线粒体基因组数据的分析方法和软件*

李雪娟 杨 婧 王俊红 任倩俐 李 霞 黄 原**

(陕西师范大学生命科学学院 西安 710062)

摘 要 线粒体基因组的研究已经普及,其正确的拼接和注释是所有后续研究的基础。本文以 Staden Package 软件为主介绍了拼接和注释的线粒体基因组的方法,同时介绍了其他常用的拼接软件 ContigExpress、DNAMAN、DNASTAR、BioEdit 和 Sequencher,以及利用不同软件(包括 DOGMA、MOSAS、MITOS、GOBASE、OGRe、MitoZoa、tRNAscan-SE、ARWEN、BLAST 和 MiTFi 等)对线粒体基因组中的蛋白质编码基因、rRNA、tRNA 和 A + T 富集区进行注释的方法,最后介绍了利用 MEGA5 软件分析线粒体基因组的组成、Sequin 软件提交序列和线粒体基因组数据绘图工具(CG view、MTviz 和 OGDRAW)。

关键词 线粒体基因组, 拼接, 注释

Methods and software tools for mitochondrial genome assembly and annotation

LI Xue-Juan YANG Jing WANG Jun-Hong REN Qian-Li LI Xia HUANG Yuan**

(School of Life Sciences, Shaanxi Normal University, Xi'an 710062, China)

Abstract With the increasing popularity of mitochondrial genome studies, the correct assembly and annotation of genomes are the basis of all subsequent research into a species. Here we describe the protocols using Staden Package software to assemble and annotate the mitochondrial genome, along with other commonly used software, such as ContigExpress, DNAMAN, DNASTAR, BioEdit and Sequencher. In addition, methods for the use of different software packages (including DOGMA、MOSAS、MITOS、GOBASE、OGRe、MitoZoa、tRNAscan-SE、ARWEN、BLAST and MiTFi) to annotate mitochondrial genomic protein-coding genes, rRNA, tRNA and the A + T region are briefly introduced. Finally, application of MEGA5 software to analyze the composition of mitochondrial genomes, Sequin software to submit sequences to GenBank, and mitochondrial genome data visualization tools (CG view、MTviz and OGDRAW) are also briefly introduced.

Key words mitochondrial genome, assembly, annotation

1 引言

线粒体基因组数据广泛应用于系统与进化生物学、群体遗传学和保护生物学等许多生物学研究领域。随着测序技术的快速发展和测序费用的下降,大量的线粒体基因组序列被很快测出,拼接和注释这些线粒体基因组是所有下游系统分析的先决条件。本文综述了目前可以利用的线粒体基因组拼接、注释、提交和绘图方法和软件。

线粒体基因组分析工具基本上可以分为本地和在线服务器二种,许多软件都是只能完成分析流程中的部分工作。Staden Package (Bonfield *et al.*,1995)是可以安装在本地计算机上进行拼接和注释的测序项目管理软件包,主要由 Pregap4、Trev、Gap4 和 Spin 等模块组成,可以进行序列拼接、突变检测、序列注释和对序列峰图及读序文件进行操作等。其中,Pregap4 是 Gap4 的前处理,可以处理原始的峰图文件,对序列进行载体和污染检查,同时也可以进行 Gap4 组装。经 Pregap4 处理所得到的结果,可以通过 Gap4 来进行查看和编

* 资助项目:国家自然科学基金(31172076,30970346)。

**通讯作者,E-mail:yuanh@snnu.edu.cn

收稿日期:2012-12-21,接受日期:2012-12-26

辑。组装后的序列以*.seq 格式输出用于在 Spin 中线粒体序列的注释。本文主要以该软件为主介绍对线粒体基因组序列进行拼接和注释的方法,同时介绍了线粒体基因组常用的其他拼接软件 ContigExpress、DNAMAN、DNASTAR、BioEdit 和 Sequencher,以及利用不同软件对线粒体基因组中的蛋白质编码基因、rRNA、tRNA 和 A + T 富集区进行注释的方法,最后介绍了序列提交软件 Sequin 和线粒体基因组数据绘图工具。

2　线粒体基因组序列的拼接

序列拼接是将测序生成的短读序片段通过重叠部分连接形成较长的片段,这样的较长片段称为叠连群(contig)。DNA 测序数据的固有特点(测序有误差、不完全覆盖性、序列所在链不确定)重复序列的干扰是解决实际序列拼接问题的难点所在。

采用传统 Sanger 法测定线粒体基因组通常需要至少 28 对以上的引物的双向测序反应,如果引物重叠区加长的话甚至需要更多对引物,这样双向测序可以获得比较高的覆盖度和准确性。传统测序的拼接软件主要有 ContigExpress、DNAMAN、DNASTAR、BioEdit、Sequencher 和 Staden Package 等。新一代测序技术中得到的读序片段长度短、数量巨大、覆盖度高,针对高通量测序开发了大量拼接软件(如 SOAPdenovo)等。

测序获得的序列首先进行同源性搜索确定其为线粒体序列,删除测序效果不好的测序文件,准备拼接。ContigExpress 作为 Vector NTI 的组件之一,是一款非常实用的序列拼接软件。它将每个测序片段视为一个叠连群,当输入多个叠连群后,软件会自动寻找其中的公共序列,然后以图形方式呈现出拼接结果。DNASTAR 软件包的 SeqMan II 可进行多序列拼接,最多支持 64 000 条序列的同时拼接,而且在拼接时可以修整质量差的序列并清除污染数据,还提供完善的编辑和输出功能。Sequencher 是序列拼接的行业标准软件,它以快速拼接叠连群、用户友好型的编辑工具以及精湛的技术支持等特点而众所周知。

Staden Package 软件拼接的具体流程为:

(1)打开 Pregap4,将全部测序片断(测序源文件为*.abi 格式)加载到 Pregap4 中。

(2)设置参数,在 Configure Modules 模块中选择各种参数,修改参数时,只需选择相应的选项,再据提示设置一定的参数,生成批处理文件*.0.aux。

(3)查看,在 Gap4 中打开文件*.0.aux,文件菜单含数据库打开和拷贝功能及保守序列文件产生的例行程序。一旦一个数据库被载入 Gap4,将以图形方式显示出 Contig Selector。当进行叠连群的比较分析时,Contig Selector 自动转变为 Contig Comparator,在这个对话框中选择所需的组装片段,依次察看相应叠连群的峰图,对峰图进行编辑。

(4)序列拼接结束后,还要对序列两头进行连接,看是否可以组成一个环状的 DNA 序列。全线粒体基因组序列拼接时,序列会在覆盖度最小的位置断开,因此拼接时输出的一致性序列的第 1 个碱基不是全线粒体基因组序列的第 1 个碱基。因为完整的线粒体基因组是一个环形结构,而断裂处的两条片断在连接时有一部分是重叠的,所以在输出的一致性全序列中,最前面和最后面一定数量的碱基是重复的,需要删除。注释线粒体基因组时确定起始碱基的过程称为调零,昆虫线粒体基因组以 tRNA-Ile 的第一个碱基设置为 1,在此位置之前的序列需要拼接在 3′末端。序列拼接完成后,要将输出的一致性序列输入到 ClustalX 等软件中计算序列长度,并记录下来。

(5)修改拼接版本。通常情况下拼接不是一次就能够完成的,在以后的序列注释过程中会发现一些错误,需要返回来仔细查看测序峰图,修改读序,或补充测序,重新拼接,产生一个新的拼接版本,直至所有注释顺利完成,才可以确定最终的拼接结果。

3　线粒体基因组序列的注释

线粒体基因组注释是确定已知的在绝大多数动物中存在的 13 个蛋白质编码基因、22 个 tRNA 基因、2 个 rRNA 基因和 1 个 D-loop 区的位置和序列,在有些物种中也存在个别 tRNA 基因和 D-loop 区的重复。线粒体基因组注释的标准源是 NCBI 的参考序列数据库(RefSeq),该数据库中基因组序列和注释结果是经过专家核对过的,可以为各种动物线粒体基因组的注释提供了很好的参考。注释方法是通过与其近缘物种的线粒体基因组比较和分析来定位蛋白编码基因、tRNA 基因、rRNA

基因和 D-loop 区。进行注释时需要建一个序列注释记录文件,记录每一次翻译过程(包括序列起始位置、碱基长度、蛋白质序列的起始密码子、tRNA 的反密码子以及其起始位置、基因位于 N 或 J 链),以便后续的提交序列、核对及查找。

线粒体基因组注释的工具很多,包括在线注释的网络服务器和本地计算机安装的注释软件。DOGMA (http://dogma.ccbb.utexas.edu/)、MOSAS(http://mosas.byu.edu)和 MITOS(http://mitos.bioinf.uni-leipzig.de/)是近年来开发的 3 个线粒体基因组注释网络服务器。GOBASE(O'Brien *et al.*,2009)试图改善现有的基因组注释,但仅仅集中于 NCBI 条目中;OGRe(Jameson *et al.*,2003)只存储 NCBI 的基因组注释和基因次序信息,并纳入了一些手动改善了的注释。MitoZoa(Lupi *et al.*,2010)使用一系列规则(如 tRNAscan-SE、ARWEN 和 BLAST)和专业知识,实现了系统的半自动化错误检查。此外,Jex 等(2010)开发了一种线粒体基因组的高通量测序、拼接和注释通道,通过内建的自动注释通道进行注释,注释序列数据被输入到 Sequin 程序中进行线粒体基因组结构的最终核实,且随后直接提交到 GenBank。

DOGMA 是一个基于 Web 的自动化注释细胞器基因组(叶绿体和线粒体)的软件包,并提供了一个图形化的用户界面用于查看和编辑注释结果。用户可以在 DOGMA 中输入 FASTA 格式的线粒体基因组完整核苷酸序列。该程序允许使用 BLAST 搜索自定义数据库以及动物线粒体 tRNAs 二级结构中保守性的碱基配对,从而识别和注释基因。最终的注释结果可以提取,并直接提交到 GenBank。DOGMA 在 BLAST 输出中构建了一系列基因,用图表为用户显示了基因列表。当一个基因被选中时,该基因核苷酸和氨基酸序列的详细视图和 BLAST 点击框位置会显示出来。由于使用与其他基因组中基因序列相似性的方法来定位假定的基因,所以,用户必须自己选择起始和终止密码子。每个 tRNA 和 rRNA 的起始和终止位置也必须验证。注释完成时,结果可能被恢复成 Sequin 格式,用于直接提交到 GenBank 中。对于进一步的分析而言,DOGMA 还允许用户提取基因组的子序列(包括间隔区,内含子,蛋白质编码基因的氨基酸序列等)。

MITOS 是后生动物线粒体基因组自动注释工具,该方法是基于蛋白质编码基因和非编码 RNAs 自动一致的重新注释(Bernt *et al.*,2012)。MITOS 允许系统误差筛选、基因名称和基因边界划定的标准化以及 tRNA 的反密码子标签,此外,MITOS 可提供框架来对基因定位的有效性进行评估。对于各种后续分析而言,MITOS 还可进行如基因组重排研究和系统发育分析等,并对现有数据的重新注释以及 De novo 注释有助于改进数据。

METAMiGA(Feijão *et al.*,2006),正式命名为 AMiGA,也可用于后生动物线粒体基因组的注释,且是一个更新速度较快的实用平台。

在这些注释工具中,DOGMA、MOSAS 和 MITOS 使用 BLASTX 对内部数据集进行搜索来确定蛋白质编码基因,后两者也采用 BLAST 来检测 rRNA 基因。来自于各种各样后生动物的数据库序列被用于 DOGMA 和 MITOS 搜索中;而 MOSAS 目前仅限于昆虫。DOGMA 和 MOSAS 使用 tRNAscan-SE 用于 tRNA 基因的鉴定,而 MITOS 使用 MiTFi 进行 tRNA 注释。MITOS 通道尝试来改善基因边界自动地预测。这 3 种工具都提供了图形和表格输出,它们允许输出 Sequin 格式的注释文件,以便将新的线粒体基因组提交到 GenBank 中。

利用 Standen Package 的注释工具 Spin 进行线粒体基因组注释的步骤详细介绍如下。

3.1 蛋白质编码基因的注释

蛋白质编码基因的注释内容主要包括 13 个基因起始与终止位点、起始密码子与终止密码子、编码链(N 或 J 链)以及基因长度等方面。可供使用的软件有 BLAST、CLUSTALX、SPIN、ORFFinder 以及 DOGMA 等。

使用 Standen Package 中 SPIN 程序注释蛋白质编码基因的具体操作步骤为:

(1)下载一条最相近物种的有详细注释信息的线粒体基因组全序列作为注释的参考序列。打开 Staden 中的 spin,在 file 主菜单中设置参考基因组和待注释基因组的所在路径。

(2)通过 file 菜单加载参考基因组和待注释基因组文件。

(3)设置坐标:在主菜单中 sequence 子菜单 Horizonal(参考基因组)和 Vertical(待注释基因组)中设置 X 轴和 Y 轴的两个序列。

(4)调零:在主菜单 Comparison 的子菜单 Local sequences 中进行参考基因组和待注释基因组全局比对。以 tRNA-Ile 的第一个碱基设置为 1,前面序列后移。

(5)在主菜单中 Translation 子菜单 Set genetic code 中设置遗传密码,如是无脊椎动物,则选择 Invertebrate mitochondrial。

(6)根据该基因在参考序列中对应的起始和终止位点确定其在待注释序列中的可能起始和终止位点。选择主菜单中 Translation 子菜单 Find open reading frames 的下拉菜单 Write protein as fasta file,将核苷酸序列翻译成氨基酸序列。若终止位点与比对时预测的可能终止位点一致或基本一致(实际终止位点与预测的终止位点相差不大,仅提前向前移动了少数几个位点,且测序效果很好),则该基因注释完成,将该基因的起始、终止位点及翻译所得的蛋白质序列等信息复制保存到注释模板中的相应表格中。若终止位点与比对时预测的可能终止位点很不一致(实际终止位点与预测的终止位点相差很大),则需对终止位点前后的碱基进行检查核对,看是否因测序误差产生了移码突变;若发现有测序效果不好的碱基,则需进行必要的手工校正,并将校正后的序列重新翻译,直到能够顺利翻译为止,手工校正后的序列文件要另外保存,给予一个新的版本编号,以防手工校正不成功而破坏前面已经完成的注释工作。若序列能够顺利翻译,但末尾不是标准的终止密码子,则需将翻译位点向后移动一段,看能否找到终止密码子;若仍找不到终止密码子,则需将翻译所得的氨基酸序列与参考物种的氨基酸序列比对,比对结果可分为两大类:① 如果序列末尾是 T 或 TA 且比对结果很好,则认为该基因的终止密码子是不完整的终止密码子,基因注释完成,② 如果序列末尾不是 T 或 TA、或比对结果很不好,则需要在 SPIN 中找到对应的位点进行检查与核对,再根据检查结果在拼接图中找到对应的位点,查看这些位点的测序结果并进行手工校正。如果在 Spin 比对后找到相似序列,但是在翻译过程中开放阅读框出现序列提早终止,搜索相似序列时有发现有 2 个或多个序列都可以找到相似程度较高,那么需要查找测序峰图,检查峰图效果好坏,有时需重新测序以矫正所测峰图。

(7)验证编码区及完整性,将注释获得的蛋白质序列在 NCBI 中的 Blast 进行相似性检索,查看是否找到相应蛋白质的相似序列。

DOGMA 也可用于注释蛋白质编码基因,这是基于数据库中与其他基因组相似序列的保守性。每个基因都在氨基酸序列数据库中使用 BLASTX 对其 6 个阅读框进行核苷酸序列校对,各种 BLAST 参数(如 E 值)可由用户自行设置。DOGMA 确定了蛋白质编码基因后,用户可为每个基因选择起始和终止密码子。对于含有内含子的基因而言,DOGMA 会基于 BLAST hit 的边界来确定内含子边界,而后由用户进行验证。

3.2　rRNA 基因的注释

rRNA 基因的注释内容主要包括 2 个 rRNA 基因位置、长度及二级结构验证等。可供使用的软件有 BLAST、CLUSTALX、DOGMA 和 Infernal 等。

利用 Standen package 中的 SPIN 程序注释 rRNA 基因的流程为:在 Comparison 主菜单的 Align sequences 子菜单中,与参考序列注释文件相比较,找到大致位置,还要考虑前后的基因终止和起始位置,最好是在画出其二级结构后,再确定 RNA 长度。

对于已注释好的 RNA 基因,可使用不同的方法检测该基因序列是否可信。例如,① 根据比对结果,查看有大量空格插入的位置(包括参考序列插入空格的位置)的测序峰图,若测序效果好,则初步认为序列可信,否则考虑进一步检测或进行手工校正。② 将该基因序列输入到 NCBI 中,用 Blast 命令进行比对,检查其同源性,由于一部分的 rRNA 基因是高度变异的,故 BLAST 参数(如间隔罚分或相同度)是需要不断优化的,若与近缘物种的相同基因的同源性很高,则认为该基因序列可信,注释完成。③ 用 rRNA 二级结构预测软件对其二级结构进行预测,若二级结构中的主要区域完整,则认为该基因序列可信,注释完成。

rRNA 二级结构通过 RNA Structure 以及与近缘物种的线体粒体 rRNA 比对进行预测。Vienna RNA Package(Hofacker *et al.*,1994)可预测和比较 RNA 二级结构,通过能量最小化来预测 RNA 二级结构。

3.3　tRNA 基因

线粒体基因组 tRNA 基因注释内容主要包括基因位置、数量、长度、二级结构与变异、反密码

子、分布链(N 或 J 链)及二级结构验证等。线粒体基因组通常编码 22 种 tRNA 基因,除丝氨酸和亮氨酸各有 2 种 tRNA 以外,其他氨基酸均只有一种相应的 tRNA。线粒体基因组 tRNA 基因的 2 个最大特点是:其位置多发生重排,而确定 tRNA 基因种类的主要依据是反密码子类型,所以 tRNA 的注释的核心是通过二级结构确定反密码子类型;缺乏典型三叶草结构的奇异二级结构的 tRNA 序列很难采用软件检测到,在计算分析和线粒体基因组注释中经常会被错过,需要通过手动注释。

线粒体基因组 tRNA 基因注释可供使用的软件有 tRNAscan-SE、CLUSTALX、ARWEN 和 MiTFi 等。

tRNAscan-SE(Lowe and Eddy,1997)是应用最广泛的 tRNA 基因注释工具,其特点是假阳性率很低,很好的结合了 tRNAscan 和 EufindtRNA 2 种算法的灵敏度,可以识别出绝大多数真实的 tRNA。使用 tRNAscan-SE 注释 tRNA 基因的步骤是:

(1)将已经调零的 fasta 格式的待注释序列输入到 tRNAscan-SE 服务器中(http://lowelab. ucsc. edu/tRNAscan-SE/)。

(2)设置以下选项:

检索模式(search mode) 选项设置一般采用默认(Default)。检索模式决定使用哪种概率模型进行搜索,每个模型都是基于从不同分类学类群中得到的 tRNA 序列训练的。不同的模型在灵敏度和速度上有所差别,在大多数情况下,默认的搜索模式速度快而且十分敏感。

来源(Source)选项设置为"Mito/Chloroplast";

格式(Format)选项使用默认项"Raw Sequence"或其他中的 FASTA

搜索结果可根据需要选择"Show results in this browser"或"Receive results by e-mail instead";

"Genetic Code for tRNA Isotype Prediction"选项根据待注释序列的物种所属的生物类群选用(昆虫选择"Invertebrate Mito")。

"Cove score cutoff"值默认为 20,具有典型三叶草结构的 tRNA 的值通常在 20 以上,默认报告 18 以上的,初学用户使用默认参数,熟练用户可以采用 13 可以获得 23 个左右的 tRNA。

其他参数都采用默认设置。

(3)点击"run tRNAscan-SE",预测结果会显示在页面上,用户可以查看 tRNA 基因的位置及二级结构等方面的信息。

(4)将运行结果复制并保存到注释文档中备查。点击每一个 tRNA 搜索结果下方的"View tRNA"按钮,打开该 tRNA 的二级结构图,将鼠标移至该图上,点击右键,使用"图片另存为(S)"选项保存图片(*.gif 文件),图片文件名可使用"[搜索结果顺序号]-[起始位点]-[终止位点]-tRNA-[tRNA 名称]"的格式,以便识别。

tRNA 注释过程中需注意的事项包括:①在默认的"Cove score cutoff"情况下,tRNAscan-SE 搜索的结果在 20 个左右。$tRNA^{Cys}$ 和 $tRNA^{Ser}$ 基因中的一个通常搜索不到,需要通过与近缘物种的比对进行手工寻找并绘制二级结构图;②搜索结果中显示出来的 tRNA 序列都是提交序列,所以在寻找与核对反链编码的 tRNA 基因序列中的反密码子时,要看对应位点的反向互补序列,绘制 tRNA 二级结构图时要将该段序列转换成反向互补序列再绘图。③在注释过程中如果对序列进行了手工插入或删除碱基的校正,那么要将校正后的序列输入到 tRNAscan 中重新搜索,以便纠正后面的 tRNA 基因的起始和终止位点。

MiTFi 可以通过调用 Infernal(Nawrocki *et al.*, 2009)来搜索含有所有 22 个 tRNA 的目标线粒体基因组,然后,采用步进式程序来评估和总结搜索结果,输出的是一个综合性的 tRNA 基因注释。由于线粒体遗传密码的可变性,且对应的反密码子和同功受体种类是含糊不清的,故 MiTFi 允许用户从 NCBI 遗传密码页面指定编码,或允许用户提供改进的编码。此外,MiTFi 提供了多种输出选项以方便对结果进行手动检查。对于动物线粒体基因组 tRNA 注释而言,MiTFi 在敏感性和准确率方面有了大幅的改善。

ARWEN (Laslett and Canbäck, 2008) 和 tRNAscan-SE 中均使用了相同的模型。然而,与 tRNAscan-SE 相比,ARWEN 首先只识别最保守的结构域和反密码子茎,随后评估可能的 D-茎和 T-茎结构以及搜索受体臂。ARWEN 在其敏感性增加的同时,也增加了错误率。

3.4 D-loop 或 A + T 富集区的注释

线粒体 D-loop 或 A + T 富集区的注释内容主要包括序列长度变化、保守基序和重复序列等,可供使用的软件有 CLUSTALX 和 SPIN 等。可根据

其两端的序列位置确定 D-loop 或 A + T 富集区位置,也可以近缘物种为参考序列,利用 Standen Package 中的 SPIN 程序来确定其位置。对于 D-loop 或 A + T 富集区中的重复序列而言,可利用 Tandem Repeats Finder(Benson,1999)在默认参数下从控制区全序列中筛选。

4　线粒体基因组的组成分析

对于线粒体基因组的组成分析,主要内容包括基因组组织特征分析、密码子特征分析、比较及进化分析和谱系基因组学分析等。所使用的软件主要有 MEGA5、DAMBE、BioEdit 和 DNASTAR 等。这些软件均可统计序列全长、碱基含量及百分比等信息。

在 MEGA5 软件中进行线粒体 DNA 序列组成及变异分析的主要流程是:

(1)导入数据,用 MEGA 进行数据分析时,输入的数据必须是"*.meg"格式文件,否则不能识别,所以在分析数据前要先将其它格式文件转换成"*.meg"格式文件,并根据研究类群选择相应的遗传密码子表。

(2)计算 DNA 序列碱基组成,在"View Sequence Data"窗口(即数据处理窗口)中点击"Satistics→Desplay Results in Text Editor",可将统计结果设置为在"Text File Editor and Format Convertor"窗口中显示(也可以根据需要将统计结果设置为以"Excel"形式或"Comma-delimited format"形式显示);点击"Satistics → Nucleotide Composition",软件将会在内置文本编辑器(built-in text editor)"Text File Editor and Format Convertor"窗口中显示碱基组成分析结果,保存文件备用(分析结果包括碱基总数,每种碱基的百分比)。

(3)计算密码子使用情况,点击"Satistics→Codon Usage",软件将会在"Text File Editor and Format Convertor"窗口中显示密码子使用分析结果,保存文件备用(分析结果包括碱基总数,每种碱基的百分比,各碱基在密码子第 1 位、第 2 位、第 3 位的使用频率)。

5　线粒体基因组数据提交

线粒体基因组数据提交的常用软件为 Sequin,它是 NCBI 为了方便参与测序工作的研究人员将序列提交到 GenBank 数据库而设计的一个客户端软件,用户可以不需要上网就能够在本地计算机上实现序列输入、格式的定制和修改,通过 Email 提交和更新序列数据。一般情况下,Sequin 所需的提交序列是 FASTA 格式的。但是在 Sequin 中进行群体学、系统发育学、突变等研究的多个序列可以通过 PHYLIP, NEXUS, MACAW 或 FASTA + GAP 等格式进行提交。Sequin 还可以对序列进行复杂的注释,并且内建某些的确认功能,从而保证所提交序列的质量可靠性。除了提供序列注册功能外,Sequin 程序也结合了 NCBI 的 ENTREZ 搜索引擎和 BLAST 搜索引擎,提供对 NCBI 的其它数据库的检索和同源性比较的程序。此外,通过开始界面上或提交界面上的 Misc 菜单即可进入其 ENTREZ 界面,从而可以方便地对 GenBank 中的 MEDLINE 文献数据库、核酸序列、蛋白质序列、蛋白质三维结构和基因组序列数据库进行检索。

Sequin 提交序列的流程为:

(1)主界面选择需要提交的数据库(GenBank、EMBL & DDBJ),点击 Start New Submission。

(2)选择提交的日期,并填写提交序列草稿的暂定题目。

(3)填写第一提交作者的个人信息,如:姓名、联系方式。

(4)填写其他提交作者的个人信息。

(5)填写提交作者的工作单位。

(6)选择所提交序列的类型:共有 7 个选项供选择(单序列、片断序列、群体学研究序列、系统发育学研究序列、突变研究序列、环境样本序列以及批量提交序列)。然后选择所提交序列数据的格式。

(7)填写所提交序列的物种名称。在"Location of Sequence"选项中选择所提交序列在生物体中的位置,如果是核基因编码的序列就选择 Genomic 选项。在"Genomic Code for Translation"选项中选择翻译的遗传秘密类型。

(8)填写核酸序列以及相关的描述信息。序列属于哪类分子,这个选项里有 genomic DNA, genomic RNA, mRNA(cDNA)等选项;所提交的序列是环状(circular)还是线状(linear)等内容。

(9)注释:在编辑窗口中的主菜单中"Annotate",可以对基因组各中特征序列进行注

释,其中,Genes and Named Region 选择 Gene, Coding Region and Transcrips 选择 CDS,进而对蛋白质基因进行注释;而 Structural RNA 选择 tRNA 和选择 rRNA,对 tRNA 和选择 rRNA 基因进行注释,包括 tRNA 基因所在位置、氨基酸、反密码子及其位置。

(10)验证与修改记录:如果错误或者不具体的地方可以进行修改和补充。具体的操作就是在需要修改的地方双击,就会弹出一个对话框,然后就可以按要求进行修改了。然后进行以下测试是否符合要求,点击"Done"按钮就是对这一记录进行合法性测试,如果成功,则可以存盘,如果不成功,则需要按要求进行修改,如果有些问题实在是难以修改,也可以提交到数据库去,他们会帮助完成修改的。

(11)提交:Sequin 所产生的文件是*.sqn 文件,可以通过 Email(gb-sub@ ncbi. nlm. nih. gov)的方式提交到 GenBank。如果一次需要提交 2 个或者更多的包含多条序列的*.sqn 文件,不要用一个邮件来发送,而是分别用不同的邮件来发送每一个*.sqn 文件。

6 线粒体结构绘图

线粒体基因组数据需要很多图形化显示,主要包括线粒体基因组结构图、rRNA 和 tRNA 二级结构图等。普通的矢量绘图软件(CorelDraw, Illustrator)可以绘制出满足发表需要的各种结构图。目前也开发出了专门用于线粒体基因组的在线绘图软件,如 CG view(http://stothard. afns. ualberta. ca/cgview_server/),MTviz(http://pacosy. informatik. uni-leipzig. de/mtviz/mtviz)和 OGDRAW(OrganellarGenomeDRAW, http://ogdraw. mpimp-golm. mpg. de/),这些软件可以根据用户提供的 GenBank 格式的数据绘制出高质量的基因组结构图。

7 小结与展望

线粒体基因组数据分析涉及很多生物信息学工具,随着越来越多的线粒体基因组序列测序工作的完成,线粒体基因组的拼接和注释成为了许多实验室的一项重要工作。然而,目前的分析工具虽然很多,但缺乏系统性,需要运行很多软件才能完成分析,还没有一个可以整合所有分析内容的综合软件。随着各个门类的物种线粒体基因组数据的快速积累,相信很快就可以开发出整合了拼接、注释、提交、绘图和分析线粒体基因组的软件包或在线服务器,能够更准确便捷地完成线粒体基因组的分析工作。

参考文献(References)

Benson G, 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*, 27(2): 573 – 580.

Bernt M, Donath A, Jühling F, Externbrink F, Florentz C, Fritzsch G, Pütz J, Middendorf M, Stadler PF, 2012. MITOS: Improved de novo metazoan mitochondrial genome annotation. *Mol. Phylogenet. Evol.*, doi: 10. 1016/j. ympev. 2012. 08. 023.

Bonfield JK, Smith KF, Staden R, 1995. A new DNA sequence assembly program. *Nucleic Acids Res.*, 23(24): 4992 – 4999.

Feijão PC, Neiva LS, deAzeredo-Espin AM, Lessinger AC, 2006. AMiGA: the arthropodan mitochondrial genomes accessible database. *Bioinformatics*, 22(7): 902 – 903.

Hofacker IL, Fontana W, Stadler PF, Bonhoeffer S, Tacker M, Schuster P, 1994. Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.*, 125: 167 – 188.

Jameson D, Gibson AP, Hudelot C, Higgs PG, 2003. OGRe: a relational database for comparative analysis of mitochondrial genomes. *Nucleic Acids Res.*, 31(1): 202 – 206.

Jex AR, Hall RS, Littlewood DT, Gasser RB, 2010. An integrated pipeline for next-generation sequencing and annotation of mitochondrial genomes. *Nucleic Acids Res.*, 38(2): 522 – 533.

Laslett D, Canbäck B, 2008. ARWEN: a program to detect tRNA genes in metazoan mitochondrial nucleotide sequences. *Bioinformatics*, 24(2): 172 – 175.

Lowe TM, Eddy SR, 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, 25(5): 955 – 964.

Lupi R, de Meo PD, Picardi E, D'Antonio M, Paoletti D, Castrignanò T, Pesole G, Gissi C, 2010. MitoZoa: a curated mitochondrial genome database of metazoans for comparative genomics studies. *Mitochondrion*, 10(2): 192 – 199.

Nawrocki EP, Kolbe DL, Eddy SR, 2009. Infernal 1.0: inference of RNA alignments. *Bioinformatics*, 25 (10): 1335 – 1337.

O'Brien EA, Zhang Y, Wang E, Marie V, Badejoko W, Lang BF, Burger G, 2009. GOBASE: an organelle genome database. *Nucleic Acids Res.*, 37: D946 – 950.