Introduction to Genomics, transcriptomics and structural databases

Dan MacLean

11/16/22

Table of contents

Preface

The primary purpose of this course is to introduce you to databases that contain genome sequence data, transcriptome data and sequence structural data, and some of the tools that you can use to interact with them. Most of the material in this course will be web-based but some work will be via command lines.

Three weeks of computational analysis can save you an hour of database searches

All scientists know well the value of the library and how getting up to speed with the literature and current knowledge will prevent you from wasting time in the lab testing (or rather re-testing) ideas that have already been tested. An analogous rule in bioinformatics is that databases contain data and results that have already been performed such that knowing about them and how to access them can save you great deal of effort. Knowing that you can just download some results can help you progress and develop your own unique analyses much more quickly. The trick therefore is simply to know what databases exist, what is in them and how to extract what you need from them.

In this short course we'll look at primary and secondary genomics and transcriptomic databases as well as looking at tools that can make downloading from them very straightforward. We'll also look at some databases and tools for examining protein structures.

Prerequisites

Knowledge prerequisites

The materials in this book assume that you already know how to make use of command line tools.

Software prerequisites

You have two options to run the software for this course, you can run it in the pre-built Colab Document which you can access by clicking this Link to Google Colab

You will need a Google account to use it. If you don't yet have one, or don't want to use your personal account for this perhaps you could sign up using your `@tsl.ac.uk` address.

Alternatively, you can install the tools yourself on your local machine, which means you need to install the following stuff:

- SRA Toolkit
- SPAdes
- prokka
- busco
- kallisto

Most can be installed using bioconda

1 Primary and Secondary Databases

1.1 About this chapter

1. Questions:

- What is the difference between a primary and secondary database?
- What are some examples of primary and secondary databases?
- How can I interact with them to find what I need?
- What is metadata?

2. Objectives:

- Understand the role of a primary and secondary database
- Know where to access and how to find training resources provided by the databases
- Understand the importance of clear structured metadata

3. Keypoints:

- Primary databases are repositories of experimental data
- Secondary databases are repositories of results based on the primary data
- Metadata is what makes data findable and useable

1.2 Primary databases

In many fields, databases can be categorized as either primary or secondary. The main difference is actually quite simple, *Primary* databases hold data that come from experiments or direct data gathering exercises and their contents are fixed. A primary database is often called an archive or a database of scientific record.

1.3 Examples of primary databases

1.3.1 ENA

ENA - https://www.ebi.ac.uk/ena/browser/home

The European Nucleotide Archive (ENA) is a comprehensive record of the world's nucleotide sequencing information, covering raw sequencing data and sequence assemblies. Access to it is provided through a genome browser, search forms and tools, and an API - an interface that allows you to write programs that extract data. Like many databases, user training is a very important part of the mission of the database, so they provide a very comprehensive manual and how to's at https://ena-docs.readthedocs.io/en/latest/

1.3.2 GenBank

GenBank - https://www.ncbi.nlm.nih.gov/genbank

GenBank is another distinct genetic sequence database with annotated sequences and it aims to hold *all* publicly available DNA sequences. Access to it is provided by the Entrez system which can find identifiers (ID's) and BLAST which can search sequence similarity. There are also FTP (bulk download) and programatic acess via the NCBI e-utilities system. The provided training materials for GenBank are very dense, and take some reading https://www.ncbi.nlm.nih.gov/guide/training-tutorials/. The quick start guides are often a better place to start https://www.ncbi.nlm.nih.gov/nuccore/

1.3.3 GEO

GEO - https://www.ncbi.nlm.nih.gov/geo/

The Gene Expression Omnibus (GEO) is a functional genomics repository of microarray and RNAseq data that accepts raw microarray, sequence and curated gene expression profiles. The normal web, FTP and programmatic access routes are provided and help is provided best through the FAQ https://www.ncbi.nlm.nih.gov/geo/info/faq.html.

1.3.4 BioStudies

BioStudies - https://www.ebi.ac.uk/biostudies/about

BioStudies is a relatively new database that consolidates some other databases. BioStudies database holds descriptions of biological studies and links to data from these studies in other databases. Here you can find transcriptome data (from ArrayExpress), literature data (from PubMedCentral), image data (from BioImages) and more. The help describes searching https:

//www.ebi.ac.uk/biostudies/help and again this database provides a web search interface, FTP and Aspera for bulk download and an API for programmatic access.

1.3.5 Protein Data Bank

Protein Data Bank https://www.ebi.ac.uk/pdbe/node/1

PDBe aims to provide a resource of high quality macromolecular structures with integration of function, taxonomy and sequence. The databank provides excellent tutorials at https://www.ebi.ac.uk/pdbe/training/tutorials in particular the course at (https://www.ebi.ac.uk/training-beta/online/courses/biomacromolecular-structures)[https://www.ebi.ac.uk/training-beta/online/courses/bioma structures] and the YouTube channel https://www.youtube.com/user/ProteinDataBank

1.4 Secondary databases

Secondary databases contain the results of analyses, often using the contents of primary databases. They will typically use information and raw data from various sources to create a new resource of information. Often highly curated, they can employ combinations of tools and analyses to present a synthesis of lower data types. These are the sort of databases molecular biologists are most used to using and many provide derived information of varying degrees of confidence about objects within the primary sequences, things like genes in genomes.

1.5 Examples of secondary databases

1.5.1 Ensembl

http://www.ensembl.org/index.html

Ensembl is a genome browser based database that contains information on comparative genomics, sequence variation, transcriptional regulation alongside annotated genes. While ensembl.org is focussed on vertebrates, there are sister databases for fungi, plants and bacteria. The help files are large and have many video lessons http://www.ensembl.org/info/website/tutorials/index.html. The BioMart tool in particular is very useful for medium scale retrieval BioMart. The APIs provided can make whole genome and comparitive genomics analysis very streamlined.

1.5.2 UniProtKB

https://www.uniprot.org/help/uniprotkb

UniProt Knowledgebase is a database of functional information on proteins with a focus on accurate and consistent annotations. The database is divided into two parts, one with manually-curated records with information from literature and curator analyses (Swiss-Prot), the other with computationally-analysed only records that are yet to be examined by a human (TrEMBL). A YouTube channel of introductory videos is provided https://www.youtube.com/channel/UCkCR5RJZCZZoVTQzTYY92aw for high level information. A large selection of on-demand training is also available through EBI Training Portal, e.g https://www.ebi.ac.uk/training/events/guide-uniprot-students/

1.6 Metadata

Data, especially raw data is absolutely useless on its own. It always need some contextual information to allow us to make use of it. This contextual data is called `metadata` - data about the data. For example genome sequencing data would need information about the organism sequenced, the machine used to sequence, the tissue type and much more. Anything that would be useful needs to be stored along with the data. Without metadata it would not possible to find data and select the correct samples for a given analysis of it.

In order for metadata to be useful we need to capture the same metadata for all data samples. This can mean not only standardising the questions we ask of, but also the words we allow to be used in answering the questions. This can seem a bit restrictive at first, but is actually essential. Consider two labs working on NLR proteins, one group is in the habit of calling them `NBS-LRR` proteins, the other `NLR`. These two terms would be completely different to a computer and anyone within the field not yet aware of all the jargon. Some searches and analyses and computer programs would completely fail to link these up.

As a result there are a wide range of accepted international standards on the metadata and this is why databases require users to annotate their data according to these standard. The standards themselves are constantly developing as new technologies and tools emerge, agreeing on data standards is an important and valuable job within research.

1.6.1 Minimum Information Standards

Many data types have a 'minimum information standard', a set of guidelines for reporting data on them. These ensure that the data of a certain type are all reported on in the same consistent way. Microarrays (MIAME), RNASeq (MINSEQE), and proteomics have data standards specific to them. More are listed at FAIRSHARE.

1.6.2 Controlled Vocabularies

To maintain a consistent set of terms for describing biological concepts and objects, we use controlled vocabularies, taxonomies and ontologies, among others. Controlled vocabularies restrict the words a human may use to describe something, a list of countries on a website would be an example of this. There are other ways to structure the controlled vocabulary. Taxonomies are ways of classifying things, often hierarchically from a more generic to a more specific group. This structure can add flexibility over the simple list, the biological taxonomy is an example of this. Ontologies represent the background knowledge of a domain, the objects are represented as terms and links between them are created to represent the relationships, such that we can analyse the ontology. The typical example is the Gene Ontology, which you can learn more about here https://www.ebi.ac.uk/training/online/course/goa-and-quickgo-quick-tour

1.6.3 Conclusion

> **Note**
>
> ```
> * Databases are important stores of data for the scientific record
> * Databases enable new analyses.
> * The metadata used to describe the data in databases is what makes the data useful.
> ```

2 Genomics Database and Genome Assembly

2.1 About this chapter

1. Questions
    - How can I use the ENA and NCBI to download sequence reads?
    - What is the overall pipeline for assembling genomic DNA reads?
    - How can I annotate an assembled prokaryotic genome?
2. Objectives
    - Download some reads identified in ENA using efficient tools
    - Run the tools needed to build a first draft genome assembly

2.2 Interactive tutorial

The work in this course is provided as a combination of interactive tutorial and command-line tasks for you to run on your own machine. All the materials you need are referenced in the interactive tutorial.

> **Note**
>
> - Run the interactive tutorial here <a href="https://tsl-bioinformatics.shinyapps.io/genome_dbs/"
>
> If you want to use the pre-prepared Colab instance, you can find that here: <a href="https://colab.research.google.com/drive/1LJffmZ4yuzEjvtWp6Sj5tSibyTKJb_b1?usp=sharing"

3 Transcriptomics Databases and Transcript Abundance

3.1 About this chapter

1. Questions
    - How can I use the ENA and NCBI to download sequence reads for transcripts?
    - What is the overall pipeline for estimating abundance of transcripts?
    - How can I annotate an assembled prokaryotic genome?
2. Objectives
    - Download some reads identified in ENA using efficient tools
    - Run the tools needed to estimate transcript abundances

3.2 Interactive tutorial

The work in this course is provided as a combination of interactive tutorial and command-line tasks for you to run on your own machine. All the materials you need are referenced in the interactive tutorial.

> **Note**
>
> - Run the interactive tutorial here <a href="https://tsl-bioinformatics.shinyapps.io/transcriptome_dbs/"
>
> If you want to use the pre-prepared Colab instance, you can find that here: <a href="https://colab.research.google.com/drive/1LJffmZ4yuzEjvtWp6Sj5tSibyTKJb_b1?usp=sharing"

4 Structural Databases

4.1 About this chapter

1. Questions:

- What is held in structural databases?
- How can I find the structure of my protein?
- Are there any solved structures like my protein?

2. Objectives:

- Understand what is held in a structural database
- Know how to get an approximate structure for a protein of interest

4.2 Interactive tutorial

The work in this course is provided as a combination of interactive tutorial and command-line tasks for you to run on your own machine. All the materials you need are referenced in the interactive tutorial.

> **Note**
>
> - Run the interactive tutorial here <a href="https://tsl-bioinformatics.shinyapps.io/structural_dbs/"