Bioinformatics Resources at a Glance

A Note about FASTA Format

There are MANY free bioinformatics tools available online. Bioinformaticists have developed a standard format for nucleotide and protein sequences that allows them to be read by a wide range of programs. This format is called FASTA format, and each nucleotide or amino acid is represented using a single letter. The first line of a FASTA is the comment line, identified with either the greater than symbol '>'. This line identifies the sequence and includes the accession number from NCBI, Genbank or another repository. The remaining lines contain the sequence, in lines of 80 or 120 characters per line. Many bioinformatics tools require that the sequence be terminated with two slashes '//' to indicate the end of a sequence. These may have to be added manually, as they are not always included in FASTA sequences online.

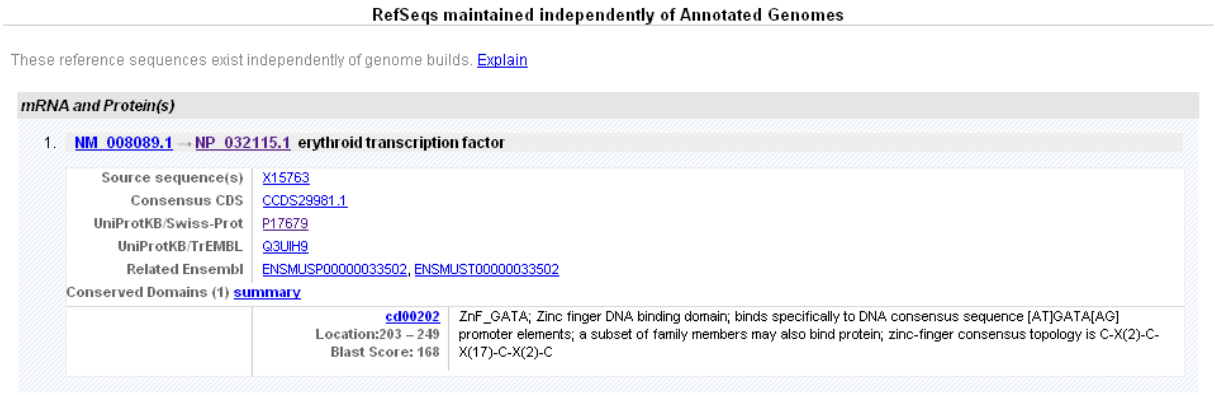
If you locate a nucleotide or protein sequence, and it is not in FASTA format, you can easily convert it to FASTA in a word processing program. Often the sequences contain numbers at the beginning of each line, as well as spaces between the numbers and the sequence. I simply run a series of 'find and replace' operations, replacing each of the digits 0 through 9 and a space with nothing. (In Word on a PC, open the find search box, enter what you are searching for in the find box, click on the replace tab, don't put anything in the replace field, and click 'replace all'.) You can then wrap the text by clicking 'delete' at the end of each line. If you have a long sequence, you'll discover that it is easier to start at the bottom and go up the sequence. I develop a rhythm – delete, up arrow, delete, up arrow, etc. (If you start at the top, you have to add an extra step: delete, down arrow, end, delete, down arrow, end…) When putting a file into FASTA format, add the comment string AFTER you've edited your sequence – or you'll lose any digits in the accession number!

I typically save nucleotide and protein sequences for a gene as separate Word documents within a single folder on my computer. This allows easy access for any of the manipulations I may do with the sequences.

Obtaining Nucleotide and Protein Sequences

1. Go to NCBI (http://www.ncbi.nlm.nih.gov/) – this is a repository of nucleotide and protein sequences, with MANY associated resources.
2. Input protein name and select 'all databases'
3. This will show the listing of all the resources in the NCBI related to your protein. Though the resources may appear overwhelming, it is worth your time and effort to do a little bit of exploring on this page just to see what is available. So many resources, so little time…

4. Click on ‘nucleotide’ or ‘protein’
5. Often there are numerous sequences; NCBI will frequently list the most likely candidates at the top of the screen. Since there are variations among species, be sure to select the protein from the appropriate species.
6. Click on the link to the protein you want to explore. This will bring up resource page in the NCBI – this is a GREAT place to get links to journal articles related to your protein, and oftentimes as you explore the site and follow the links, you’ll find a list of mutants and their impact, or even other proteins your protein associates with.
7. For nucleotide sequences, there is usually a gene map that depicts introns and exons, with an accession number on the left side of the map. If you click on the accession number and follow the links, you’ll typically get the WHOLE clone that includes your gene…but other genes as well. Depending on what you plan to do with the sequence, you may want to select the sequence in various forms:
    a. The clone will show the gene in context of other nearby genes on the chromosome. Though you won’t use the WHOLE clone, if you intend to create an activity that explores the regulatory sequences (promoters, for instance), you’ll need this information. This is often referred to as a ‘genomic’ sequence. You’ll need this sequence to show introns as well.
    b. An mRNA sequence contains the protein coding sequence (with introns removed), as well as 5’ regulatory sequences and 3’ sequences through to the polyadenylation site. If you want to map the introns, you can easily overlay the mRNA sequence on the genomic sequence. (When the files are in FASTA format, it is easy to locate the start of the mRNA on the genomic sequence by searching for 5-10 nucleotides from the mRNA. You can begin to annotate the genomic sequence by highlighting the mRNA sequence.
8. Be sure to read the accompanying description for the sequences you find. Is the sequence complete, or just a part of the gene? Are there any mutations in the gene? Often the notes will identify introns.
9. If you want to start with mRNA and protein sequences, scroll down towards the bottom of the screen. Often you’ll see something that looks like this:



10. This reference sequence includes two accession numbers. The one beginning with NM is the mRNA sequence; the one beginning with NP is the protein sequence. If you plan to look at both nucleotide and protein sequences, it is important that they come from the same gene. This reference sequence map ensures that the two sequences correspond.

Once You Have Your Sequence(s)…

What do you want to do with your sequence(s)? There are many tools and resources available online. This is just a brief overview of some of the things you can do.

1. Look for similar sequences – perhaps you'd like to see if your protein (or something similar) is found in other species. Or you want to compare the protein from different species to identify conserved regions (often important in function). If you have the nucleotide and/or protein sequence, you can do a BLAST search. This will identify similar sequences in the database. Note that, due to the degeneracy of the genetic code, you may find protein sequences that are highly conserved…but that there is significant variation in the nucleotide sequence.
2. Align two sequences – to compare their similarity. You may know two sequences are similar from the literature – or you may discover a sequence similar to your protein from a BLAST search. Sequence alignments can be done with either nucleotide or protein sequences. You may align two sequences, or, if you have multiple sequences, you may develop a phylogenetic
3. Annotate a gene – identifying promoters, introns, polyadenylation sequences. There are two ways to approach this task. You should start with a genomic sequence. If you have an mRNA sequence, you can easily highlight the processed mRNA on the genomic map. There may be gaps in the mRNA that are expressed in the genome – these are introns. If you wish to identify the 3 reading frames and highlight the correct one, you would TRANSLATE the mRNA sequence. It helps to have the protein sequence to identify the correct reading frame for your map.

BLAST Search: http://blast.ncbi.nlm.nih.gov/Blast.cgi

Search using:	Search databases:	Use tool:
Nucleotide sequence	Nucleotide database	Nucleotide blast
Protein sequence	Protein database	Protein blast
Nucleotide sequence – translated to protein	Protein database	Blastx
Protein sequence	Translated nucleotide database	Tblastn
Nucleotide sequence – translated to protein	Translated nucleotide database	Tblastx

TRANSLATE

1. Go to http://bio.lundberg.gu.se/edu/translat.html
2. Cut and paste DNA sequence

Aligning sequences

1. Aligning protein sequences: www.expasy.ch/tools/sim-prot.html
2. Multiple sequence alignment (nucleotide or peptide) – slow server: http://bio.biomedicine.gu.se/edu/msf.html

Other Bioinformatics Tools (a few of many similar resources!)

1. http://www.expasy.ch/tools/ - scroll down to DNA -> Protein
2. http://bio.lundberg.gu.se/
3. http://molbiol-tools.ca/ - includes background information

Generating a Gene Map of a Protein

1. Go to NCBI http://www.ncbi.nlm.nih.gov/
    a. Search for protein in all databases
    b. Select nucleotide databases
    c. Locate gene map (will show introns and exons)
    d. Download DNA sequence (left click on black code above sequence map)
        i. Select FASTA – save in Word file
    e. Download mRNA sequence (left click on blue code to left of sequence map)
        i. Select FASTA – save in Word file
    f. Download protein sequence (left click on red code to left of sequence map)

        i. Select FASTA – save in Word file
2. Translating DNA Sequence
    a. Go to http://bio.lundberg.gu.se/edu/translat.html
    b. Cut and paste DNA sequence into the box…omit the header information, and make sure the end of the sequence includes // (this is the FASTA code that says 'I'm done') – It's okay to leave in the numbers and spaces.
    c. Keep all the settings as they are, and click 'Translate'
    d. Cut and paste the results into another Word document – this is your gene map, with the three forward reading frames
    e. If you want to have Mark generate a paper bioinformatics strip – after translating to get all 3 reading frames – translate each one separately:
        i. Use the back arrow on the web browser to get back to the translate page. Just change the settings:
            1. Translate entire sequence and select reading frame – select 1
            2. Output options: uncheck 'Map of DNA sequence'
        ii. Copy the results in a Word document called frame 1
        iii. Repeat i and ii for reading frames 2 and 3
3. Formatting data to generate info for bioinformatics strips
    a. DNA:
        i. Use find and replace iteratively to replace the numerals 1-0 and a space with NOTHING. This will remove all the numbers and spaces. Do a find and replace to change t to u if you want to have an RNA map.
        ii. Go to the bottom of the file, next to last row, at the right side of the row. Iteratively click 'Delete' followed by 'up arrow' until you get to the top of the file. This will wrap all the text into a long string.
    b. Protein:
        i. Paste the three frames into the same file (to avoid having to run the steps three times). Place them in order (1,2,3) and separate with a divider that is non-alphanumeric. DO NOT label until after completing the find and replace steps! Save early and often.
        ii. Use find and replace iteratively to replace the numerals 1-0 and a space with NOTHING. This will remove all the numbers and spaces.
        iii. Go to the bottom of the file, next to last row, at the right side of the row. Iteratively click 'Delete' followed by 'up arrow' until you get to the top of the file. This will wrap all the text into a long string.
        iv. Next, to get the spacing correct, use find and replace to change each amino acid to be followed by two spaces:
            1. Find 'A' replace 'A ' – A followed by 2 spaces

2. Do this for all amino acids and the stop codon: A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y and *
3. Label the three reading frames and save the file!