genes encoding bacterial (p) ppGpp synthetases/hydrolases [J]. J Mol Micr Biot, 2001, 3 (4): 585－600.

Development of Eucaryote Intron Dtabase (EID)

HE Miao, LI Ji-dong, ZHANG Shang-hong

(School of Life Sciences, Sun Yat-sen University, Guangzhou 510275, China)

Abstract: Base on downloading the data of sequences from GenBank, a database of Eucaryote Intron was developed, which includes a main base and several sub-databases. The sub-databases were made up of karyon intron database, organelle intron database, and the intron databases of plants, invertebrates, primates, rodents, and other vertebrate. The methods of data fetching and filtering from GenBank were studied. The file formats of DNA-EID, Protein-EID, Header-EID and the structure of the main database had been introduced. Some of the skills of EID development process were presented. Also the querying methods and some of the analysis tools had been provided. The goal of this project is developing a bioinformatics platform for Eucaryote Intron function, evolution and origination study.

Key words: eucaryote; intron; database