中国中药杂志
China Journal of Chinese Materia Medica

生物信息学在中药材分子鉴别中的应用

黄家乐[1], 邵鹏柱[1,2*]

(1. 植物化学与西部植物资源持续利用国家重点实验室 香港中文大学 中医中药研究所, 香港;
2. 香港中文大学 生命科学学院, 香港)

[摘要] 随着 DNA 技术鉴定中药材日渐普及,相关研究资料与日俱增,研究人员必需掌握生物信息学知识作有效序列信息搜集和分析。文章综述了生物信息学在 DNA 鉴定中药材的应用范围,包括查找物种分类条目、下载 DNA 序列数据、比对数据库序列、多重序列排序、及序列分析等。除逐一介绍外,也指出当中应注意事项。

[关键词] 中药材;鉴定;生物信息学

生物信息学(bioinformatics)是一门利用计算机和信息科学分析生物数据的学科。“Bioinformatics”一词最先由 Paulien Hogeweg 和 Ben Hesper 于 1978 年创造发表[1],原指探讨生物系统中信息传递过程的研究,但 1980 年后期演变成现今的定义。自 1950 年 DNA 的双螺旋结构发表后,生物学取得重大进展。20 世纪 90 年代,第一个细菌全基因组序列发表,其后有更多物种基因组被解读。为了处理和分享这些成果,科研人员建立了各种数据库,加入各种搜查、统计及分析等功能。生物信息学已是现今生物医学研究不可或缺的一部分,在中药材分子鉴定技术方面亦渐担当重要角色。生物信息学有 3 个主要功能:一是将实验所得数据整理储存,方便分析。《中国药典》2010 年版收载了约 700 个生物品种[2],当中超过 500 个品种的 DNA 序列已收载在美国国立生物技术信息中心(National Center for Biotechnology Information, NCBI)数据库[3]。DNA 序列数量庞大,如果没有数据库根本难以管理。二是透过计算机程序,将所得结果加以分析和比对,从而辨别受测样本的身份。理论上,每一个生物个体的 DNA 序列都是独一无二的,DNA 鉴定技术就是通过各种方法,去分析解读生物基因组数据。三是协助设计实验程序和选取合适鉴定工具。DNA 鉴定技术可分为三大类:DNA 指纹图谱技术;DNA 测序;及基于 PCR 的高通量核酸杂交技术(DNA 微阵列)。这些技术分辨出物种间的序列突变;插入突变;缺失突变;重复单元数目的突变,以及单核苷酸突变等差异。利用生物信息工具可分析物种间的各种差异,依此设计鉴定测试的方法和选择合适的材料。生物信息的数据库和软件日趋先进,加上互联网的普及,都令分析和储存数据变得更容易。《中国药典》2010 年版首次纳入 DNA 分子技术,用作中药材蕲蛇、乌梢蛇和贝母的品种鉴定,预计不久将来,DNA 鉴定中药材技术会日趋普及而成为常规检测方法,因此从事中药材鉴定和质控的人员须有基本的生物信息学知识。

1 生物信息学在中药材 DNA 鉴定的应用范围

生物信息学在中药材 DNA 鉴定的应用范畴主要包括:查找生物品种的分类条目和学术名称,下载物种的 DNA 序列作分析及设计实验,进行序列排序,及根据排序结果计算相似度或建立亲缘树,以鉴定待测品。本文从鉴定中药材的角度和需要,重点讨论常用的数据库、在线分析工具和分析软件。

2 生物品种分类信息

中药材主要分为动物、植物和矿物三大类,当中又以植物药居多。除了《中国药典》外,中国其他地方药用品种估计超过 1 万,具备生物分类学知识对鉴定中药材大有帮助。NCBI (http://www.ncbi.nlm.nih.gov/)是一个医学和生物学研究成果交换平台,由不同类型数据库和分析软件所组成。Taxonomy 数据库(http://www.ncbi.nlm.nih.gov/taxonomy/)是当中一员,其主要功能是管理生物品种名称和分类条目,让用户连接到 NCBI 等数据库内有关该物种的所有数据。当物种的 DNA 或蛋白序列收载到核酸(nucleotide)或蛋白(protein)数据库后,NCBI 便会新增该物品种的资料到 Taxonomy 数据库内。中药材鉴定经常比较正品和伪品的差异,透过 Taxonomy 数据库便可检视正品,亲缘种和伪品的所有数据。《中国植物志》是中国植物分类学的权威[4],记载了我国超过 3 万种植物的科学名称、形态特征、生态环境、地理分布和经济用途等,网上电子版网址为(http://frps.plantphoto.cn/)。《中国植物志》可让研究人员了解中国境内植物的品种数量,例如要开发鉴定人参 *Panax gisneng* 的方法,应先知道人参的近缘种,利用《中国植物志》查明中国境内 *Panax* 属内的品种,采集各品种的样本。The International Plant Names Index (http://www.ipni.org/)是一个植物品种分类的查询数据库[5],收载世界各地的植物资料。此

[稿件编号] 20111116009

[通信作者] *邵鹏柱,研究员,博士生导师,研究方向为中药材 DNA 鉴定技术开发与应用, Tel:(852)26961363, Fax:(852)26035646, E-mail:pcshaw@cuhk.edu.hk

外,由本组建立的全世界第一个针对中药材而建成的 DNA 条形码数据库(Medicinal Materials DNA Barcode Database)亦提供药材的基本数据和相片,对中药材的分子鉴定,很有参考价值[6]。

3 下载 DNA 序列信息

3.1 DNA 序列数据库的介绍 DNA 序列数据库是生物信息学的重要工具。中药材鉴定技术主要利用物种间 DNA 序列差异的现象来达至区分物种的目的。鉴定中药材时,经常要从 DNA 数据库中下载序列跟待测品比较或设计特异性 PCR 引物。NCBI 的 Nucleotide 数据库、欧洲生物信息学研究所(European Bioinformatics Institute, EBI)的 EMBL-DNA 数据库[7]和日本 DNA 数据库(DNA Databank of Japan, DDBJ)[8]为世界三大序列数据库,当中又以 NCBI 最为广泛采用。以下集中讨论从 NCBI 的 Nucleotide 数据库下载 DNA 序列的方法和要点。

3.2 从 Nucleotide 数据库下载序列 NCBI 的 Nucleotide 数据库(http://www.ncbi.nlm.nih.gov/nuccore/)接受各研究组所提供的序列数据。跟 NCBI 的 Taxonomy 数据库一样,Nucleotide 数据库和其他 NCBI 数据库互相连系,用户可从搜寻序列开始,追踪找出和该项序列有关的其他类型数据。最简单的下载方式,是在版面中输入序列或基因名称,例如输入 internal transcribed spacer(植物核糖体 DNA 内转录间隔区序列),数据库便会显示所有符合输入条件的记录。如果只需下载属于个别物种的序列(例如人参 *Panax ginseng* 的所有 internal transcribed spacer),可在查询语法中加入物种名称变成"(internal transcribed spacer) 和 panax ginseng [Organism]"。为方便用户,Nucleotide 数据库提供 Nucleotide Advanced Search 功能版面,无需强记语法便可自行组合搜寻条件。搜寻后,数据库会列出每项序列记录的摘要。版面左上方的"Display Settings"下拉选单中,用户可转换显示的格式、每页显示记录数量,以及排列次序。而右上方"Send To"下拉选单让用户导出数据,而且可选择多种格式,最常用的是 FASTA。另外,亦推荐 3 种 XML 格式。

3.3 FASTA 与 XML 档案的运用 FASTA 原本是一个计算机排序软件的名称,因这个软件得到广泛采用,其格式已成为业界标准。现时大部分序列分析软件都接受 FASTA 格式,包括 BLAST[9],Clustal W[10]等,所以一般序列分析会优先考虑。虽然使用 FASTA 较为简便,但所含信息量有限,如无法显示序列长度、提供者姓名和所属单位、文献出处等。为此,NCBI 采用 Abstract Syntax Notation 1 (ASN1)语言为每项数据记录,下载档中包含与该序列有关的所有讯息,让用户以计算机语言如 Pearl 等,按所需从 ASN1 原始檔中选取数据。近年 NCBI 提供可延伸标记式语言(eXtensible Markup Language, XML)档下载服务,为 ASN1 的另一选择。XML 属于标记式语言[11],档案内只含标记和信息,其最大优势是可作跨平台之间数据互换,利用扩展样式表单转换语言(Extensible stylesheet language transformations, XSLT)对 XML 的标记和信息进行转化,便可从大量信息中选取目标数据。随着互联网的普及,XML 已逐渐成为在线交换数据的标准。XML 内的标记和信息就像英文字母积木方块,而 XSLT 则是英文生字的串法。用户可编写不同 XSLT 文件,命令程序组合字母积木成为一个英文生字。如果要另一个英文生字,仅需编写新的 XSLT 档。因此,同一份的 XML 档案,在不同用户手上可各自转化成所需信息。现时 NCBI 提供 3 种 XML 导出格式,包括 XML,INSDSeq XML,TinySeq XML。XML 包含最多数据,但所占空间亦最多,其次是 INSDSeq XML,最小的是 TinSeq XML。TinSeq XML 因描述结构简单,只需 Microsoft Excel 内置的 XML 转化功能亦可将档内数据格式化显示。而其余两者的结构繁杂,必须自行建立 XSLT 将之转化。

4 DNA 序列排序

4.1 排序方式的种类 序列排序是一种比较序列的方法,序列相互比对,以最接近方式对齐,从而比较两者差异。如果超过 2 条序列,便是多序列排序(multiple sequence alignment)。排序可分为全局性排序(global alignment)和区域性排序(local alignment),前者会将整个序列比对,而后者只会对序列的局部进行比对。

4.2 区域性排序计算机程序 BLAST BLAST (Basic Local Alignment Search Tool,局基本区域排比搜寻工具),是 NCBI 提供的一个 DNA 和蛋白序列比较工具,功能是将检索序列(query)与数据库内的序列互相比较,找出与检索序列相似的配对序列(subject),并按两者相似程度由高至低排列显示。BLAST 所采用的搜寻方法属于区域性排序,检索序列会先被分割成小片段,然后跟数据库比对。BLAST 之所以采用区域性排序,原因是被检索序列可能经过重组,又或者包含一些蛋白结构区域(protein domain),这些特异性区域通常只占整个序列的一小部分。如果以全局性方法去寻找,数据库内某些拥有这种特异性区域的序列,往往因只有一小部分相似,而被排除出搜寻结果。NCBI 提供三大类别针对不同需要的 BLAST 程序。第一类是全基因组查询,第二类是基本序列查询,第三类是特别序列查询。中药材 DNA 鉴定技术主要利用第二类别中"nucleotide blast"功能进行序列比较。为加快搜寻速度,NCBI 将 nucleotide blast 里的序列数据库分成人类、鼠及其他物种 3 个选项,搜寻前必需选择合适的数据库,否则会出现假阴性结果。实际操作时先将检索序列贴在输入栏,如查询多条序列,只需将每条序列以 FASTA 形式输入,即可免去逐一查询的工夫。之后在数据库选项中点选"其他",然后按"BLAST"键。比对完成后,BLAST 会将结果分成 4 个部分显示。最顶部是查询的条件设置记录,例如检索序列长度和用作比对的数据库名称。第二部分是 1 个图表,图表上半部有 5 个不同颜色的条带,由左至右顺序是黑、蓝、绿、粉红和红。最左边的黑色代表检索序列和配对序列相似度为最低,右边的红色为最高,其余的蓝、绿和粉红色

介乎两者之间。紧接是 1 条代表检索序列的红色条带,左侧有"Query"一字,并有刻度表示序列碱基位置。图表下半部显示跟检索序列最相似的配对序列。当进行分析时应注意两点:第一,这个配对序列的条带是何种颜色,如果是红色,即代表这个配对序列和检索序列相似程度高;但如果配对序列是黑色,即使排在第 1 位亦不可信。第二,这个配对序列是全部还是局部跟检索序列相似,假设检索序列长度为 600 个碱基,而比对结果排第 1 位,属于物种 A 的配对序列整条都呈红色,而第 2 位的物种 B 的其中 2 段是红色,但这两段中间却是粉红色,即代表这个序列并非完全和查询序列相似,只有当中的 2 段相似度较高,而中间较低。BLAST 结果第三部分是一个表格,显示比对后的配对序列 NCBI 的编号、序列描述等,右边是比对数据,可自行选择排列次序,例如要按序列相似度由高至低的排列显示可点选"Max ident"超连接。结果最下部分是查诣序列和配对序列的排序图。

实际操作时将待测品的 DNA 序列透过 BLAST 作比对后,便可初步估计其身份。但如果配对序列和检索序列的重叠率(query coverage)太低,就算比对的相似度达 100%,亦切勿将 BLAST 的结果作为鉴定依据。举例,1 条 500 个碱基的检索序列经 BLAST 比对后得出 2 条配对序列,排第 1 位的序列来自物种 A,而第 2 的是物种 B,比对结果显示两者跟检索序列的相似度都同样是 100%。但在 BLAST 的结果图表上,物种 A 的整条序列都呈现红色,但物种 B 只有第 1 ~ 200 位点是红色,其余是黑色。如果研究人员未有检查检索序列和配对序列的重叠率,便将 BLAST 所显示物种 A 和 B 的相似度作为依据,就会错误地将检索序列判定为物种 A 或物种 B,但事实上检索序列只跟物种 A 一致,和物 B 存在差异。因此,当需要计算序列间的整体差异,应采用全局性排序。

4.3 全局性排序计算机程序 ClustalW 全局性排序可完全反映序列之间整体差异。ClustalW 由 European Bioinformatics Institute (EBI)提供的多序列比较程序,兼容核酸和蛋白质序列,属全局性排序(global alignment)算法。用户将序列以 FASTA 格式上载到服务器上,程序便会排比及计算 2 条序列之间的相似度,连同序列比较图及系统树显示。除了在线分析外,一些序列分析软件包亦内置 ClustalW 功能。

程序计算后所得排序结果,未必完全正确。如果序列长度一致、物种同源性高、或属于蛋白质编码基因(如叶绿体 ribulose-bisphosphate carboxylase, *rbc*L),序列排序的结果一般不需要微调。相反,序列长度不一,物种同源性低,又或是基因间隔区(如核糖体 ITS 或叶绿体 *trn*H-*psb*A),往往需要作出修正。以 *trn*H-*psb*A 为例,这个区域由 3 部分组成,第一部分是 tRNA-His (*trn*H)基因的末端,紧接是 *trn*H 和 photosystem II protein D1 (*psb*A) 2 个基因之间的区域,第三部分是 *psb*A 基因的开端。*trn*H 和 *psb*A 基因是编码区,所以较为保守。相反两者之间的区域因不受进化限制,所以常有碱基的插入/缺失等情况,导致长度不一。事实上,本小组曾对多种植物的 *trn*H-*psb*A 测序,发现大部分植物的 *trn*H-*psb*A 序列介乎 400 至 600 bp,最短的只有 300 bp 左右,最长超过 1 000 bp,可见其变化之大。当序列长度不一时,程序会在较短的序列中加入空隙(gap),其法则是先分析序列中保守区位置,然后在较短序列的 2 个保守区之间加入空隙,而这个动作是取决于用户输入参数。如果 2 个保守区域相隔太远,程序未必能辨认出,所加入的空隙不正确,最明显的例子是序列间的开端或末端并非对齐。*trn*H 和 *psb*A 是保守的编码区,所以序列间的开端和末端必定是排在一起,如果出现较短序列的末端,跟较长序列的中间位置而非末端对齐,即代表所加入的空隙不足。解决方法之一是改变 ClustalW 的空隙罚分(gap penalty)。另外,亦可在分析软件手动调整。

5 分析 DNA 序列

5.1 序列分析的种类 当完成物种分类调查、下载 DNA 序列、BLAST 比对、以及排序,最后一步是分析序列。如果目的是设计特异性 PCR 引物,应寻找序列之间的特异性位点。如果目的是鉴定待测品,可选择两种计算方法。第一种是计算待测品和标准品之间的序列相似度,第二种是计算两者之间的遗传距离(genetic distance)。

5.2 序列相似度的计算法 序列相似度的计算法,是将 2 条序列之间相同碱基的数目除以序列长度,求出百分比值[12]。当分析 3 条或以上序列时,会得出相似度阵列(similarity matrix),3 条序列有 3 个数值,4 条序列有 6 个,5 条序列增加至 10 个,如此类推。使用在线 ClustalW 会显示 2 条序列间的相似度,将之输入其他分析软件作后续分析。但如果分析的序列数量十分多,手动输入方式变得费时。BioEdit 是一个免费的序列分析软件[13],提供图形用户界面,内建 ClustalW 和亲缘树功能,并可自动生成序列相似度阵列,阵列导出后可于 Microsoft Excel 开启,免除手动逐一输入数值。

5.3 遗传距离计算法 另一类计算法是基于遗传距离(genetic distance),最小距离代表物种间在遗传关系上相较最近,较大距离则相反。鉴定物种一般采用 uncorrected p-distance 和 Kimura 2-parameter (K2P)模型[14-17]。经计算后,研究人员更可将遗传距离建构种系发生树,描述各个物种和待测品之间可能有的亲缘关系。Molecular Evolutionary Genetics Analysis (MEGA)是一个免费的分子系统学分析软件[18]。初版于 1993 年发行,以图形用户界面操作。和 BioEdit 一样,MEGA 内建 ClustalW,亦可作简单的排序结果微调。透过 MEGA 可计算 uncorrected p-distance 和 K2P 等遗传距离,使用者可更改种系发生树的各项参数,如 Bootstrap 抽样、选择不同树形图类型等。最新版本是 5.0 Beta 6.1,首次加入 maximum likelihood 算法,令功能更完备。

6 结语

本文说明生物信息学在分子鉴别中药材的应用,介绍有关数据库和分析工具的操作策略和重点。在众多生物信息学工具当中,数据库的角色至为重要。现时序列数据库所存

储的纪录,物种繁多,序列数目庞大。因此,NCBI 可容许用户 BLAST 搜寻个别的物种,如人和鼠。此外,为研究特定类别生物的序列数据库也应运而生。本组已于 2010 完成世界上第 1 个以鉴定中药材为目标的 DNA 序列数据库,推动中药材的分子鉴定。随着新颖快速的分子检测方法面世和执法单位愈来愈重视利用分子方法鉴定中药材,将有利推动分子鉴别中药材的发展和壮大相应生物信息学的研究。

[参考文献]

[1] Hogeweg P. The roots of bioinformatics in theoretical biology[J]. PLoS Comput Biol,2011,7(3):e1002021.
[2] 中国药典. 一部[S]. 2010.
[3] Editorial Team. National center for biotechnology information [DB/OL]. http://www.ncbi.nlm.nih.gov,2011-12-14.
[4] 中国科学院植物研究所系统与进化植物学国家重点实验室. 中国植物志[DB/OL]. http://frps.plantphoto.cn,2011-12-14.
[5] Editorial Team. The international plant names Index[DB/OL]. http://www.ipni.org,2011-12-14.
[6] Lou S K,Wong K L,Li M,et al. An integrated web medicinal materials DNA database: MMDBD (Medicinal Materials DNA Barcode Database)[J]. BMC Genomics,2010,11:402.
[7] European Bioinformatics Institute. EMBL Nucleotide Sequence Database[DB/OL]. http://www.ebi.ac.uk/embl,2011-12-14.
[8] Center for Information Biology and DNA Data Bank of Japan. DNA Data Bank of Japan[DB/OL]. http://www.ddbj.nig.ac.jp, 2011-12-14.
[9] Altschul S F,Madden T L,Schäffer A A,et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs[J]. Nucleic Acids Res,1997,25(17):3389.
[10] Thompson J D,Higgins D G,Gibson T J,et al. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice[J]. Nucleic Acids Res,1994,22(22):4673.
[11] W3C. XML technology[DB/OL]. http://www.w3.org/standards/xml,2011-12-14.
[12] European Bioinformatics Institute. ClustalW help file[DB/OL]. http://www.ebi.ac.uk/tools/msa/clustalw2/help/faq.html#21,2011-12-14.
[13] Hall T A. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT[J]. Nucl Acids Symp Ser,1999,41:95.
[14] Ha W Y,Reid D G,Lam W L,et al. Genetic differentiation between fake abalone and genuine *Haliotis* species using the forensically informative nucleotide sequencing (FINS) method[J]. J Agric Food Chem,2011,59:5195.
[15] Kress W J,Wurdack K J,Zimmer E A,et al. Use of DNA barcodes to identify flowering plants[J]. Proc Natl Acad Sci USA,2005,102(23):8369.
[16] Guo X,Wang X,Su W,et al. DNA barcodes for discriminating the medicinal plant *Scutellaria baicalensis* (Lamiaceae) and its adulterants[J]. Biol Pharm Bull,2011,34(8):1198.
[17] 刘震,陈科力,罗焜等. 忍冬科药用植物 DNA 条形码通用序列的筛选[J]. 中国中药杂志,2010,35(19):2527.
[18] Tamura K,Peterson D,Peterson N,et al. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods[J]. Mol Biol Evol, 2011,28(10):2731.

Application of bioinformatics in molecular authentication of traditional Chinese medicinal materials

WONG Kalok[1], SHAW Pangchui[1,2]*
(1. *State Key Laboratory of Phytochemistry and Plant Resources in West China (CUHK), Institute of Chinese Medicine, The Chinese University of Hong Kong, Hong Kong, China*;
2. *School of Life Science, The Chinese University of Hong Kong, Hong Kong, China*)

[**Abstract**] Benefiting from various DNA technologies, DNA markers have now become a popular means for the identification of Chinese medicinal materials. Facing the huge amount of valuable data that has been produced, researchers need to understand the bioinformatics tools for analyzing the obtained DNA information. This paper summarizes the applications of bioinformatics in molecular authentication of Chinese medicinal materials, including checking phylogenetic information of the samples, searching and retrieving DNA sequence data, matching of similarity between the sequences and performing multiple sequence alignment.

[**Key words**] Chinese medicinal materials; authentication; bioinformatics

doi:10.4268/cjcmm20120805

[责任编辑 吕冬梅]