UiT Norges fiskerihøgskole

Practical 6. GENETIC DATA AND DATABASES

- **Learning outcome from this exercise**

1. To know what different genetic databases are there and how to access them (NCBI, BOLD and KEGG)
2. To know how to download the nucleotide/protein data from databases.
3. What is the format of nucleotide or protein data?
4. How to identify unknown sequences through similarity search tools.

Genetic database is one or more sets of genetic data stored together with a software to retrieve, supplement and extract information from them. In this exercise, we will focus on the genetic database called NCBI (National Center for Biotechnology Information). In the database, data are either stored in physical hard drives in a location or many locations.

We will quickly browse couple of databases such as KEGG and NCBI. We will dive deep into NCBI in a while.

Let´s start using the NCBI browser to retrieve information. First go to the website: https://www.ncbi.nlm.nih.gov. Look at the home page. Home page has lot of information, and it is comprehensive: submitting, downloading, different databases (nucleotide, protein, etc…), different software, literature information. We will spend some time on browsing.

- **Downloading the sequence information**

The NCBI is very user-friendly database and data search platform. All the basic operations are meant for traditional biologists. So, anyone with minimum biology knowledge can go in there and retrieve information. The only condition is that as a user you need to know what you want, which depends on the objective of your question. We will retrieve COI (cytochrome c oxidase subunit I) sequences of some salmonids (family Salmonidae), or you can choose the species you want.

On top of the web page there is a search bar to search information about the species of your interest. If you select *all the databases* in the search menu and type the scientific name (common name do work sometime). You will be directed to a page where the information about your species of interest is available all possible databases. This is one way to find desired information for your species.

Another way is by choosing specific database line one below.

Use *genus species or family or any other taxonomic* classification information along with the gene name (COI) to find the sequences. Now you choose *nucleotide* instead of *all the database* example: search term could be *salmonidae coi*

1. Go through the result table and choose a few sequences that belong to different species of the Salmonidae family
2. Download by clicking right corner “send to” > “complete record” > “file” > “fasta” > “download”.

Now sequences are downloaded in the *fasta* format to your local system (your laptop). Generally, it is downloaded to the folder Downloads.

Now Open downloaded file using a text editor. In windows there is default text editor called notepad. But it is not a best text editor. Rather you can download notepad++ (https://notepad-plus-plus.org/downloads/). macOS users can download BBedit (https://www.barebones.com/products/bbedit/).
Now you have COI sequences belong to different genus of salmonids/your species or family of interest. If you have time just, try different genes. Example: Severe acute respiratory syndrome coronavirus 2 spike protein

Using blast (basic local alignment search tool) (online version)

The BLAST is the most famous and powerful application of NCBI. As the name suggests, it is a match searching tool. Locate the tab in NCBI webpage where the blast application is shown.
Please look into various blast modules. We will go one by one and discuss.
This is how whole thing works: Just feed an “unknown” nucleotide or protein sequence to the blast search box and do blast’ing (important: choose right blast module, based on search molecule type and your expectation). It will give you possible hit to your sequence with statistical support (E-VALUE and query coverage). It finds the similarities between sequences you provided (which is query), and sequences stored in the database (which is subject). Blast also calculates the statistical significance of that comparison (E-value, which is like p-value, tells the probability of the query matching to the subject being random). Let’s try a few unknown sequences in blast and try to find the possible identity of those unknown sequences. The file with unknown sequences is in your virtual machines in this address: /home/adminfsk2053/data/unknown1.txt
You get a table as output. Now we see what’s there on that table.

Tips for search sequences:

1. Choose right blast module based on the type of sequence input and importantly, your objective

2. Try to set the parameters if needed. Generally, default settings work better in most of the situations. So, leave them alone if you don’t have any compelling reason to change them or if you know what you are doing.

One disadvantage with online BLASTing is waiting time. Longer the sequence (s) and larger the database you are comparing it against, more time it takes to finish the search operation. The genbank has doubled in size in the last year, this means slower searches and slower database updates. It also depends on the time of the day you are using the Blast. NCBI servers located in USA. Hence it is obliged to serve national users first. So, what is the solution if you have lot of sequences?

Because of above reasons it is convenient to use a local blast when you have lot of sequence to identify. We will use a local blast application in our system. Using local blast, we can get more customised result table too. We can construct our own blast commands according to our desired filtering parameters.

2.4 Linux based command line BLAST (Standalone blast)

What do you need,

1. NCBI BLAST+ command line tool installed
2. Data base to search against
3. Query sequence(s)

You don’t do this in this practical. I have done installation part for you:

Install NCBI BLAST+:

In linux OS: apt-get ncbi-blast+ or conda install

It will download the precompiled latest version of blast+ tools to your system.

Or

https://www.ncbi.nlm.nih.gov/books/NBK52640/

go to this link to install in Linux.

How to prepare a database

Don’t do this either (if you want to try, go to blast manual on how to make blast database). I made this for you. Database is the one you blast your query against.

I made a database specific for teleost.

Go to FTP server of the NCBI and look for premade blast database.

https://ftp.ncbi.nlm.nih.gov/blast/db/

or

download nucleotide or protein (nr) database to your system and make database locally.

The command for creating your own database looks like this:

```
makeblastdb -in your_db_sequences.fasta -input_type fasta -dbtype nucl
```

Tip:

1. Basis to download premade database is to filter the sequences for desired family or genus or so on. You may not need full database. Full database is very huge in disk size.
2. Filtering is not easy if you download just the sequences in fasta instead of premade database.

Just type blastn in the command line. You get some warning related to parameters of blastn. If this happened, you are good to launch.

Now we know how the command line system works in a Linux OS. We will utilise that knowledge here. To run a local blast. We can a similar output as we got from the web-based blast. But the command line version lacks some graphical features.

We will write the blastn script together (NB! there are spaces between the different parameters)

```
blastn -query name of the query file -db
path_to_database/nt_teleost_16112020 -max_target_seqs 1 -outfmt 6 -out
results2.txt -num_threads 1 -evalue 0.001
```

Also try blastp.

Default column names (look into online help or terminal help:

```
-outfmt "6 qseqid sseqid pident length mismatch gapopen qstart qend sstart
send evalue bitscore"
```

Now use this parameter in the same command.

```
-outfmt "6 qseqid qlen qaccver sseqid slen saccver sacc stitle salltitles
length pident nident mismatch gapopen qstart qend sstart send evalue bitscore
qcovs qcovhsp"
```

BLASTn tabular output format 6

Column headers:

```
qseqid sseqid pident length mismatch gapopen qstart qend sstart send
evalue bitscore
```

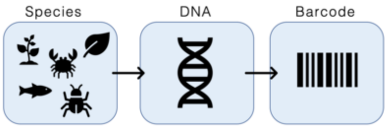
1. **qseqid** query or source (e.g., gene) sequence id
2. **sseqid** subject or target (e.g., reference genome) sequence id
3. **pident** percentage of identical matches
4. **length** alignment length (sequence overlap)
5. **mismatch** number of mismatches
6. **gapopen** number of gap openings
7. **qstart** start of alignment in query
8. **qend** end of alignment in query
9. **sstart** start of alignment in subject
10. **send** end of alignment in subject
11. **evalue** expect value
12. **bitscore** bit score

*Complete description of blast output fields

https://www.metagenomics.wiki/tools/blast/blastn-output-format-6

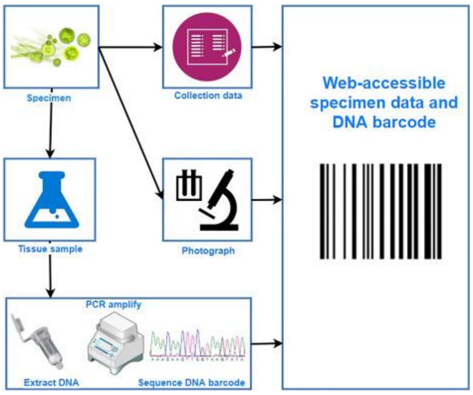
Field application of Blast and blast like tools

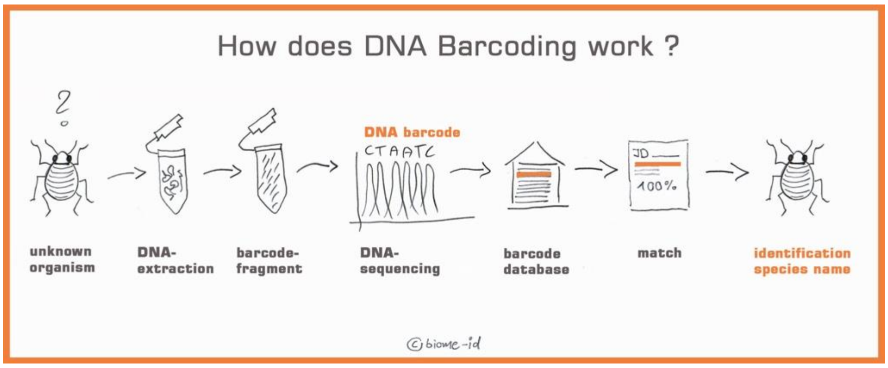
COI is used as a genetic marker to barcode the species. Barcode is the same thing you see on a packet or box in the shop.



Let’s say you have DNA from an unknown sample. You want to find out what species this DNA belongs. Simple way is to amplify COI region, sequence it and do a database search. You “may” find the answer. Then submit the sequence to appropriate database and that piece sequence will remain as barcode for that species for ever.

Food safety and quality of the food are of major concern. They play an important role in public opinion especially when food alterations or food adulterations get media publicity. There is an increasing demand for the improve food quality by identifying the commercial frauds. DNA barcoding is one of the few means to identify fish/meat sold in retail both due to insufficient labeling requirements or rampant mislabeling of the product or willful mixing of meat from other animals. However, success of the DNA barcode depends on the presence of highly curated sequence database. BOLDSYSTEMS (https://www.boldsystems.org) is one of them. COI is the standard genetic marker used in DNA barcoding technique.





COI is the standard genetic marker used in DNA barcoding.

Exploration of BOLD system (Barcode of Life Data System)

Go to https://www.boldsystems.org/index.php

What is the largest group of animals having barcode information and from where they are reported the most? How much of these records come from Norway?

To get that information go to taxonomy tab and start looking into different groups animals and the numbers of available records are in brackets. Press the links associated with group and you will get lot of metadata about the group.

Application of BOLD systems

The sequences (database search fasta) come from a publication where the authors studied whether sushi restaurant goers in NYC eat any IUCN red listed (https://www.iucnredlist.org). Tuna species, as many times fish species (not only tuna) are either mislabelled purposefully or lack of taxonomical knowledge. We will identify the species of Tuna and deliberate the status of these species in IUCN red list (https://www.iucnredlist.org) using BOLD system and if time permits, we can also try them in NCBI web.

We need to select 'identification' tab to do blast like search (https://www.boldsystems.org/index.php/IDS_OpenIdEngine). Copy and paste (one by one or whole sequence set, it needs login account)). Choose species level barcode information as database. Discuss the result from BOLD identification analysis.