Sequence Formats and Sequence Database Searches

Gloria Rendon
SC11 Education
June, 2011
A sequence is the primary structure of a biological molecule. It is a chain of residues that form a precise linear order.
Other types of structures

Secondary structure: is the general 3D form of local segments (alpha helices, beta sheets) and their interactions mediated by hydrogen bonds.
Other types of structures

Tertiary structure: is the 3D structure as defined by its atomic coordinates
Other types of structures

Quaternary structure: many proteins are actually assembles of more than one polypeptide chain. The assemblage of those protein subunits are known as the quaternary structure of the protein.
Plain Sequence Format

DNA’s alphabet: \{ \text{A, C, T, G} \}
RNA’s alphabet: \{ \text{A, C, U, G} \}

Plain sequence format. It is an ORDERED string of characters where each character represents not only the residue but also its location.

The characters must belong to any of these three alphabets depending on the type of sequence, DNA, RNA or protein.
**Fasta Sequence Format**

> gi|455952|gb|S67398.1| 18S rRNA {3' region} [Lemna minor, entire plant, rRNA Partial, 111 nt]
CTCCTACCGATTGAATGGTCCGGTGAAGCGCTCGGATCGCGGCGACGAGGGCGGTCCCCCGCCCGCGACGTCGCGAGAAGTCCGTTGAACCTTATCATTTAGAGGAAGGAG

**Fasta format.** For each sequence there is a description line and the contents (or body) of the sequence.

The description line is always the the first line of the sequence. The description line begins with the > symbol followed by the description itself; all in one line.

The description must include the identifier or name of the sequence. But there is no universal convention for what else to specify and in which order.

The body of the sequence can occupy as many lines as necessary. Like in the plain format; this segment is an ORDERED string of characters where each character represents not only the residue but also its location.
There exist many other sequence formats which arrange information about the sequence in a more verbose manner.

Here we see the same sequence of the previous page but in GenBank format.
Sequence Databases

• The GOOD news:
  • There are many Repositories of sequences.
  • Some of them are publicly available and can be freely obtained (no fee).
  • Megasites like NCBI (at NIH), UniProt (at EBI), KEGG (Japan) provide storage of and access to their sequence databases. They also provide additional services like curation, regular updates, software tools for mining the databases, etc.

• The BAD news:
  • There is great proliferation of formats and conventions that are site and/or database specific
  • There is little effort at each site to provide cross-referencing with the rest of the world
  • Be prepared to spend a great deal of time dealing with format conversions and cross-referencing of databases
Searching Sequence Databases

First a word of warning!

Most publicly available sequence databases provide web servers for users to search and access their content.

Most of them, if not all, are still a long way from being user-friendly during search, retrieval and download activities.

These sites do not yet have the capability of conducting meaningful searches based on natural language queries. In other words, do not expect the search engine to behave like Google.
Searching Sequence Databases

These three scenarios will illustrate the most common approaches to performing searches for sequences in databases. They are ordered by level of difficulty.

Scenario 1: you know the database and the sequence identifier; for example, the database is GenBank, the sequence identifier is FJ985763

Scenario 2: you know the name of the gene or protein and the organism; for example, 16s ribosomal RNA in E. Coli

Scenario 3: you have a segment of the sequence but nothing else; for example, a segment from an assembly after shotgun sequencing on a soil sample.
For scenarios 1 and 2, we will be using search tools that try to match your query to the information in the description line of the sequence. These tools assume your query will be specified in plain English but with control vocabulary.

For Scenario 3 we will be using search tools that try to match your query to the body of the sequence. These tools, like Blast, perform the match by doing sequence alignment.
Searching Sequence Databases
Scenario 1 – (seq-id, db) known

Let us suppose that you were extremely lucky and read a paper that included a full reference of the sequence it was reporting on. That is, it provided the database name and the sequence identifier.

For example:

The supplementary material includes a table with sequence ids on page 31. http://www.nature.com/nature/journal/v459/n7250/extref/nature08182-s1.pdf

Let us search and retrieve any of the sequences in that table from the NCBI url
Scenario 1 – (seq-id,db) known

1. Extract the sequence id and database from the paper
2. Go to the web server of the database it belongs to
3. In the search box type the sequence id and run the search
4. Examine the results and download the sequence
5. Done
Scenario 1 – (seq-id, db) known

1. Extract the sequence id and DB from the paper
2. Go to the web server of the database it belongs to
3. In the search box type the sequence id and run the search
4. Examine the results and download the sequence
5. Done
Scenario 1 – (seq-id, db) known

1. Extract the sequence id and DB from the paper
2. Go to the web server of the database it belongs to
3. In the search box type the sequence id and run the search
4. Examine the results and download the sequence
5. Done
Scenario 1 – (seq-id, db) known

1. Extract the sequence id and DB from the paper
2. Go to the web server of the database it belongs to
3. In the search box type the sequence id and run the search
4. Examine the results and download the sequence
5. Done
Scenario 1 – (seq-id, db) known

1. Extract the sequence id and DB from the paper
2. Go to the web server of the database it belongs to
3. In the search box type the sequence id and run the search
4. Examine the results and download the sequence
5. Done

In the Results page, click here to retrieve the result found in this particular database of nucleotide sequences
Scenario 1 – (seq-id, db) known

1. Extract the sequence id and DB from the paper
2. Go to the web server of the database it belongs to
3. In the search box type the sequence id and run the search
4. Examine the results and download the sequence
5. Done

This is the result displayed in GenBank format. To change to FASTA format click on the `fasta` link.
Scenario 1 – (seq-id, db) known

1. Extract the sequence id and DB from the paper
2. Go to the web server of the database it belongs to
3. In the search box type the sequence id and run the search
4. Examine the results and download the sequence
5. Done

This is the result in FASTA format. To download it to your computer click on Send.
Scenario 1 – (seq-id, db) known

1. Extract the sequence id and DB from the paper
2. Go to the web server of the database it belongs to
3. In the search box type the sequence id and run the search
4. Examine the results and download the sequence
5. Done

You can also access the Taxonomy page of the Organism that contains this gene by clicking this link.
Scenario 1 – (seq-id, db) known

1. Extract the sequence id and DB from the paper
2. Go to the web server of the database it belongs to
3. In the search box type the sequence id and run the search
4. Examine the results and download the sequence
5. Done

From the box select **File** as destination; **FASTA** format
Then click on **Create File**
Searching Sequence Databases

 Scenario 2 – (gene name) known

Let us suppose that you read a paper written some time ago and gives the name of the gene; but it does not give any additional information like database it is found in and sequence id.

For example:
http://www.pnas.org/content/74/11/5088.full.pdf+html

We can choose a likely database that can contain the gene and formulate a query with as much information about the gene as we can gather from the paper. In this case, the database is GenBank at NCBI and the query could be:

16s/18S ribosomal RNA of each one the species listed on the table of p 5089
Scenario 2 – (gene name) known

1. Extract the gene name from the paper
2. Go to the web server of a megasite like NCBI, EBI, etc.
3. In the search box type a query with the gene name and run the search
4. Select the MOST likely hit(s) from the list of results.
5. Examine the results and download the sequence
6. Done
Scenario 2 – (gene name) known

1. Extract the gene name from the paper
2. Go to the web server of a megasite like NCBI, EBI, etc.
3. In the search box type a query with the gene name and run the search
4. Select the MOST likely hit(s) from the list of results.
5. Examine the results and download the sequence
6. Done

<table>
<thead>
<tr>
<th>Table 1. Association coefficients ($S_{AB}$) between</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>1. Saccharomyces cerevisiae, 18S</td>
</tr>
<tr>
<td>2. Lemna minor, 18S</td>
</tr>
<tr>
<td>3. L cell, 18S</td>
</tr>
<tr>
<td>4. Escherichia coli</td>
</tr>
<tr>
<td>5. Chlorobium vibrioforme</td>
</tr>
<tr>
<td>6. Bacillus firmus</td>
</tr>
<tr>
<td>7. Corynebacterium diphtheriae</td>
</tr>
<tr>
<td>8. Aphanocapsa 6714</td>
</tr>
<tr>
<td>9. Chloroplast (Lemna)</td>
</tr>
<tr>
<td>10. Methanobacterium thermoautotrophicum</td>
</tr>
<tr>
<td>11. M. ruminantium strain M-1</td>
</tr>
<tr>
<td>12. Methanobacterium sp., Cariaco isolate JR-1</td>
</tr>
<tr>
<td>13. Methanosarcina barkeri</td>
</tr>
</tbody>
</table>

The 16S (18S) ribosomal RNA from the organisms (organelles) list.
Scenario 2 – (gene name) known

1. Extract the gene name from the paper
2. Go to the web server of a megasite like NCBI, EBI, etc.
3. In the search box type a query with the gene name and run the search
4. Select the MOST likely hit(s) from the list of results.
5. Examine the results and download the sequence
6. Done

Type the information in the search box

The advantage of going to a megasite, like NCBI, is that their repository has good coverage and is updated regularly.
Scenario 2 – (gene name) known

1. Extract the gene name from the paper
2. Go to the web server of a megasite like NCBI, EBI, etc.
3. In the search box type a query with the gene name and run the search
4. Select the MOST likely hit(s) from the list of results.
5. Examine the results and download the sequence
6. Done

Alternatively, you can go to specialized databases like this one for ribosomal RNA sequences only

http://www.arb-silva.de/search/
Scenario 2 – (gene name) known

1. Extract the gene name from the paper
2. Go to the web server of a megasite like NCBI, EBI, etc.
3. In the search box type a query with the gene name and run the search
4. Select the MOST likely hit(s) from the list of results.
5. Examine the results and download the sequence
6. Done

This is a partial view of the results page. As you can see from the figures (representing number of hits); it could easily turn into "looking for a needle in a haystack".
Scenario 2 – (gene name) known

1. Extract the gene name from the paper
2. Go to the web server of a megasite like NCBI, EBI, etc.
3. In the search box type a query with the gene name and run the search
4. Select the MOST likely hit(s) from the list of results.
5. Examine the results and download the sequence
6. Done

In the Nucleotide results page we will examine the first hit. To do that, you can click on it directly.
If you want to examine several entries, select them first and then click on any of the selected items to retrieve the corresponding record
Scenario 2 – (gene name) known

1. Extract the gene name from the paper
2. Go to the web server of a megasite like NCBI, EBI, etc.
3. In the search box type a query with the gene name and run the search
4. Select the MOST likely hit(s) from the list of results.
5. Examine the results and download the sequence
6. Done

Notice that the description is occupying two lines rather than a single line for this type of format. Make sure that the file you download contains the sequence(s) in the proper format.
Searching Sequence Databases

Scenario 3 – (partial seq) known

Let us suppose that you read a paper that presents only a segment of the sequence; but does not provide clear details about the source (database and sequence id).

For example:

Figure 2 shows the (sequence) segment of interest.

We have the choice of using a) the GENE name or b) the SEGMENT of the sequence to find the entire sequence of the gene.

a) It will be similar to what we did in Scenario 2
b) It requires us to use BLAST to search for the complete sequence by matching the segment of the sequence to one contained in the database.
Scenario 3 – (partial seq) known

1. Extract the partial sequence from the paper
2. Go to a BLAST web server.
3. In the sequence box type the partial sequence
4. Select a database and run the search
5. Select the MOST likely hit(s) from the list of results.
6. Examine the results and download the sequence
7. Done
Scenario 3 – (partial seq) known

1. Extract the partial sequence from the paper
2. Go to a BLAST web server.
3. In the sequence box type the partial sequence
4. Select a database and run the search
5. Select the MOST likely hit(s) from the list of results.
6. Examine the results and download the sequence
7. Done

This is the partial sequence for the first gene 5-HT3A as it appears on Figure 2 of the paper.

NSGERVSFKITLLLLGYSVFLIIIVSDTLPATA
Scenario 3 – (partial seq) known

1. Extract the partial sequence from the paper
2. Go to a BLAST web server.
3. In the sequence box type the partial sequence
4. Select a database and run the search
5. Select the MOST likely hit(s) from the list of results.
6. Examine the results and download the sequence
7. Done

www.uniprot.org
Scenario 3 – (partial seq) known

1. Extract the partial sequence from the paper
2. Go to a BLAST web server.
3. In the sequence box type the partial sequence
4. Select a database and run the search
5. Select the MOST likely hit(s) from the list of results.
6. Examine the results and download the sequence
7. Done

1. Select the Blast tab first
2. Paste the partial sequence in the box
3. Leave default parameters unchanged
4. Then click on the Blast button
Scenario 3 – (partial seq) known

1. Extract the partial sequence from the paper
2. Go to a BLAST web server.
3. In the sequence box type the partial sequence
4. Select a database and run the search
5. Select the MOST likely hit(s) from the list of results.
6. Examine the results and download the sequence
7. Done

The most likely hit is NOT always the top hit on the list of results.

We will come back to this results page after we learn how to read global and local alignments, which are marked here by the green rectangles in columns 3 and 4 respectively.
Scenario 3 – (partial seq) known

1. Extract the partial sequence from the paper
2. Go to a BLAST web server.
3. In the sequence box type the partial sequence
4. Select a database and run the search
5. Select the MOST likely hit(s) from the list of results.
6. Examine the results and download the sequence
7. Done

1. Check that the identity is 100%
2. Check that it is the right gene
3. A perfect match (of the local alignment)
4. Retrieve the sequence by clicking on its Accession
Scenario 3 – (partial seq) known

1. Extract the partial sequence from the paper
2. Go to a BLAST web server.
3. In the sequence box type the partial sequence
4. Select a database and run the search
5. Select the MOST likely hit(s) from the list of results.
6. Examine the results and download the sequence
7. Done
Scenario 3 – (partial seq) known

From the previous page, click on fasta, written in yellow, to get to this page which displays the sequence in fasta format.

The name of the gene is 5HT3A_HUMAN
Its sequence id in this database (called UniProt) is P46098

You can download the sequence by saving this page to a text file.

You can also bookmark this url for future reference. We will revisit this page in the next section.
Additional Resources

- Sequence Formats
  http://emboss.sourceforge.net/docs/themes/SequenceFormats.html

- NCBI introductory video
  http://www.youtube.com/watch?v=X7s1zewqeLI

- UniProt introductory video
  http://www.youtube.com/watch?v=TCF3qWn7sil