中药材参考 DNA 序列库
参考 DNA 序列的一般使用指引

目的

政府中药检测中心(“检测中心”)建立中药材参考 DNA 序列库，载列各种中药材的参考 DNA 序列。本指引旨在概述参考 DNA 序列的用法，介绍其三大用途，并说明如何通过两种常见做法发挥上述三大用途。

背景

为控制中药材的质量及确保其安全稳当，中药材的鉴别工作必须准确无误。众所周知，DNA 技术是适用于鉴别物种的方法之一，原因是 DNA 一般不受生物的年龄、生理状况和生境所影响。DNA 条形码技术利用来自一个或多个特定 DNA 区域(又称“DNA 条形码”)的信息进行鉴别，是最为广泛用于鉴别生物的 DNA 技术之一。就特定 DNA 区域内的 DNA 而言，不同生物之间存有显著差异，但在同一物种的不同个体之间，这方面的差异则较小。因此，在区分近缘中药材物种、形态上易于混淆的中药材和没有独特化学指标的中药材时，DNA 条形码技术尤其有用。

中药材参考 DNA 序列库所载列的参考 DNA 序列，均源自分析已知分类位置的中药材标本。建立此参考 DNA 序列库的目的，是希望透过提供多样化的检测服务，提升本地检测服务的水平。检测中心会拟备资料表，记述所选用 DNA 条形码的参考 DNA 序列及其相关资料。每种中药材的资料表均有一个部分是按 DNA 条形码划分，分别以 FASTA 格式列出所有标本的参考 DNA 序列，这有助快速处理参考 DNA 序列，以便进行数据分析(有关详情，请参阅中药材参考 DNA 序列库的“通用公告”)。

参考 DNA 序列的用法

中药材参考 DNA 序列库的参考 DNA 序列均来自凭证标本和经专家鉴

定的中药材，并利用检测中心内部验证的方法产生，发挥以下三大用途，计有(1)鉴别物种、(2)显示生物的亲缘关系及(3)传递遗传信息，以便制订全新的检测方法。所选用的 DNA 条形码适用于鉴别物种，已获多部药典采用，并得到科学界广泛认可。

(1) 鉴别物种

鉴别物种的方法是进行序列比对：比较样本所产生的序列与参考序列来判断两者的相似度。两者具有愈多共同的核苷酸，相似度愈高，在功能、结构及 / 或演化方面有关联的机会愈大。

(2) 亲缘关系分析

生物的演化史和关系对物种分类至关重要。科学家通过研究生物的可遗传特征(例如 DNA 序列、形态、行为特征)推断生物的亲缘关系。我们只要根据生物的特征为生物建立巢状羣组，便可绘制出显示上述假定关系的图表，称为亲缘关系图。使用一组或多组 DNA 条形码绘制的亲缘关系图有助区分高遗传相似度的近缘物种。

(3) 制订检测方法

序列比对亦可找出中药材与相似物种的多型性位置。种特异性鉴别方法(例如特异性聚合酶链式反应(“特异性 PCR”)和聚合酶链式反应－限制性片段长度多态性)是根据物种的多型性核苷酸制订，可用来鉴别目标中药材物种或区分中药材真品与常见的伪品，这正是制订特异性 PCR 测试的要旨。与 DNA 条形码技术比较，特异性 PCR 测试的处理速度较快，结果易于分析，因此成本较低。

要使参考 DNA 序列发挥上述用途(包括鉴别物种及制订检测方法)，两种常见的做法是(1)进行序列排比及(2)利用公开资料库进行序列比对。

(1) 序列排比

序列排比可透过比较两个或以上 DNA 序列，找出其中相似之处。双序列排比务求以最佳方式，对两个查询序列的字元进行排列和并列分析。多序列排比与双序列排比类近，但在每次比对中同时分析两个以上 DNA 序列。由于 DNA 条形码的长度大多为 300 至 600 个碱基对，因此以人手对 DNA 条形码进行并列分析并不可行。下文将推介数个用作序列排比的电脑程式和软件。

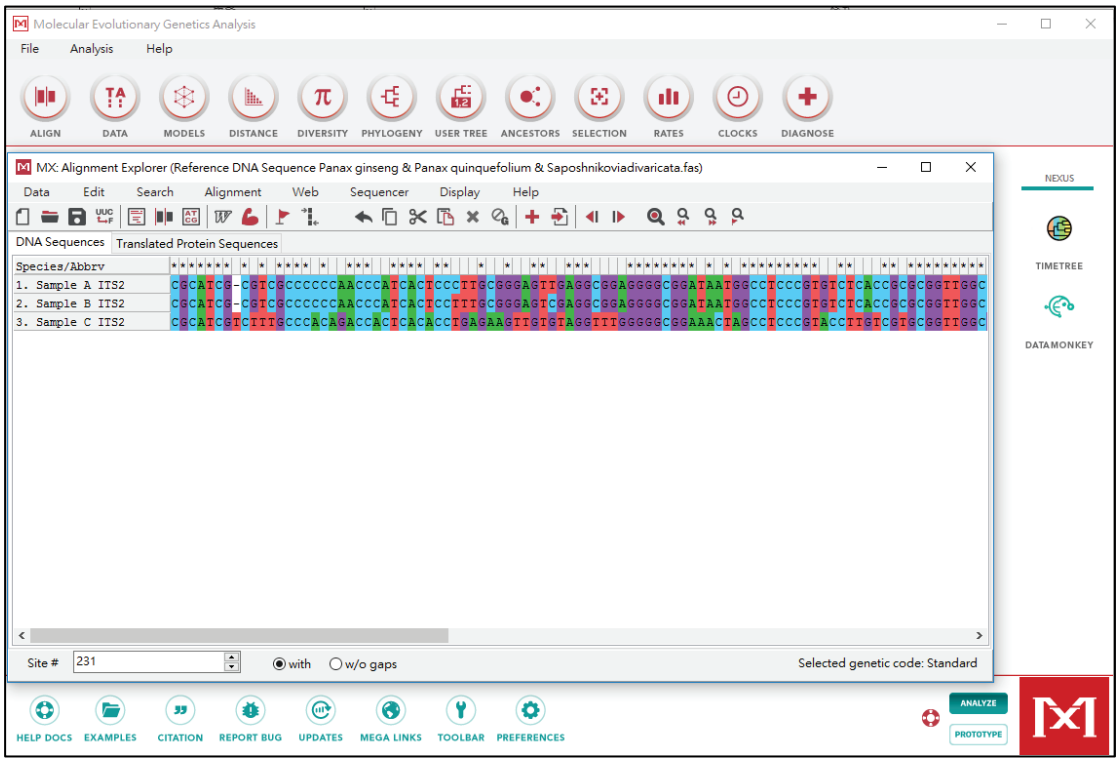
用作序列排比的电脑程式

双序列排比和多序列排比可透过使用序列排比程式进行。擅于进行多序列排比的电脑程式包括 Clustal Omega、ClustalW、MUSCLE(MUltiple Sequence Comparison by Log-Expectation，即基于对数期望的多重序列比较)、MACSE (Multiple Alignment of Coding SEquences，即编码序列的多重比对)、T-Coffee(Tree-based Consistency Objective Function for Alignment Evaluation，即以树形基础的一致性作多重序列比对)等。Clustal Omega 和 ClustalW(网址：http://www.clustal.org/)是通用的多序列排比程式，用于分析蛋白质和 DNA／RNA。Clustal Omega 是最新版本的 Clustal 系列电脑程式，与之前的版本相比，可进行较大规模的并列分析。MUSCLE(网址：https://www.drive5.com/muscle/)的优点则在于其速度和进行并列分析的准确度。上述程式大多适用于 Windows、Mac OS 和 Unix／Linux 电脑，并可以指令列命令模式运行。

用作序列排比的软件套装

分子进化遗传学分析(Molecular Evolutionary Genetics Analysis, MEGA)是免费的序列排比软件，可用作比较分析 DNA 序列。MEGA 配备了 ClustalW 和 MUSCLE 程式，以图形使用者介面操作。使用者可将包含测试样本的 DNA 序列和参考 DNA 序列的 FASTA 档案汇入 MEGA 以进行序列排比。MEGA 以不同颜色标示四种核苷酸(分别以“A”、“G”、“C”和“T”代表)，因此使用者

可轻易从分析结果找出置顶序列(通常是参考序列，如图一所示)与其余序列之间在核苷酸排列上的差异。除了可用于进行序列排比，MEGA 更可用作绘制亲缘关系图和计算遗传距离。



图一：利用 MEGA 软件进行序列排比的例子

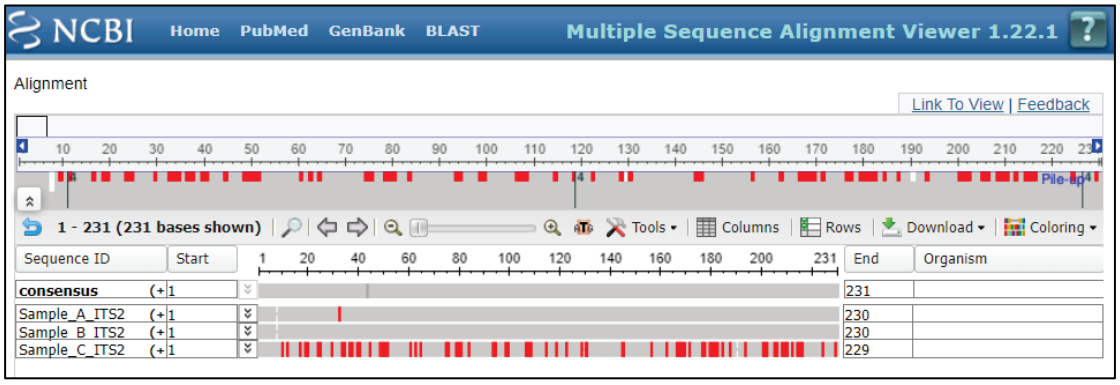
BioEdit 生物序列排列编辑器和 CodonCode Aligner 也是深受研究人员欢迎的序列排比软件。BioEdit 生物序列排列编辑器是适合入门者使用的免费软件，内置多种编辑功能，例如序列排序、计算序列一致性百分比和绘制亲缘关系图等。另外，此软件的使用者介面简单易用，可汇入、汇出、操控和检视以 ClustalW 程式进行序列排比所得出的结果。CodonCode Aligner 则是商业软件套装，擅于编辑核苷酸的顺序和监控原始序列数据的质量分数。使用者可以之修剪序列的两端部分、组装序列、编辑片段重迭羣和侦测突变，亦可利用内置程式 MUSCLE、MACSE 或 Clustal Omega 进行序列排比。

可进行序列排比的网上平台

网上也有多个平台提供可进行多序列排比的电脑程式。欧洲分子生物

学实验室欧洲生物资讯研究所(EMBL-EBI)旨在协助研究人员互相分享和分析生物数据，其官方网站(网址：https://www.ebi.ac.uk/)提供了各种网上工具，例如 Clustal Omega、MAFFT(Multiple Alignment using Fast Fourier Transform，即是使用快速傅立叶变换之多重比对)、MUSCLE、T-coffee 等。在该网页上的表格中，使用者可以“复制和贴上”的方式输入多个 DNA 序列进行分析，更可按照需要和工作流程调节各项参数，例如汇出序列分析结果的排列顺序和格式等。

NCBI Multiple Sequence Alignment Viewer 是由美国国家生物技术资讯中心(National Center for Biotechnology Information，简称 NCBI)管理的网上应用程式(网址：https://www.ncbi.nlm.nih.gov/projects/msaviewer/)，以图表方式展示多序列排比分析。如样本 DNA 序列中的核苷酸与参考 DNA 序列(置顶序列)中的核苷酸有所不同，差异会以红色标示(如图二所示)，让使用者直观地从并列分析结果看出 DNA 序列之间的差异。



图二：利用网上平台 NCBI Multiple Sequence Alignment Viewer 进行序列排比的例子

(2) 利用公开资料库比对 DNA 序列

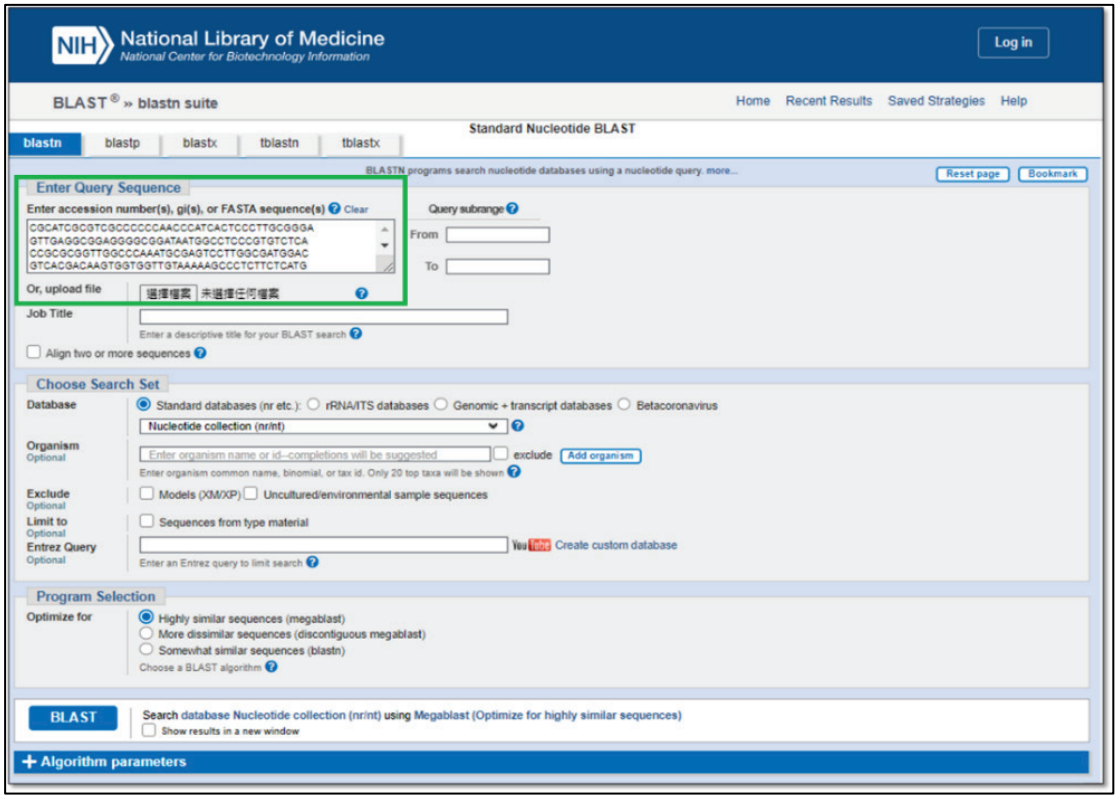
GenBank

NCBI GenBank 是由 NCBI 管理的综合资料库，收录采自 50 多万个物种的超过 25 亿个序列，是最大型的序列资料库之一。GenBank 会为收录的每个序列记录编配一个称为登录号(accession number)的专用识别号，而每

个序列记录包括有关生物来源、基因座定性、作者、序列特征等。NCBI 接纳全球科学界提交的序列。须予留意的是，部分被提交的序列是来自未获专家检定的生物物料(例如中药材商业制品)。总括而言，产自凭证标本或可追溯标本的核苷酸序列是比较可靠的参考来源。

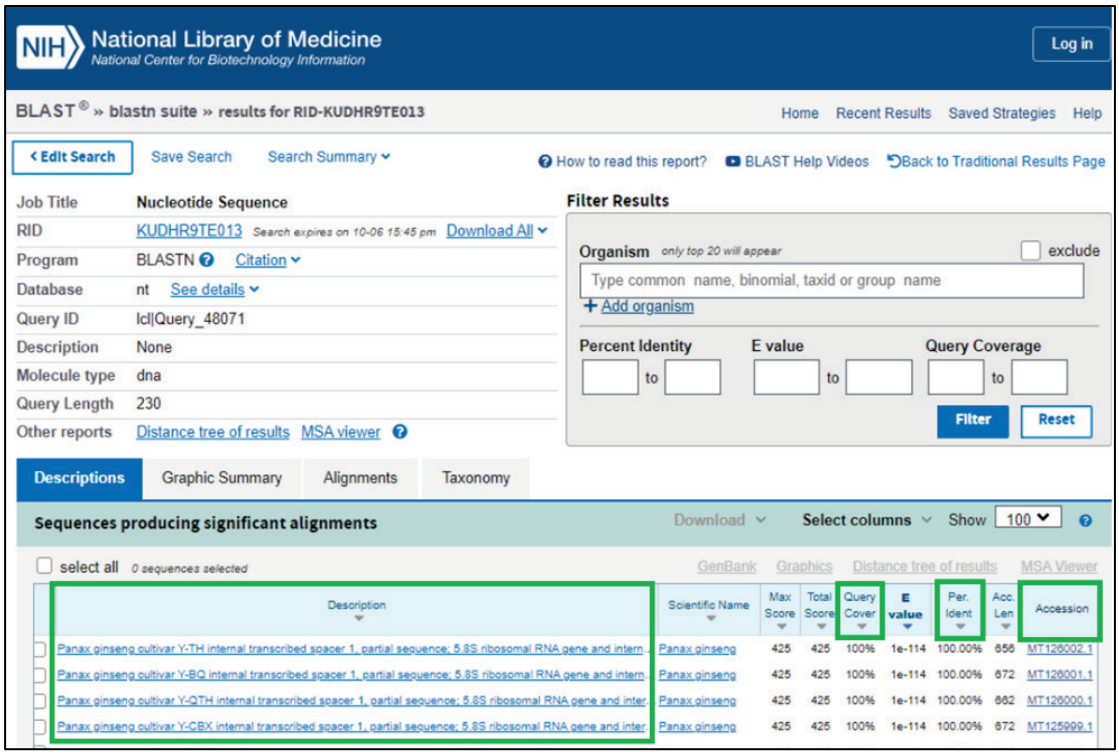
GenBank 的基本局部排比搜索工具(Basic Local Alignment Search Tool，简称 BLAST)

BLAST 是用于比照不同数据库进行序列相似度搜索的程式。GenBank 有多个版本的 BLAST 程式，专为比照个别数据库进行序列搜索而设。就中药材的 DNA 条形码技术而言，最常用的版本是比照核苷酸资料库进行搜索的 BLASTn (网址：https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch)。使用者可于网上介面输入查询序列及调节参数，然后 BLAST 会找出查询序列与 NCBI GenBank 收录核苷酸序列在统计学上的相似度(图三)。



图三：上载序列到 BLAST 搜索平台的例子

BLAST 搜索结果会胪列 NCBI GenBank 中与查询序列相符的核苷酸序列，并将最相符者置顶(图四)。分析 BLAST 搜索结果时，使用者应着眼于两个数字："一致性百分比"(percent of identity)和"查询覆盖度"(query coverage)。"一致性百分比"代表查询序列与 GenBank 内有关核苷酸序列的相似度百分比，而"查询覆盖度"则代表查询序列与 GenBank 内核苷酸序列重迭的百分比。



图四：参考 DNA 序列比对 NCBI GenBank 收录核苷酸序列的 BLAST 搜索结果例子

其他公开资料库

生命条形码数据系统(Barcode of Life Data Systems，简称 BOLD；网址：https://www.boldsystems.org/index.php/IDS_OpenIdEngine)是用途广泛的网上平台，提供核苷酸序列(特别是有关动物者)数据储存和分析服务。

参考资料

1. C Notredame, DG Higgins, J Heringa. T-Coffee: A novel method for multiple sequence alignments. Journal of Molecular Biology, 2000, 302(1): 205-217.

2. CodonCode Corporation. CodonCode Aligner (Version 9.0). 2019. Available from: https://www.codoncode.com/index.htm

3. EW Sayers, EE Bolton, JR Brister, K Canese, J Chan, et al. Database resources of the national center for biotechnology information. Nucleic Acids Research, 2022, 50(D1): D20-D26.

4. K Katoh, K Misawa, K Kuma, T Miyata. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research, 2002, 30(14): 3059-3066.

5. K Tamura, G Stecher, S Kumar. MEGA11: Molecular Evolutionary Genetics Analysis Version 11. Molecular Biology and Evolution, 2021, 38(7): 3022-3027.

6. MA Larkin, G Blackshields, NP Brown, R Chenna, PA McGettigan, et al. Clustal W and Clustal X version 2.0. Bioinformatics, 2007, 23(21): 2947-2948.

7. National Center for Biotechnology Information. Multiple Sequence Alignment Viewer (Version 1.22.0). 2022. Available from: https://www.ncbi.nlm.nih.gov/projects/msaviewer/

8. RC Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research, 2004, 32(5): 1792-1797.

9. S Ratnasingham, PDN Hebert. BOLD: The Barcode of Life Data System (www.barcodinglife.org). Molecular Ecology Notes, 2007, 7(3): 355-364.

10. SF Altschul, W Gish, W Miller, EW Myers, DJ Lipman. Basic local alignment search tool. Journal of Molecular Biology, 1990, 215(3): 403-410.

11. TA Hall. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic acids symposium series, 1999, 41: 95-98.

12. V Ranwez, S Harispe, F Delsuc, EJ Douzery. MACSE: Multiple Alignment of Coding SEquences accounting for frameshifts and stop codons. PLoS One, 2011, 6(9): e22594.