

# 中藥材參考 DNA 序列庫

## 參考 DNA 序列的一般使用指引

### 目的

政府中藥檢測中心(“檢測中心”)建立中藥材參考 DNA 序列庫，載列各種中藥材的參考 DNA 序列。本指引旨在概述參考 DNA 序列的用法，介紹其三大用途，並說明如何通過兩種常見做法發揮上述三大用途。

### 背景

為控制中藥材的質量及確保其安全穩當，中藥材的鑒別工作必須準確無誤。眾所周知，DNA 技術是適用於鑒別物種的方法之一，原因是 DNA 一般不受生物的年齡、生理狀況和生境所影響。DNA 條形碼技術利用來自一個或多個特定 DNA 區域(又稱“DNA 條形碼”)的信息進行鑒別，是最為廣泛用於鑒別生物的 DNA 技術之一。就特定 DNA 區域內的 DNA 而言，不同生物之間存有顯著差異，但在同一物種的不同個體之間，這方面的差異則較小。因此，在區分近緣中藥材物種、形態上易於混淆的中藥材和沒有獨特化學指標的中藥材時，DNA 條形碼技術尤其有用。

中藥材參考 DNA 序列庫所載列的參考 DNA 序列，均源自分析已知分類位置的中藥材標本。建立此參考 DNA 序列庫的目的，是希望透過提供多樣化的檢測服務，提升本地檢測服務的水平。檢測中心會擬備資料表，記述所選用 DNA 條形碼的參考 DNA 序列及其相關資料。每種中藥材的資料表均有一個部分是按 DNA 條形碼劃分，分別以 FASTA 格式列出所有標本的參考 DNA 序列，這有助快速處理參考 DNA 序列，以便進行數據分析(有關詳情，請參閱中藥材參考 DNA 序列庫的“通用公告”)。

### 參考 DNA 序列的用法

中藥材參考 DNA 序列庫的參考 DNA 序列均來自憑證標本和經專家鑒

定的中藥材，並利用檢測中心內部驗證的方法產生，發揮以下三大用途，計有(1)鑒別物種、(2)顯示生物的親緣關係及(3)傳遞遺傳信息，以便制訂全新的檢測方法。所選用的 DNA 條形碼適用於鑒別物種，已獲多部藥典採用，並得到科學界廣泛認可。

### (1) 鑒別物種

鑒別物種的方法是進行序列比對：比較樣本所產生的序列與參考序列來判斷兩者的相似度。兩者具有愈多共同的核苷酸，相似度愈高，在功能、結構及／或演化方面有關聯的機會愈大。

### (2) 親緣關係分析

生物的演化史和關係對物種分類至關重要。科學家通過研究生物的可遺傳特徵(例如 DNA 序列、形態、行為特徵)推斷生物的親緣關係。我們只要根據生物的特徵為生物建立巢狀羣組，便可繪製出顯示上述假定關係的圖表，稱為親緣關係圖。使用一組或多組 DNA 條形碼繪製的親緣關係圖有助區分高遺傳相似度的近緣物種。

### (3) 制訂檢測方法

序列比對亦可找出中藥材與相似物種的多型性位置。種特異性鑒別方法(例如特異性聚合酶鏈式反應(“特異性 PCR”)和聚合酶鏈式反應－限制性片段長度多態性)是根據物種的多型性核苷酸制訂，可用來鑒別目標中藥材物種或區分中藥材真品與常見的偽品，這正是制訂特異性 PCR 測試的要旨。與 DNA 條形碼技術比較，特異性 PCR 測試的處理速度較快，結果易於分析，因此成本較低。

要使參考 DNA 序列發揮上述用途(包括鑒別物種及制訂檢測方法)，兩種常見的做法是(1)進行序列排比及(2)利用公開資料庫進行序列比對。

## (1) 序列排比

序列排比可透過比較兩個或以上 DNA 序列，找出其中相似之處。雙序列排比務求以最佳方式，對兩個查詢序列的字元進行排列和並列分析。多序列排比與雙序列排比類近，但在每次比對中同時分析兩個以上 DNA 序列。由於 DNA 條形碼的長度大多為 300 至 600 個鹼基對，因此以人手對 DNA 條形碼進行並列分析並不可行。下文將推介數個用作序列排比的電腦程式和軟件。

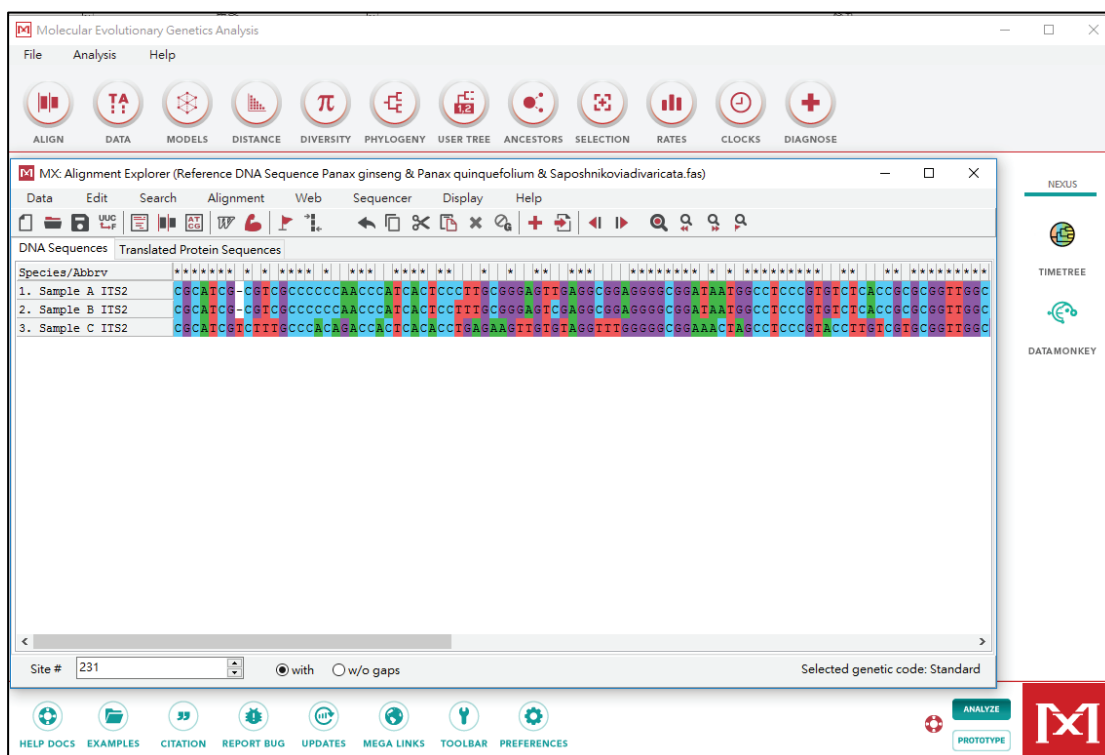
### *用作序列排比的電腦程式*

雙序列排比和多序列排比可透過使用序列排比程式進行。擅於進行多序列排比的電腦程式包括 Clustal Omega、ClustalW、MUSCLE(Multiple Sequence Comparison by Log-Expectation，即基於對數期望的多重序列比較)、MACSE (Multiple Alignment of Coding SEquences，即編碼序列的多重比對)、T-Coffee(Tree-based Consistency Objective Function for Alignment Evaluation，即以樹形基礎的一致性作多重序列比對)等。Clustal Omega 和 ClustalW(網址：<http://www.clustal.org/>)是通用的多序列排比程式，用於分析蛋白質和 DNA/RNA。Clustal Omega 是最新版本的 Clustal 系列電腦程式，與之前的版本相比，可進行較大規模的並列分析。MUSCLE(網址：<https://www.drive5.com/muscle/>)的優點則在於其速度和進行並列分析的準確度。上述程式大多適用於 Windows、Mac OS 和 Unix/Linux 電腦，並可以指令列命令模式運行。

### *用作序列排比的軟件套裝*

分子進化遺傳學分析(Molecular Evolutionary Genetics Analysis, MEGA)是免費的序列排比軟件，可用作比較分析 DNA 序列。MEGA 配備了 ClustalW 和 MUSCLE 程式，以圖形使用者介面操作。使用者可將包含測試樣本的 DNA 序列和參考 DNA 序列的 FASTA 檔案匯入 MEGA 以進行序列排比。MEGA 以不同顏色標示四種核苷酸(分別以“A”、“G”、“C”和“T”代表)，因此使用者

可輕易從分析結果找出置頂序列(通常是參考序列，如圖一所示)與其餘序列之間在核苷酸排列上的差異。除了可用於進行序列排比，MEGA 更可用作繪製親緣關係圖和計算遺傳距離。



圖一：利用 MEGA 軟件進行序列排比的例子

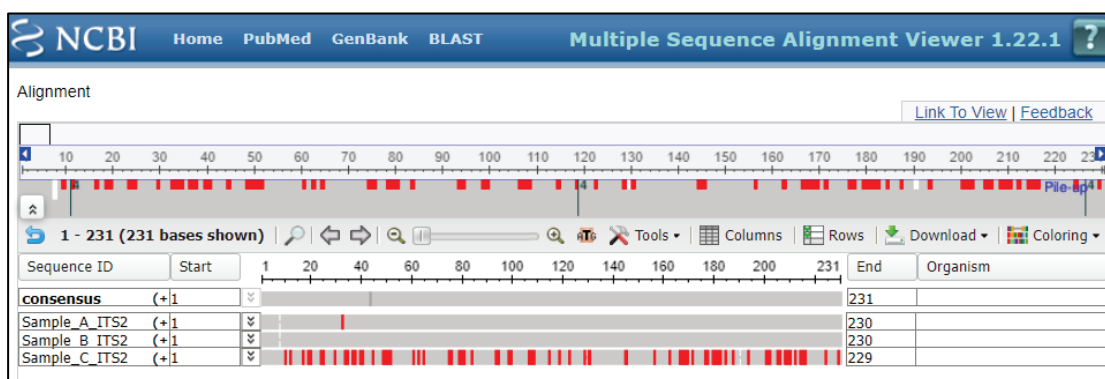
BioEdit 生物序列排列編輯器和 CodonCode Aligner 也是深受研究人員歡迎的序列排比軟件。BioEdit 生物序列排列編輯器是適合入門者使用的免費軟件，內置多種編輯功能，例如序列排序、計算序列一致性百分比和繪製親緣關係圖等。另外，此軟件的使用者介面簡單易用，可匯入、匯出、操控和檢視以 ClustalW 程式進行序列排比所得出的結果。CodonCode Aligner 則是商業軟件套裝，擅於編輯核苷酸的順序和監控原始序列數據的質量分數。使用者可以之修剪序列的兩端部分、組裝序列、編輯片段重疊羣和偵測突變，亦可利用內置程式 MUSCLE、MACSE 或 Clustal Omega 進行序列排比。

### 可進行序列排比的網上平台

網上也有多個平台提供可進行多序列排比的電腦程式。歐洲分子生物

學實驗室歐洲生物資訊研究所(EMBL-EBI)旨在協助研究人員互相分享和分析生物數據，其官方網站(網址：<https://www.ebi.ac.uk/>)提供了各種網上工具，例如 Clustal Omega、MAFFT(Multiple Alignment using Fast Fourier Transform，即是使用快速傅立葉變換之多重比對)、MUSCLE、T-coffee 等。在該網頁上的表格中，使用者可以“複製和貼上”的方式輸入多個 DNA 序列進行分析，更可按照需要和工作流程調節各項參數，例如匯出序列分析結果的排列順序和格式等。

NCBI Multiple Sequence Alignment Viewer 是由美國國家生物技術資訊中心(National Center for Biotechnology Information，簡稱 NCBI)管理的網上應用程式(網址：<https://www.ncbi.nlm.nih.gov/projects/msviewer/>)，以圖表方式展示多序列排比分析。如樣本 DNA 序列中的核苷酸與參考 DNA 序列(置頂序列)中的核苷酸有所不同，差異會以紅色標示(如圖二所示)，讓使用者直觀地從並列分析結果看出 DNA 序列之間的差異。



圖二：利用網上平台 NCBI Multiple Sequence Alignment Viewer 進行序列排比的例子

## (2) 利用公開資料庫比對 DNA 序列

### *GenBank*

NCBI GenBank 是由 NCBI 管理的綜合資料庫，收錄採自 50 多萬個物種的超過 25 億個序列，是最大型的序列資料庫之一。GenBank 會為收錄的每個序列記錄編配一個稱為登錄號(accession number)的專用識別號，而每

個序列記錄包括有關生物來源、基因座定性、作者、序列特徵等。NCBI 接納全球科學界提交的序列。須予留意的是，部分被提交的序列是來自未獲專家檢定的生物物料(例如中藥材商業製品)。總括而言，產自憑證標本或可追溯標本的核苷酸序列是比較可靠的參考來源。

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

BLAST 是用於比照不同數據庫進行序列相似度搜索的程式。GenBank 有多個版本的 BLAST 程式，專為比照個別數據庫進行序列搜索而設。就中藥材的 DNA 條形碼技術而言，最常用的版本是比照核苷酸資料庫進行搜索的 BLASTn (網址：[https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE\\_TYPE=BlastSearch](https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch))。使用者可於網上介面輸入查詢序列及調節參數，然後 BLAST 會找出查詢序列與 NCBI GenBank 收錄核苷酸序列在統計學上的相似度(圖三)。

The screenshot shows the NCBI BLASTn search interface. The main heading is "BLAST® -> blastn suite" with sub-heading "Standard Nucleotide BLAST". The "blastn" tab is selected. The "Enter Query Sequence" section is highlighted with a green box and contains a text area with a nucleotide sequence: "CGCATCGGTCGCCCCCAACCCATCACTCCCTTGGGGGA GTTGAGGCGGAGGGGCGGATAATGGCCCTCCCGTGTCTCA CCGCCGGTTGGCCCAATGCGAGTCCTTGGCGATGGAC GTACGACAAAGTGTGTTTAAAAAGCCCTTCTTCATG". Below the text area are fields for "Job Title", "Align two or more sequences", "Choose Search Set" (Database: Standard databases (nr etc.), Organism: Nucleotide collection (nr/nt), Exclude: Models (XM/XP), Limit to: Sequences from type material), "Program Selection" (Optimize for: Highly similar sequences (megablast)), and a "BLAST" button. A footer section shows "+ Algorithm parameters".

圖三：上載序列到 BLAST 搜索平台的例子

BLAST 搜索結果會臚列 NCBI GenBank 中與查詢序列相符的核苷酸序列，並將最相符者置頂(圖四)。分析 BLAST 搜索結果時，使用者應着眼於兩個數字：“一致性百分比”(percent of identity)和“查詢覆蓋度”(query coverage)。“一致性百分比”代表查詢序列與 GenBank 內有關核苷酸序列的相似度百分比，而“查詢覆蓋度”則代表查詢序列與 GenBank 內核苷酸序列重疊的百分比。

The screenshot displays the NCBI BLAST search results for query RID-KUDHR9TE013. The search parameters are as follows:

- Job Title: Nucleotide Sequence
- RID: KUDHR9TE013
- Program: BLASTN
- Database: nt
- Query ID: lcl|Query\_48071
- Molecule type: dna
- Query Length: 230

The filter results section shows the following criteria:

- Organism: (empty)
- Percent Identity: (empty) to (empty)
- E value: (empty) to (empty)
- Query Coverage: (empty) to (empty)

The table of sequences producing significant alignments is as follows:

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc Len	Accession
Panax ginseng cultivar Y-TH internal transcribed spacer 1, partial sequence, 5.8S ribosomal RNA gene and intern...	Panax ginseng	425	425	100%	1e-114	100.00%	656	MT128002.1
Panax ginseng cultivar Y-BQ internal transcribed spacer 1, partial sequence, 5.8S ribosomal RNA gene and intern...	Panax ginseng	425	425	100%	1e-114	100.00%	672	MT128001.1
Panax ginseng cultivar Y-QTH internal transcribed spacer 1, partial sequence, 5.8S ribosomal RNA gene and inter...	Panax ginseng	425	425	100%	1e-114	100.00%	662	MT128000.1
Panax ginseng cultivar Y-CBX internal transcribed spacer 1, partial sequence, 5.8S ribosomal RNA gene and inter...	Panax ginseng	425	425	100%	1e-114	100.00%	672	MT125999.1

圖四：參考 DNA 序列比對 NCBI GenBank 收錄核苷酸序列的 BLAST 搜索結果例子

### 其他公開資料庫

生命條形碼數據系統(Barcode of Life Data Systems, 簡稱 BOLD; 網址：[https://www.boldsystems.org/index.php/IDS\\_OpenIdEngine](https://www.boldsystems.org/index.php/IDS_OpenIdEngine))是用途廣泛的網上平台，提供核苷酸序列(特別是有關動物者)數據儲存和分析服務。

### 參考資料

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/msviewer/>
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](http://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.