Retrieval of Sequence(s) from the NCBI Nucleotide Database

CHAPTER 1

CS Mukhopadhyay and RK Choudhary
School of Animal Biotechnology, GADVASU, Ludhiana

1.1 INTRODUCTION

The NCBI nucleotide database (http://www.ncbi.nlm.nih.gov/nucleotide/) is an archive of gene, transcript, and fragments of genomic DNA sequences. It combines several online public repositories, including *GenBank* (the genetic sequence database of NIH), *RefSeq* (annotated, non-redundant reference sequence from genomic, transcript and protein), *TPA* (third-party annotated data on nucleotide sequences), and *PDB* (protein databank: a repository of 3D structures of proteins and nucleic acids). The International Nucleotide Sequence Database Collaboration (INSDC) maintains the liaison between the three major molecular data repositories – namely, NCBI, DDBJ, and EMBL – to share the nucleotide data present in any of those databanks.

A brief description of the NCBI databases has been given in Appendix A "NCBI Database: A Brief Account" at the end of this book.

1.2 COMPONENTS OF THE NCBI NUCLEOTIDE DATABASE

- **GenBank**: An annotated collection of all publicly available nucleotide and *in silico* translated protein sequences.
- **EST database**: Maintains expressed sequence tags (ESTs) and short, single-pass reads (the sequence-fragments/reads obtained by loading the reaction in a lane only once and, hence, obtained after analyzing the input sequence by the sequencer only once) from mRNA (cDNA).
- **GSS database**: A database of genome survey sequences (GSS), or short single-pass genomic sequences (TTS, Exon Trapped, BAC/YAC, etc.)

Basic Applied Bioinformatics, First Edition. Chandra Sekhar Mukhopadhyay, Ratan Kumar Choudhary and Mir Asif Iquebal.

1.3 OBJECTIVES

To search and download nucleotide sequences from NCBI Nucleotide database and save as a text file (*.txt). The sequence of interest for downloading could be complete or partial gene/mRNA/coding sequence, non-coding RNA (rRNA, tRNA), non-coding and repeat sequences (VNTR) in the genome, partial genomic DNA sequences, and so on.

1.4 PROCEDURE

1.4.1 Nucleotide sequence search

a. Open the NCBI nucleotide page: http://www.ncbi.nlm.nih.gov/nucleotide/

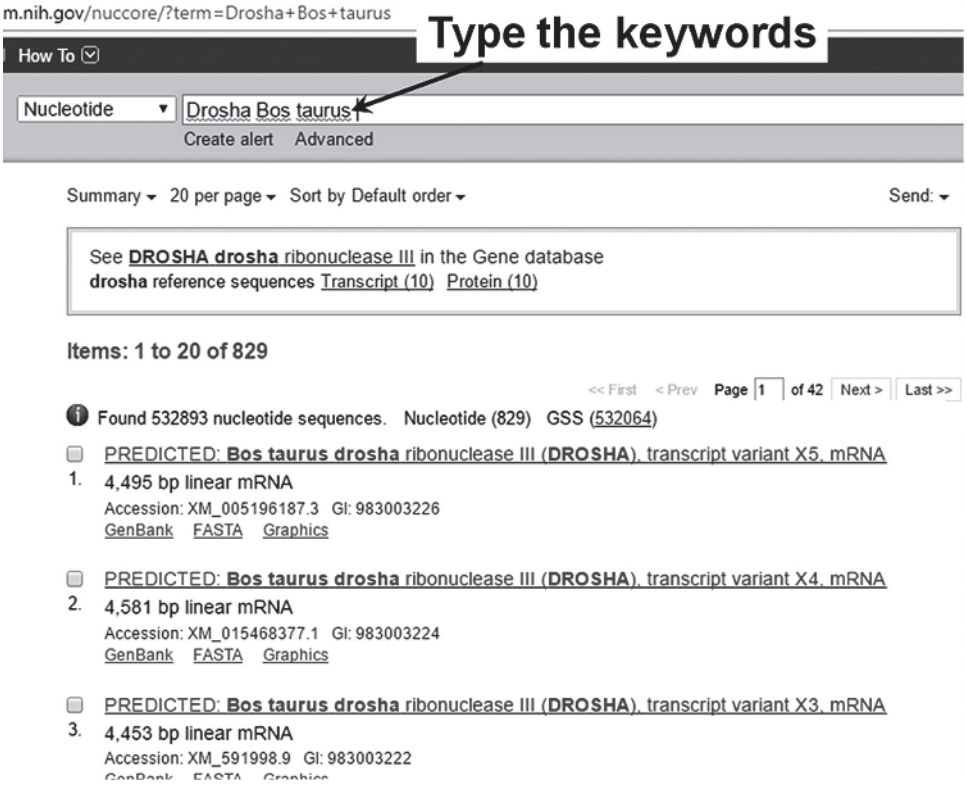

FIGURE 1.1 Main search window of NCBI Nucleotide page and list of hits for nucleotide sequences of taurine *Drosha* (gene/mRNA). (*See insert for colour representation of the figure.*)

b. Search the target sequence by providing the name of the gene and keywords – say, for example, the *Drosha* gene sequence in taurine cattle (*Bos taurus*) (Figure 1.1). Thus, the keywords are: "Drosha Bos taurus" (type your keywords without quotes, or else the quotes will instruct the search engine to find out the exact phrase within quotes, which ultimately limits your search). Then click on the "Search" button.

c. The nucleotide sequence(s) can also be searched by specifying the *accession number*(s), separated by a space (or comma). Please note that from September 2016 onwards, NCBI has phased out the sequence gi numbers. The accession numbers are *unique codes* assigned as an *identifier* to each nucleotide sequence in the database.

1.4.2 Downloading the selected sequences

a. Now, for example, select the first three sequences (depending on your requirement) by checking the small checkboxes on the left-hand side of each of the sequences.
b. Click on the "Send to" button, located at the top-right side of the page (Figure 1.2). Choose the destination of the selected sequences (to a *.txt file or to the clipboard for copying and pasting to a separate file, or collection in your NCBI account). Register yourself to NCBI and get your account-Id and password. Select the sequence format (Summary, GenBank, FASTA, etc.), the items per page and mode of sorting the selected sequences from the drop-down menus before downloading in a text file.
c. Finally, click on the "Create File" button to download the nuccore_result.txt file (default name) (see Figure 1.2, below). Open the file to obtain the sequences in the specified format and order.

FIGURE 1.2 Click on the "Send to" button to download and save (in a text file) the first three Drosha mRNA sequences in "Summary" format. (*See insert for colour representation of the figure.*)

1.5 SOME USEFUL NUCLEOTIDE SEQUENCE DATABASES OF NCBI

One can search other NCBI databases that archive nucleotide sequences:

a. species-wise or chromosome assembly search (WGS or other assembly of chromosome or genome, likeBos_taurus_UMD_3.1.1);
b. clone (clones associated with genomics, cDNA and cell-based libraries, viz. BTDAEX-80 K11, HWYUBAC-1-028-04-H12);
c. dbGaP (interaction of genotype and phenotype, viz. phs000287.v4.p1Cardiovascular Health Study (CHS) Cohort);
d. dbVar (large-scale genomic variation, nsv836042, nsv836041 etc), SNP, etc., among many other databases. The process of downloading the data as a text file is the same.

1.5.1 Modifying the search with the "Limits" option (currently not available)

The user can *narrow down the search* by using the parameters available after clicking on the hyperlink "Limits". However, NCBI has removed this option nowadays. The available options are:

a. Published in the last (specify the available days or mention date range).
b. Modified in the last (specify the available days or mention your own date range).
c. Search *Field Tags* (different fields of GenBank flat file Accession, Author, Bioproject, etc.).
d. Segmented sequences (master of set or part of set).
e. Source database (RefSeq or GenBank or EMBL or DDBJ or PDB).
f. Molecule (Genomic DNA/RNA, mRNA, rRNA or cRNA).
g. Gene Location (Genomic DNA/RNA or Mitochondrion, Chloroplast or "any" of the above types).
h. Exclude (STSs and/or working draft and/or TPA and/or patents).

1.5.2 Modifying the search with "Nucleotide Advanced Search Builder"

Click on the hyperlinked word "Advanced" just below the text box. The new page enables you to build your search settings.

Please note that the *search builder* enables us to specify the keywords according to their type (i.e., accession, assembly, author, journal and so on); in turn, this instructs the search engine to pinpoint the keywords from the database, depending on its feature or type.

Let us take our previous example: "Drosha Bos taurus". In the search builder, click on the drop-down menu (shown as "All Fields") and select "*Gene Name*" and type "Drosha". The role of "Show index list" is discussed in the next paragraph. Next, click on the drop-down list of the second-row field and select "*Organism*" and then

type "Bos taurus" (without quotes). If you have more keywords, then add more rows accordingly, and select the specific field before typing the keyword(s).

The "Show index list" button will show the list of indexes from which you can specify your index. To move further along the index, you can use "Previous 200" or "Next 200" options. The "+" and "–" symbols beside each of the text boxes allow you to add (new) or delete the corresponding text boxes.

Please note that we can also do the advanced search without using the "Advanced Search Builder". Type the keywords in the form as given below: "term [field] OPERATOR term [field]". So, in the case of our current example, it will be:

Drosha[Gene name] AND Bos taurus[Organism]

1.6 QUESTIONS

1. Download the following sequences in FASTA format, and save these sequences in a single text file: NCBI Nucleotide Accession numbers are AB909393.1; AB909392.1; AB909391.1; AB906338.1; AB906337.1; KF773864.1; AB898237.1
2. Suppose you need to download the "Drosha" full-length sequence of taurine cattle. How will you proceed with the "Advanced" option of the Nucleotide search?
3. Download the following sequences in NCBI (full length) sequences and save them in a text file: NCBI Nucleotide Accession numbers: KF021228.1; KC831578.1; KC822646.1; KC758965.1; KC758964.1; KC424594.1; KC424593.1
4. How will you search and save the bubaline SRY coding sequence in NCBI, Nucleotide and check only the full-length cds or mRNA, and then save the FASTA sequences in a text file?
5. What are the differences between Genome Survey Sequence (GSS) and Nucleotide sequences in the NCBI database? In your search result, if you get both types of sequences, which one will you download to use as a template for primer designing?