Tutorial: Bioinformatics Basics

Eric Holloway

DOI: 10.33014/issn.2640-5652.2.2.holloway.1

1 Introduction

Bioinformatics can appear to be a daunting field, since it combines the complex science of biology with the complex theory of computer science. However, the basics are surprisingly simple.

Essentially, bioinformatics is the discipline of analyzing symbol strings. The symbol strings represent DNA, RNA, and protein sequences, with each symbol representing a component of the molecule. The strings are analyzed by extracting the structure within the strings and comparing the relationships between the strings. The DNA and RNA strings are composed with four different symbols and protein strings have twenty different symbols.

2 Genetic Code and Sequence Translation

The letters of DNA and RNA strings are called nucleotides.

The DNA letters are G (guanine), A (adenine), T (thymine), C (cytosine).

The RNA letters are G, A, U, C. It is almost directly copied from DNA, except the T is turned into U (uracil).

The protein letters are A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y. See Figure 1 for their meanings.

Each protein letter is translated from three RNA letters. Each group of three RNA letters is called a **codon**. The translation table is known as the **genetic code**. The START and STOP symbols are used to mark where a **gene** begins and ends. Figure 1 shows the standard mapping between codons and amino acids.

A single biological DNA sequence is a molecule, and this molecule is called a **chromosome**.

3 Sequencing and Assembly

These symbol strings are digital abstractions, and show how fundamental the notion of information is to biology. The process of extracting these abstractions is called **sequencing**.

To extract genetic sequences, current technology must first break a long sequence into many little fragments, usually on the order of a few hundred nucleotides long or tens of proteins long. Once these fragments are digitized, then they must be reassembled back into the full genome. This assembly process is very computationally intensive and error prone. Sometimes the full genome cannot be assembled, and the best that can be done is to construct larger fragments.

The fragments are known as **reads**. A collection of fragments is known as a **run**. When the fragments are pieced together into a longer fragment, this is known as a **contig**. If the full genome is put together, this is known as an **assembly**. If there are gaps, but the contigs are lined up against a known genome, this is called a **scaffold**.

4 Accessing Data

The data for each step in the process is usually stored in databases, so that experiments and assembly can be reproduced by other scientists. The databases described in this article are the databases maintained by the National Center for Biotechnology Information, since these are what the author is most familiar with.

All items of data in the database have a unique ID number known as an **accession**. Once you know the accession for a piece of data, you can use NCBI's tools to download the data. The main tool you can to search for an accession and download the data is NCBI's website: `www.ncbi.nlm.nih.gov/`.

If you want to access data from the command line for an automated workflow, there are a couple command line tools available.

Amino Acid	Symbol	Codon
Alanine	A	GCU, GCC, GCA, GCG
Asparagine/Aspartic Acid	B	AAU, AAC, GAU, GAC
Cysteine	C	UGU, UGC
Aspartic Acid	D	GAU, GAC
Glutamic Acid	E	GAA, GAG
Phenylalanine	F	UUU, UUC
Glycine	G	GGU, GGC, GGA, GGG
Histidine	H	CAU, CAC
Isoleucine	I	AUU, AUC, AUA
Lysine	K	AAA, AAG
Leucine	L	CUU, CUC, CUA, CUG, UUA, UUG
Methionine	M	AUG
Asparagine	N	AAU, AAC
Proline	P	CCU, CCC, CCA, CCG
Glutamine	Q	CAA, CAG
Arginine	R	CGU, CGC, CGA, CGG, AGA, AGG
Serine	S	UCU, UCC, UCA, UCG, AGU, AGC
Threonine	T	ACU, ACC, ACA, ACG
Valine	V	GUU, GUC, GUA, GUG
Tryptophan	W	UGG
Tyrosine	Y	UAU, UAC
Glutamine/Glutamic Acid	Z	CAA, CAG, GAA, GAG
Start		AUG
Stop		UAA, UGA, UAG

Figure 1: The Standard Codon Table

4.1 File Format

The main file format used for DNA and protein sequences is the FASTA format. The format is pretty straightforward. There are metadata lines and lines that contain the genetic data. The metadata lines start with the > character, and can be used to break up multiple reads in a run file. The rest of the lines contain DNA or protein letters. Sometimes, there are extra letters that stand in for multiple possible DNA letters. Most often this is the letter N, which means any DNA letter can go in that particular spot. It signifies a sequencing error. See the next section for an example of this file format.

4.2 Sequence Read Archive

The initial read data is stored in NCBI's Sequence Read Archive (SRA): `www.ncbi.nlm.nih.gov/sra`. To access this data there is the sra-toolkit: `github.com/ncbi/sra-tools`.

Here in an example session of using the toolkit to then download a run.

```
$ fastq-dump --fasta -Z DRR001793 | head -n 8
>DRR001793.1 FC30UF2AAXX:6:1:10:1453 length=35
ACCAGCTATCACCGAGTTTNNNTATCCTTTCACCC
>DRR001793.2 FC30UF2AAXX:6:1:10:1692 length=35
TATTATTTAACTGATAATTANNCTAGATATATTAT
>DRR001793.3 FC30UF2AAXX:6:1:10:1896 length=35
AGACCAATTCATTAATTTTTTNTTTATTATACTAT
>DRR001793.4 FC30UF2AAXX:6:1:10:1845 length=35
AAAGGCAGAGTACATTAAGACNATAGATTTAGTTT
```

You can see from the above the downloaded file consists of lines of DNA spaced with identifier lines. Each of these lines of DNA is a read.

4.3 Eutils

The assembled genomes and proteins can be accessed with the Eutils web API. The documentation for the API is available at `https://www.ncbi.nlm.nih.gov/books/NBK25500/`. There is a Python library available for using Eutils in Python available at `https://pypi.org/project/eutils/`. Additionally, there are also command line tools for Eutils, and a great tutorial on them is available at `https://www.ncbi.nlm.nih.gov/books/NBK179288/`.

The general workflow for Eutils is a query tool returns an XML report, which is then fed to a data access tool to retrieve the record.

Eutils is the most versatile tool, since it can access data from all the different NCBI databases of processed genetic data.

4.4 Datasets

Finally, NCBI has recently released a beta version of a more user friendly search and data access tool called Datasets: `ncbi.nlm.nih.gov/datasets`. The tool's main feature is the ability to download many pieces of data altogether in a single 'bag'. The idea is to make accessing data like a grocery shopping store, where you fill your cart with data items, and then check them all out at once. There is also a command line tool that can be used for programmatic workflows.

5 Finding Things With BLAST

BLAST (basic local alignment search tool) is the Google of bioinformatics, hosted at the National Center for Biotechnology Information. You can enter DNA or an amino acid sequence into the tool, and BLAST will search the NCBI databases for matches.

Since DNA is mutated—with swapped, deleted, and added nucleotides—there is never an exact match. BLAST searches heuristically using a variant of the edit distance with a substitution matrix, representing the probability one nucleotide mutates into another nucleotide.

There are four main variants of BLAST, representing the four possible combinations of DNA and amino acid searches.

1. blastn: enter DNA sequence to find DNA sequence
2. blastx: enter DNA sequence to find amino acid sequence
3. tblastn: enter amino acid sequence to find DNA sequence
4. blastp: enter amino acid sequence to find amino acid sequence

5.1 BLAST Interface

The basic interface is straightforward, and the web interface at `https://blast.ncbi.nlm.nih.gov/Blast.cgi` is

self descriptive. You enter a reference to a sequence, such as the accession number, or paste the sequence itself. A specific subrange of the sequence can be queried.

To constrain the search, different database can be searched, ranging from databases of carefully currated genomes to user submission databases. Searches can be within specific organisms, or exclude specific organisms.

The search sensitivity is tuned by setting the expect threshold, which will return only matches that do not exceed the expected number of random matches.

The results come back with metadata expressing in which genetic data the match was found, an 'e' score based on the logarithm of probability of matching, and what percentage of the returned sequence matches the query. The text of the sequence itself is also returned, along with any associated publications.

5.2 Example: Finding the TOP2A Protein

The simplest way to find a known protein is through the NCBI search bar, available from the NCBI front page at `ncbi.nlm.nih.gov`. Typing in TOP2A will bring up an information card, with links to the known orthologs in other species as well as a prepopulated BLAST search. You can then run the BLAST search to retrieve matches. Note, you can only exclude results at the taxonomic level, so if there is another protein that is similar, it will show up in the search as well.

6 Summary

As you can see, while the user interface for accessing data may be a bit complicated, and the naming convention for various aspects of genetic data obscure, the fundamentals are straightforward. And, once you know that genetic data is composed into DNA, RNA, and protein sequences, you just need to download the data to start analyzing it from the comfort of your living room. The NCBI website is a good pathway to becoming a bioinformatics scientist, all without the need for an expensive lab and fancy equipment. All you need is a computer with an internet connection and the mind between your ears.