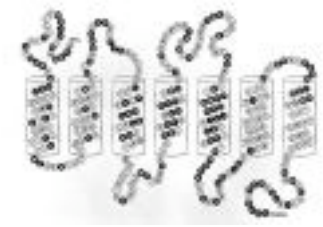
BLASTing "seven transmembrane" proteins and Preparing Sequence Files for Perl

A standard technique for any researcher interested in learning more about a particular gene is to "blast it" (i.e., run it through a program called **BLAST** which stands for "**B**asic **L**ocal **A**lignment **S**earch **T**ool").

What will we be searching and analyzing (and why?)



We will search for genes that encode "seven transmembrane" proteins. This is large family of proteins involved with all sorts of information or signal pathways in cells. One type of information processed by seven transmembrane proteins is olfactory. Some of these proteins bind specifically to odor molecules and pass information via neurons to the brain. Seven transmembrane proteins are found embedded in cell membranes looped through like thread " through cloth. The parts of the protein within the membrane are mostly " hydrophobic while the parts looping or dangling free are mostly "hydrophilic." Hydrophobic means water-hating and therefore likely to be avoiding water and deep within a membrane. Hydrophilic means water-loving and free of the membrane.

DETAILED PROCEDURE:

☛ **SNAG THE SEQUENCE**

1. **Go to NCBI** (the home page is full of useful and interesting buttons that you may want to explore later).
   http://www.ncbi.nlm.nih.gov/

2. Meanwhile click " **all databases**", your entry point for many types of database searches.

3. Type in (for example) the search words, where * is any letter: ***olfact* receptor***

4. Notice that the search results include lots of information besides sequence data. Look at some of the other buttons later.

5. Press "**core nucleotide**" to get 1000's of hits for DNA sequences. Look at the names as you scroll down and try to select a vertebrate animal (rather than an invertebrate) and an olfactory receptor rather than some related molecule.

6. Select a file and scroll down to the bottom where you will find a sequence like this. Be sure that it begins with ATG. **Capture (copy) the sequence AND include in the capture its name and identification numbers from the top**. The example sequence shown below has had its line numbers removed. However, depending on what you do with the sequence, you may have the option of running it through a Perl program to automatically remove the numbers and punctuation.

XM_357989. Mus musculus olfa...[gi:38090297]

```
atgaatgata tgacatcagg aaactattgc acagtgactg agttcttctt
ggcagggctc tcagagaagc cagaactcca gctgcccctc ttcttccttt
tcataggaat ttatatgatc actgtagcag ggaatttggg catgatcata
ctgattgggc tcagttccca cctgcacaca cccatgtact atttcctcag
cagtctgtcc ttcattgact tctgtcagtc cacagtcgtt acccctaaaa
tgctagtgaa ctttgtgaca gagaagaaca tcatatccta ccctggatgc
atgactcagc tctacttctt cctcatattt gcaattgcgg agtgttacat
tttagctgca atggcatatg accgctatgt tgctatctgt aacccattgc
tttacaatgt aaccatgtcc tatcaaattt acattttcct aatttcagga
gtgtatatta ttggtgtgat ctgtgcatca gctcacacag gcttcatggt
tagaattcga ttctgcaaat tagatgtgat caaccactat ttctgtgacc
ttcttcccct cttgaagctt gcatgctcta atacctatat caatgaaatg
ttgattctat tttttgggac actgaacatc tttgtcccaa tcctgaccat
tattacttcc tacatcttca ttattgccag catcctccgc attcgctcca
ctgaaggcag gtctaaagcc ttcagtactt gcagttctca catcttggct
gttgctgtct tctttggatc tctagcattc atgtaccttc agccatcatc
agtcagctcc atggaccaag ggaaagtgtc ctctgtgttt tataccattg
ttgttcccat gctgaacccc ttgatctaca gtctgaggaa taaagatgtt
gctgttgcct tgaaaaaaat aattgaaaga aaaacattta tgtag
```

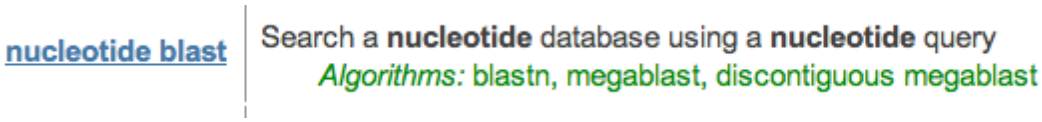
7. **Open a text editor and paste your header and sequence into this file**. [WINDOWs users: we recommend you use the Word Pad editor. MacOS users: use TextEdit – On either platform, if you use Microsoft Word, save this file as TEXT ONLY (ask someone if you are not sure how to do this)]. Because you will open/read this file in your next Perl program, you can *not* save it as a "regular" Microsoft Word file, it must be a "plain" text file). Save the file with a meaningful filename, e.g., for our example above, we might use: `Mouse_XM357989.fna`. Keep this file as you will need it with your next Perl program.

After you have made a file , go back and BLAST part of your sequence

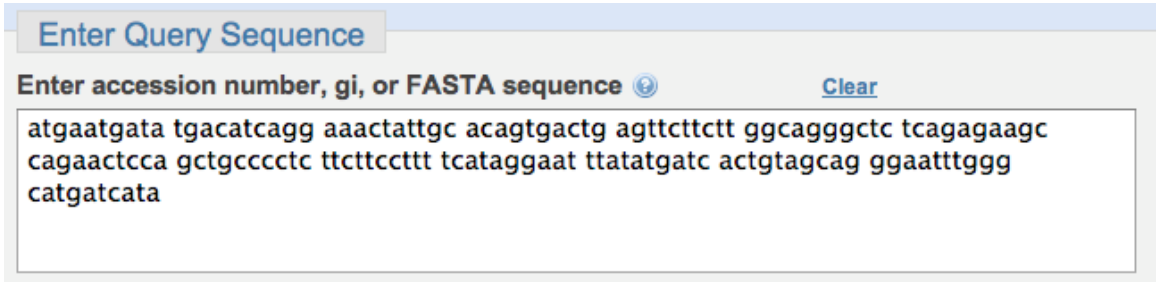
☛ **BLAST -- compare your sequence to other sequences**

1. Capture a few lines of the sequence and go back to the **NCBI homepage** (try clicking the NCBI icon)

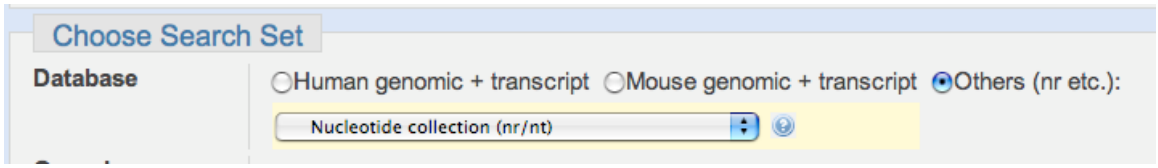
2. Click BLAST and then **"nucleotide BLAST"**



3. **Paste** in the DNA sequence and **remove the header line (name), any leading numerals (line numbers) and punctuation**. NOTE: You do *not* have to edit the spacing; spaces in the sequence will be ignored by BLAST.



4. Choose a database. A good choice is **"nr"**, which looks at several databases. Click "**Others**" in the Choose Search Set to select nr *etc*.



5. Press **BLAST** and begin the search. (We suggest getting your results in a new window)

6. If many other scientists are BLASTing too, you may have to wait a bit….

7. And *voila*, there are your results, ready for your interpretation!

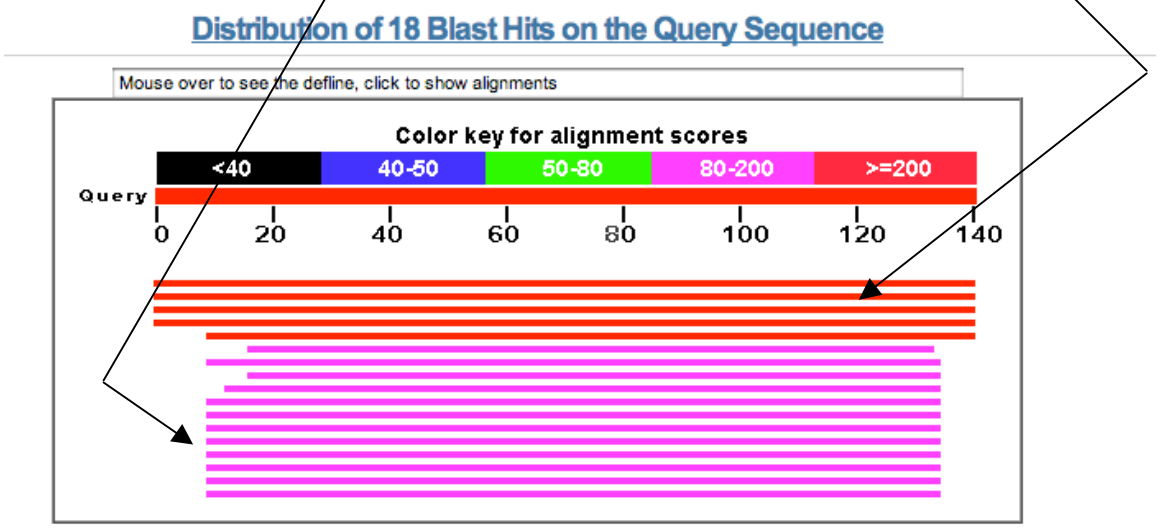
Interpreting Your Results

Job Title: Nucleotide sequence (150 letters)

BLASTN 2.2.17 (Aug-26-2007)

RID: H7F5V79701R

Scroll down. You will see a colored graph, showing " hits" (matching sequences). **Long red** lines are perfect matches. **Shorter colored lines** are genes in which parts of the sequence matches to some degree.



Note that the colored lines on the graph are click-able and will send you immediately further down in the file to specific lines of information. Next you will see written information about those same genes.

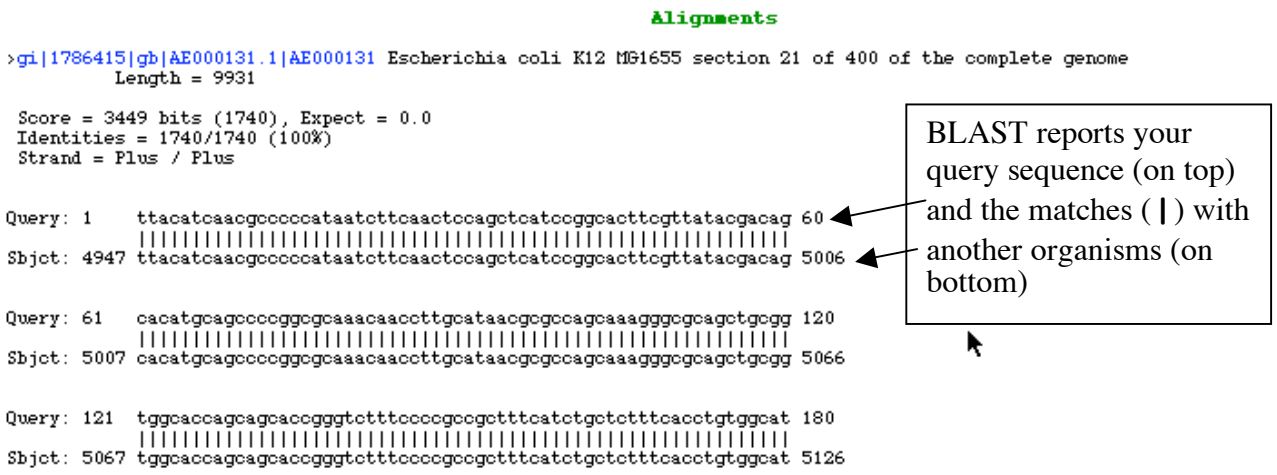
E-values

Sequences producing significant alignments:
(Click headers to sort columns)

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
NM_146830.2	Mus musculus olfactory receptor 44 (Olfr44), mRNA	259	259	100%	2e-66	100%	U E G
AF282271.1	Mus musculus odorant receptor K11 gene, complete cds	259	259	100%	2e-66	100%	E G
AC073434.42	Mus musculus strain C57BL/6J chromosome 9 clone rp23-159l13, comple	259	259	100%	2e-66	100%	
AY073263.1	Mus musculus olfactory receptor MOR171-5 gene, complete cds	254	254	100%	8e-65	99%	G
AY318147.1	Mus musculus olfactory receptor Olfr44 (Olfr44) gene, complete cds	243	243	93%	2e-61	100%	G
NM_001000960.1	Rattus norvegicus olfactory receptor 1292 (predicted) (Olr1292_predicted	161	161	83%	5e-37	91%	U G

In the right-most columns you will also see the results of a statistical analysis telling you **how well the sequences matched**. The " **E-value**" (Expected Value) is an indicator of the stringency of the search. The lower the E-value, the fewer mismatches there should be. **E= 0.002 is considered to be a good threshold** below which the results are good; E-values > 0.002 are the results that are less reliable and perhaps impossible to interpret. Please look at E-values as you attempt to make sense of the results.

Further down you can see exactly which letters of sequence aligned with your gene.



Scroll to the very bottom to see the most poorly aligned sequence with higher E-values to get an idea of what BLAST allows as a minimal alignment.

Going further with interpretation:

Why might a researcher be BLASTing " olfactory" genes?

One reason is to get closer to an understanding of how such a complex protein evolved and the varied ways that cells use this and similar proteins. Our knowledge of a fruit fly mutation in seven transmembrane proteins can further our understanding of mutations in other organisms.

FAQs

Q_1: What else is at NCBI?

A_1: Lots of great stuff! This is the most important database for the biological sciences. Here are three examples (if you have time, TRY THESE RIGHT NOW!)

1. If you do lots of searching of the biolological abstracts via "MedLine", go directly to NCBI (book mark it!) and into " PubMed". It's *the* fastest way! Try searching "olfaction".
2. For a quick, friendly key word search of human genetic diseases, click " OMIM" (Online Mendelian Inheritance in Man). Try it now! You don't have to know the official name for a disease. Ordinary key words work fine. For example, try "olfaction" to see what sorts of disorders there are.
3. If you have time for a lengthy download and if you have the right plug in, try "Structure" from the NCBI homepage. Type in "olfaction" or some other key

word associated with a protein. If you download their image playing software, it is possible to see and rotate and manipulate images with this tool. Impress your friends!

Q_2: How might I use a BLAST search from the points of view of a biologist or a computer scientist?

A_2: for a biology class, consider when writing your next lab report or biology paper including a little sequence analysis or a figure showing actual sequence for a relevant gene or protein. For example, let's say you are writing a paper on how insulin works. How about searching and retrieving the actual gene and protein sequences for insulin and whatever other annotation you can find?

When you discuss insulin as a protein, you can show the sequence too. If you are discussing mutations to the insulin gene, you might be able to show exactly where they are!

For any paper in which you are comparing organisms, consider comparative BLAST searches. How do insulin genes differ from one mammal to another?

For any paper about evolution, BLAST searches are powerful ways to show relatedness. In fact entire family trees are constructed from BLAST-type searches of important (conserved) genes.

A_2: for a computer science class, you are writing software! You can probably imagine a little Perl program that would help you compare strings in "BLAST"-like fashion, although with a much simpler output. However keep in mind all of the pre-existing software. Do not "re-invent the wheel" [1] unless it seems appropriate. Instead, see what you can incorporate from simple " word" utilities (like word finders), spread sheet commands, and (for genomics) all of the software that is available at NCBI.

A_2: From the p.o.v. of a biologist *or* a computer scientist, keep in mind how UNEXPLORED genome sequences are *and* how many questions have not yet been asked. There is plenty of room to be CREATIVE in genomics and even to make ORIGINAL DISCOVERIES. If you are interested in doing GENOMICS RESEARCH, please see Mark LeBlanc or Betsey Dyer for more information and encouragement. Take a look at our research web site:

`http://genomics.wheatoncollege.edu`

[1] Actually, one of my rules as a researcher is "Sometimes you DO have to reinvent the wheel". That is, when "starting from scratch", you may get a deeper understanding of a problem and its relatedness to other problems *and* you may come up with a better wheel. Do not assume that past researchers have arrived at the perfect solution. Take ownership of the problem from the inside out. -BDD