RETRIEVING SEQUENCE INFORMATION

Nucleotide sequence databases

Database search

Sequence alignment and comparison
Biological sequence databases

Originally – just a storage place for sequences. Currently – the databases are bioinformatics work bench which provide many tools for retrieving, comparing and analyzing sequences.

There are many different types of sequence databases.

Global nucleotide sequence storage databases:
GenBank of NCBI (National Center for Biotechnology Information)
The European Molecular Biology Laboratory (EMBL) database
The DNA Data Bank of Japan (DDBJ)

Genome-centered databases
NCBI genomes
Ensembl Genome Browser
UCSC Genome Bioinformatics Site

Protein Databases
SwissProt
Web browsers for accessing genome data

Entrez Life Sciences Search Engine (US National Institutes of Health)

Ensembl Genome Browser (European Bioinformatics Institute)

UCSC Genome Bioinformatics Site (University of California at Santa Cruz)
Entrez


A service of the US National Center for Biotechnology Information (NCBI)
29 core search pages interconnecting different types of information
Among the most valuable tools are nucleotide, EST and protein sequence databases, the Human Genome Resources, Microbial Genome Resources, Complete Genomes tools, etc.
GenBank and PubMed

The NIH also maintains GenBank and PubMed databases (both linked to Entrez)

GenBank is a genetic sequence database that provides an annotated collection of all genomic and cDNA sequences that are in the public domain. In August 2009, there were 106,533,156,756 bases in 108,431,692 sequence records in the traditional GenBank divisions.

PubMed is a literature search engine providing access to all published medically related articles, abstracts and journals.
Ensemble

Ensembl is a joined project between European Bioinformatics Institute (EMBO) and the Sanger Institute in Cambridge. Its major focus is the genomes of animals and other eukaryotes. The annotation of most of the genomes in Ensembl database have been processed automatically though annotation pipelines. Genomes of five species (human, mouse, zebrafish, pig and dog) have been carefully annotated by manual curation.
Ensemble

A nice feature is customizable home page that keeps track of your searches even if done from different computers. Ensemble has many tools for extracting and downloading user-specific aspects of the data in various formats suitable for high-end bioinformatics analysis.
UCSC Genome Bioinformatics Site

A highly sophisticated genome browser allowing for visualization of many genome features: mapping and sequencing, phenotype and disease association, genes and gene predictions, transcript evidence, gene expression data, comparative genomics, sequence variations.
Any of the genome databases can be used to extract simple sequence information.

Depending of the question asked by a researcher, one or another database may be more suitable to specific needs.
What to keep in mind when working with sequence databases

NCBI (The National Center for Biotechnology Information) accepts all submitted sequences and does not check whether the sequence is accurate or not.

Annotations of most genes and genomes are done by researchers and except for specific ‘Reference Sequences’ (RefSeqs) are not curated by Genbank experts.

As a result, the structure of the same gene reported by different labs can be different!

Furthermore:

All genomes are full of SNPs and other polymorphisms.

Gene finding algorithms are not perfect, especially when it comes to predicting splice sites.

Alternative splicing can lead to different transcripts (different versions) of the same gene.

The implication is: there is often no single correct sequence for a given gene.
GenBank

Extracting a segment of a sequence which contains the gene of interest.

Task: find the sequence of the *Yersinia pestis* gene *lacZ*.

Point your browser to the NCBI main site: http://www.ncbi.nlm.nih.gov

select database
(“nucleotides” for GenBank)

search criteria

click the Go button
GenBank

In the results page, find the link to the entry of interest. The list may have many pages.
The link you are looking for, might not be the first entry.

A complete genome of a *Yersinia pestis* strain. Should have the gene you are looking for.

the genomic database

a standard Genbank entry

essentially the same info but slightly different format
A complete genome page contains lots of info and thousands of genes.

Use the ‘find’ function of your browser to find “lacZ”.

GenBank
GenBank

read info about your gene

click either ‘gene’ or ‘CDS’ link to go to the gene sequence
GenBank

Scroll to the bottom of the file to see the nucleotide and amino acid sequence of the gene.

Note, that this gene is encoded in a complementary strand relative to the one shown in the database. Therefore, the sequence shows the template strand.
Choose the FASTA format at the top of the entry page to get a ‘pure’ sequence that you can cut and paste into another application.
Extracting sequence of a human gene

How to extract sequence of a human gene:

select ‘nucleotides’ for GenBank access

In the search window, type the name of the gene, identifier, or GenBank accession number
Extracting sequence of a gene

Among several pages of matching entries find the one you want to examine and click on the accession number.
Extracting sequence of a gene

**LOCUS**: length, genomic/mRNA, date of submission
**DEFINITION**: species, common gene name
**ACCESSION**: GenBank accession number
**REFERENCE**: authors, publication (if any), history

**FEATURE TABLE**
- Coding sequence (CDS) coordinates
- The encoded predicted protein
- Miscellaneous features (exon boundaries, identified protein domains, mutations, etc.). Features can be associated with links to other databases: protein domain database (PFAM), Mendelion inheritance database (OMIM), etc.
Extracting sequence of a gene

The sequence of the gene is shown at the bottom. The default format is blocks of 10 nucleotides, 60 bases per row. At the top of the page, one can choose the format in which sequence is displayed. Sequence can be simply copied form the web page and imported into any of the bioinformatics applications or directly exported into some of them.
Homologous sequences

If sequences of two proteins are similar, they most likely are derived from the same ancestor and therefore, have similar structures and similar biological functions.

Science by analogy:
If something is true for one such sequence it is probably true also for the other.

Studying a protein (a gene) in the lab takes years, searching a database takes seconds.

Proteins (genes) that have the same ancestor are called **homologous**. Homologous proteins (genes) usually have similar sequences, structures and functions. A general rule: two proteins are likely homologous if their sequences are at least 25% identical. DNA sequences are likely homologous if they are at least 70% identical.

Do not confuse homology and similarity. Homology is a binary parameter (homologous vs. non-homologous) and reflects evolutionary relationship between the sequences. Similarity is a quantitative measure – two sequences can be more similar or less similar. Similarity does not necessarily reflect evolutionary relationship between sequences.
Sequence comparison

A pair-wise comparison of sequences is a very common task in bioinformatics and in genomics.

By aligning two sequences we are testing the hypothesis that the sequences are homologous, meaning that each pair of aligned residues are descendants of a common ancestral residue.

If two sequences have some similarity simply by chance – this does not have any biological significance. But if two sequences are homologous, by comparing them we can gain information about their structure, function and evolution.

The most common purpose of sequence comparison in bioinformatics is identifying homologous sequences.
Sequence comparison

Sequence comparison is not trivial.

Consider two alignments of two sequences:

Alignment 1                                      Alignment 2

Sequence 1                                      ACGCTGA
Sequence 2                                      A--CTGT

Which alignment is better?

The frequency of common evolutionary events determines the alignment parameters

- **Identity**
- **Substitution**
- **Deletion**
- **Insertion**

The frequency of substitutions is higher than the frequency of indels (insertions or deletions)
Sequence comparison

Computational alignment of two sequences is based on assigning a score to each possible alignment.

The score is composed of bonuses (for a match) and penalties for a mismatch or indel.

Let us assign bonus of +3 points for a match, a penalty of -1 for a mismatch and a penalty of -5 for an indel.

Align two sequences: ACGCTTGA and ACCCGTTA

ACGCTTGA
| | | | | | Score: 3 + 3 − 1 + 3 − 1 + 3 − 1 + 3 = + 12
ACCCGTTA

ACGC−TTGA
| | | | | | Score: 3 + 3 − 1 + 3 − 5 + 3 + 3 − 5 + 3 = + 7
ACCCGTT−A

AC--GCTTGA
| | | | | | Score: 3 + 3 − 5 − 5 + 3 − 1 + 3 − 1 − 5 − 5 = - 10
ACCCGTTA--

The choice of parameters depends on the biological model. The choice of parameters defines the score of the alignment.
Searching sequence databases for the entries with the sequence similar to the sequence of your protein or gene is the most frequently used bioinformatics procedure in genomic sciences.

The most frequently used tool for that is **BLAST – Basic Local Alignment and Search Tool**.
(Altschul et al., J.Mol.Biol. 1990)

BLAST is the most popular bioinformatics data-mining tool.
BLAST is a tool that searches large sequence databases returning sequences that have regions of similarity to a query sequence provided by the user.

BLAST finds isolated regions in sequence pairs that have high levels of similarity.

BLAST concatenates all the sequences in the database into one continuous VEEEEEEEEEEERY long sequence and then looks in this sequence for the matches to segments of the query sequence.
BLAST

BLAST comes in many different flavors. The choice of the BLAST protocol depends on your sequence and the question you ask.

For protein searches:

blastp – compares a protein sequence with a protein database (‘p’ – for protein)

tblastn – compares a protein sequence with a ‘translated’ nucleotide database (‘t’ – for translated)

Choice of BLAST:

I want to find out something about my protein: use blastp. Find proteins similar to yours and what is known about them.

I want to discover genes in other organisms that code for a protein similar to mine: use tblastn to compare your protein to the nucleotide sequences translated in 6 reading frames.
BLAST

There are several servers which make BLAST accessible world-wide to the researchers.

Server at NCBI (hosted in the US):

Highly pre-configured for the most common (basic) searches. A great place to start if you do not ask any very specific question about your sequence.

Server at EMBnet-ExPASy (hosted in Europe)
http://www.ch.embnet.org/software/bBLAST.html

Gives many choices to configure your search. Very useful if you ask specific bioinformatics questions.

Both servers are freely accessible to the academic researchers!
Running a blastp search

Q: Are there any protein similar to the hamster nucleolin?

Approach: find a protein similar to hamster nucleolin protein in the database Swiss-Prot which contains a wealth of information about all the studied proteins.

Point the browser to

choose the type of blast you want to run
(protein blast, in this case)
Put either your protein accession number or cut and paste its sequence. ‘Your sequence’ (retrieved via its accession number or written in the window) is called the query sequence.

Choose the database in which you want to search.

If needed, limit the search to a specific organism.

Click the BLAST button.
An intermediate screen will appear for few seconds to few minutes while the search is in progress.

Then the results page opens. The results page may have lots of information and many pages. Always examine it before you choose to print!

The results page has three important areas: a graphic display, a hit list, and the alignments.
A graphic display shows which segment of your query is similar to the sequences in the database. The similarity bars are color-coded to reflect the degree of similarity:
- **red** – the most similar
- **magenta** – still very good
- **green** – intermediate
- **blue** – twilight zone
- **black** – match, but a really bad one.

Some matches do not extend across the entire sequence: they may represent conserved domains in different, even unrelated proteins. Thus, nucleolins, like many other proteins, contain an RNA binding domain.
### blastp

#### The hit list

<table>
<thead>
<tr>
<th>Description</th>
<th>Sequence Accession Number</th>
<th>Name</th>
<th>Bit Score</th>
<th>E-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>The E-value (the expectation value). Statistical significance of the match to the hit: how often a match like this would be found if there is no real match in the database. Matches with the E-value below 0.001 (e-3) is worth exploring.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The sequence accession number and the name</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The bit score ('goodness' of the match). Ignore anything below 50.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genomic link</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The sequence accession number and the name (clickable!)

The percentage of identity (anything more than 25% is worth exploring)

The ‘Positives’: identity plus similar residues

The ‘Gaps’ – the non-aligned residues

The coordinates in the query sequence and in the hit (subject) sequence

top line – the query
middle line – consensus
bottom line – the hit
One you choose which hit sequence from the found list you want to work with, you can click on the name of that sequence.

The browser will bring you to the database entry of that sequence.

From there you can retrieve some primary info (origin, name, length).

You can find the relevant publications or take the accession number and look for the info about the protein in other databases.
Blasting DNA sequence

Two major modes of searching nucleotide sequence databases are

blastn – searches nucleotide sequence databases using nucleotide sequence queries

blastx – searches protein databases using a translated nucleotide query

Choice of BLAST:

I want to find gene regulatory regions similar to mine: use blastn. Find nucleotide sequences similar to yours in other genomes.

I want to find out whether my sequence corresponds to a protein gene: use blastx. Find known proteins similar to the one encoded in your sequence.