Reference DNA Sequence Library for Chinese Materia Medica
General Note on the Use of Reference DNA Sequences

Purpose

This note aims to provide an overview of the use of reference DNA sequences from the Reference DNA Sequence Library for Chinese Materia Medica (CMMRSL) established by the Government Chinese Medicines Testing Institute (GCMTI). Three aspects of using the reference DNA sequences and two common approaches to achieve these aspects are introduced in this note.

Background

Accurate identification is important for the quality control and safety assurance of Chinese Materia Medica (CMM). It is well recognised that DNA technique is a suitable approach for species identification as in general, DNA is not affected by ages, physiological conditions and habitats of organisms. DNA barcoding, which utilises information from one or a few specified DNA regions, also known as DNA barcodes, is one of the most widely used DNA techniques for organism identification. These specific DNA regions usually contain significant DNA variations among organisms but low variability across individuals of the same species. Hence, DNA barcoding is particularly useful for discrimination of closely related CMM species, morphologically confused CMMs and CMMs without unique chemical markers.

The CMMRSL is a repository of reference DNA sequences of CMM specimens with known taxonomic identity. The development of CMMRSL aims to elevate local testing standards through providing versatile testing services. The reference DNA sequences and information of the selected DNA barcodes are reported in the form of a fact sheet. The fact sheet of individual CMM will include an assembly of reference DNA sequences of a DNA barcode from all specimens in FASTA format, which provides a quick way for manipulation of the reference DNA sequences for data analysis (refer to the "General Notice" of CMMRSL for details).

Use of Reference DNA Sequences

All reference DNA sequences in the CMMRSL are generated, according to GCMTI in-house validated methods, from voucher specimens and CMMs authenticated by experts. Therefore they can be used in three aspects: (1) species identification, (2) revealing phylogenetic relationships of organisms and (3) providing genetic information

for the development of new test methods. The selected DNA barcodes have been adopted in Pharmacopoeias and are generally accepted by the scientific community for the purpose of species identification.

(1) Species identification

Species identification is performed by comparing the generated sequences from the samples to the reference sequences, known as sequence comparison. In sequence comparison, the degrees of similarity among DNA sequences will be observed. The more the common nucleotides between the sequences, the higher their similarity. The compared sequences with higher similarity are in higher chance to contain functional, structural and/or evolutionary relationships.

(2) Phylogenetics analysis

The evolution history and relationship among organisms are important information for species assignment. In phylogenetics, such relationships are inferred by scientists through inspecting the heritable traits of organisms, such as DNA sequences, morphology, behavioural features. After arranging organisms into nested groups based on their traits, a diagram showing the hypothetical relationships can be built, which is known as a phylogenetic tree. A phylogenetic tree constructed by single- or multiple-DNA barcodes is useful for the differentiation of closely related species with high generic similarity.

(3) Development of test methods

Polymorphic sites between CMM and related species can also be spotted through sequence comparison. Species-specific identification methods, such as specific-polymerase chain reaction (specific-PCR) and polymerase chain reaction-restriction fragment length polymorphism (PCR-RFLP), are developed based on the polymorphic nucleotides among species. These methods allow identification of target CMM species or differentiation between genuine CMM and its common counterfeits, which are the fundamental basis for the development of specific-PCR methods. In comparison with DNA barcoding, specific-PCR methods are faster in processing and simpler in result interpretation, and thus offer a lower cost.

To achieve the above-mentioned use of reference DNA sequences, two common approaches for species identification and development of test methods are (1) sequence alignment and (2) comparison against public databases.

(1) Sequence Alignment

Sequence alignment is a method of comparing two or more DNA sequences to identify similarities. Pairwise sequence alignment aims to find the best way to align and arrange the characters of two query sequences. Multiple sequence alignment, similar to pairwise sequence alignment, analyses more than two sequences in each comparison. As DNA barcodes are mainly 300-600 bp in length, it is not practical to align them manually. Some programmes and software for sequence alignment are suggested in the following paragraphs.

Programmes for sequence alignment

Sequence alignment programme is a tool for pairwise and multiple sequence alignment. Clustal Omega, ClustalW, MUSCLE (MUltiple Sequence Comparison by Log-Expectation), MACSE (Multiple Alignment of Coding SEquences), T-Coffee (Tree-based Consistency Objective Function for Alignment Evaluation) are examples of programmes that perform multiple sequence alignment very well in practice. Clustal Omega and ClustalW (http://www.clustal.org/) are general-purpose multiple sequence alignment programmes for protein and DNA/RNA. Clustal Omega is the latest member in the Clustal family which can process larger alignment than the previous versions. The advantages of MUSCLE (https://www.drive5.com/muscle/) are its speed and alignment accuracy. These programmes are mostly applicable for Windows, Mac OS and Unix/Linux, and can be run in the command line mode.

Software package for sequence alignment

Molecular Evolutionary Genetics Analysis (MEGA) is a free sequence alignment software that allows comparative analysis of sequences. MEGA is equipped with ClustalW and MUSCLE programmes accessed by a graphical user interface. Users can input a FASTA file, which contains the sequences of tested samples and the reference DNA sequences, into MEGA and perform sequence alignment. MEGA displays the four types of nucleotides (A, G, C and T) in different colours and users can easily detect the variations from the alignment results if there are any mismatched nucleotides against the top sequence (normally the reference sequence as shown in Figure 1). Other than aligning sequences, MEGA allows phylogenetic tree building and genetic distance calculation.

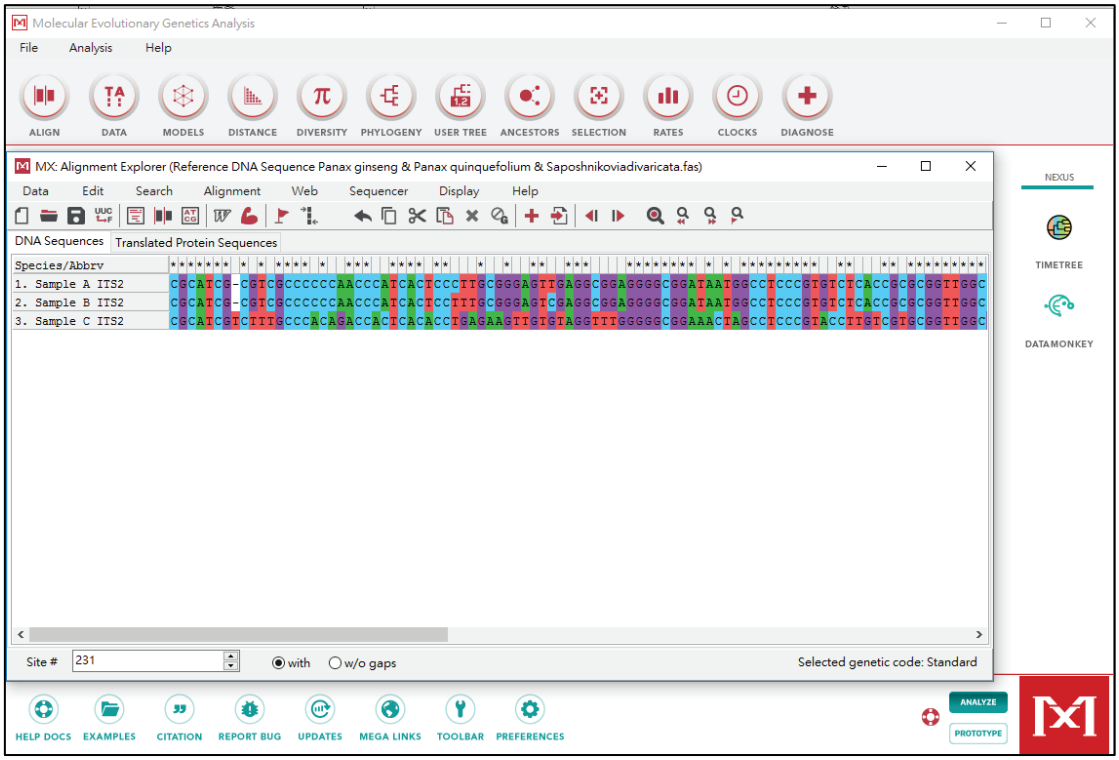

Figure 1. Example of sequence alignment using the sequence alignment software MEGA.

BioEdit and CodonCode Aligner are also popular choices for sequence alignment among researchers. BioEdit is a free and beginner-friendly software with various built-in editing functions, such as sorting of sequences, calculation of sequence percent identity and phylogenetic tree building. It has an instinctive user interface for importing, exporting, manipulating and inspecting the result of sequence alignment generated by using ClustalW. CodonCode Aligner is a commercial package strong in editing the order of nucleotides and inspecting the quality scores of raw sequence data. Users can perform sequence end clipping, sequence assembly, contig editing and mutation detection. Aligning sequences can be done by using the built-in MUSCLE, MACSE or Clustal Omega programmes.

Online platform for sequence alignment

Multiple sequence alignment programmes are also available on online platforms. The European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI) (https://www.ebi.ac.uk/) is an organisation that aims at helping researchers to share and analyse biological data. A variety of online tools are provided on its official website, such as Clustal Omega, MAFFT (Multiple Alignment using Fast Fourier Transform), MUSCLE, T-coffee, etc. The web form allows users to input multiple sequences by "copy and paste". Parameters like output sequence order and format can be adjusted to suit the needs and workflow of users.

NCBI Multiple Sequence Alignment Viewer (https://www.ncbi.nlm.nih.gov/projects/msaviewer/), managed by the National Center for Biotechnology Information (NCBI), is an online application that displays multiple sequence alignment graphically. If the nucleotides of the sample sequence are different from that of the reference DNA sequence (top sequence), the differences will be displayed in red colour (Figure 2), allowing users to detect the variations from the alignment result visually.

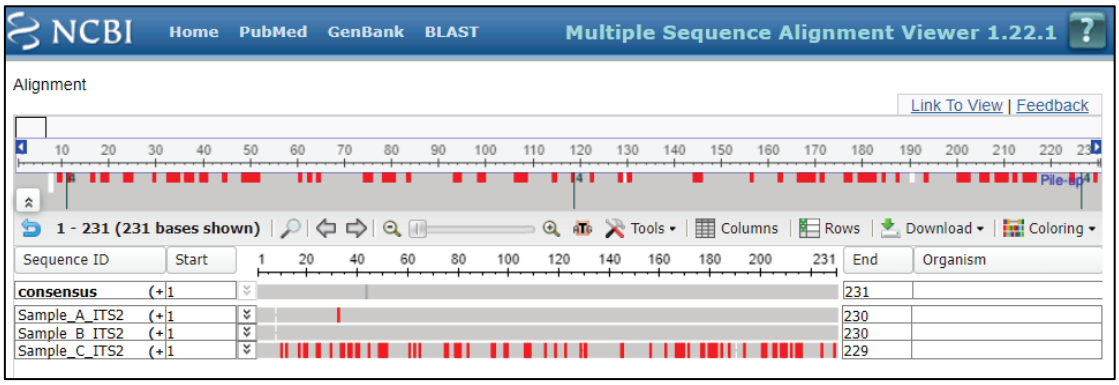

Figure 2. Example of sequence alignment using the online sequence alignment platform NCBI Multiple Sequence Alignment Viewer.

(2) Comparison of DNA Sequences using Public Databases

GenBank

NCBI GenBank is a comprehensive database managed by NCBI. It is one of the largest sequence databases and contains over 2.5 billion sequences collected from more than 500 000 species. Each sequence record in GenBank is assigned with a unique identifier called accession number. The sequence record comprises the source of organism, locus definition, authors, features of sequence, etc. NCBI accepts the submission of sequence by scientific communities around the world. It is worth noting that some submitted sequences come from biological materials that may not have been examined by experts (e.g. CMM commercial products). In general, nucleotide sequences generated from voucher specimens or with traceable specimens are a relatively reliable source of reference.

BLAST in GenBank

Basic Local Alignment Search Tool (BLAST) is a programme for performing sequence similarity search against databases. In GenBank, there are several versions of BLAST programmes designed specifically for searching sequences in a particular

database. For DNA barcoding of CMM, the most commonly used version is BLASTn which searches against the nucleotide database (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch). Users can input the query sequence and adjust the parameters in the web interface. BLAST then determines the statistical similarity between the query and nucleotide sequences available in NCBI GenBank (Figure 3).

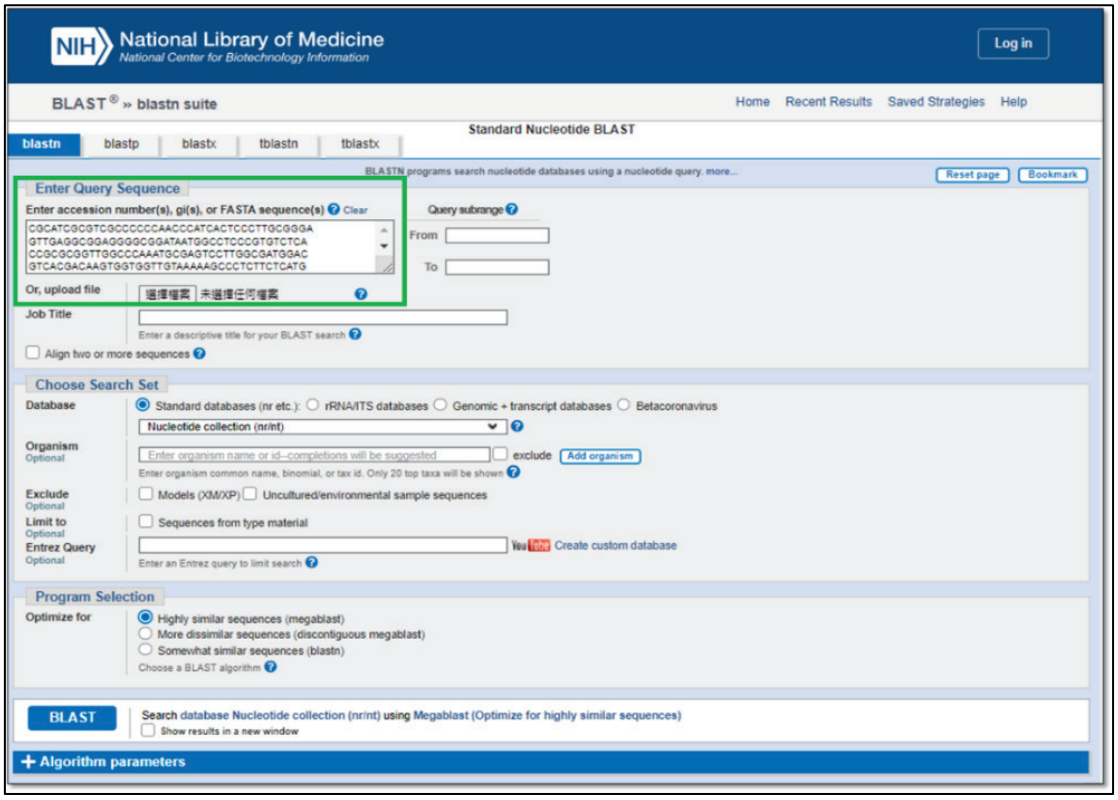

Figure 3. Example of uploading a sequence to the BLAST search platform.

The BLAST result will show a list of nucleotide sequences in NCBI GenBank that matches the query sequence with the best-matched one at the top of the list (Figure 4). To interpret the BLAST search result, there are two numbers that users should focus on: "percent of identity" and "query coverage". "Percent of identity" represents the percentage of similarity between the query and the subject nucleotide sequences in GenBank while "Query coverage" represents the percentage of the query sequence that overlaps the nucleotide sequence in GenBank.

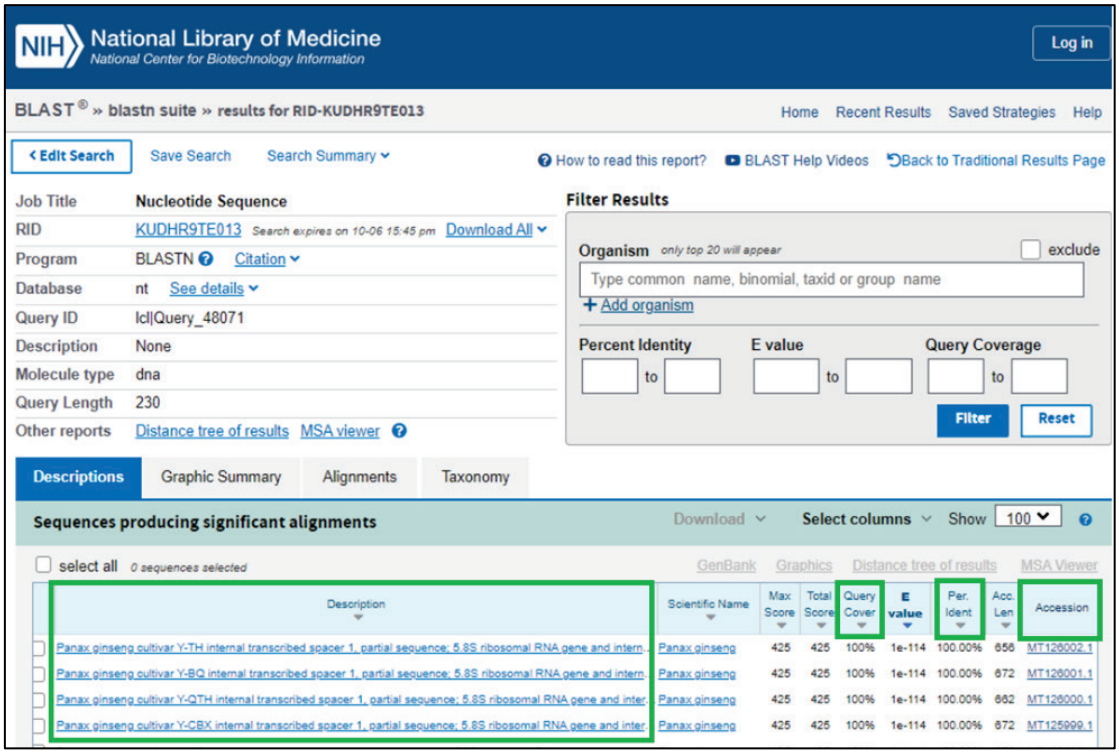

Figure 4. Example of the BLAST search result of a reference DNA sequence against nucleotide sequences available in NCBI GenBank.

Other public databases

Barcode of Life Data Systems (BOLD https://www.boldsystems.org/index.php/IDS_OpenIdEngine) is a widely applicable online platform that provides data storage and analysis of nucleotide sequences, especially for animals.

References

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.