Bioinformatics laboratory

(Written by Alexandra Shoes, UCSF and used with permission)

What is 'bioinformatics'?

(My thoughts)

Bioinformatics is the utilization of computation for biological investigation and discovery. It is the process by which you unlock the biological world through the use of computers.

What should I read for this lab?

1. Read briefly about viruses in Campbell (pgs. 334-344). Pay special attention to the virus figure 18.8 on page 341.

2. Skim the Wikipedia page on influenza viruses.

3. Skim the Wikipedia page on the 1918 flu pandemic (aka Spanish Flu Pandemic).

Where can I explore bioinformatics?

(!! are recommended sites)

Science Magazine Netwatch !!
(Weekly list of cool science sites. Not always bioinformatics sites per se. A fun way to see how computers are being used in science and to communicate science)

ExPASy Life Sciences Directory !!
(This is the standard web site to go to for bioinformatics programs)

NCBI Portal !!
(**The** US site for bioinformatics related resources. Maintained by the National Center for Biotechnology Information, the National Library of Medicine and the National Institutes of Health. You can do literature searches, examine genomes, get programs to do sequence analysis and download or search the current protein and nucleotide databases... among other things)

Bioinformatics. Org
(General bioinformatics information)

Bioinfo Online

Bioinformatics Links Directory !!
(Links, links and more links to bioinformatics resources)

Google Bioinformatics Directory

Lab:

The Blame Game

The greatly feared pandemic flu virus has finally broken out. Millions are sick and thousands have already died. It is almost impossible for the Centers for Disease Control and Prevention (CDC) to keep track of the new cases that are reported each day. Contrary to everyone's expectations, the first reported cases appeared in San Francisco and not in Asia or Eastern Europe. At the same time, the New York Times is reporting that the recently reconstituted and extremely dangerous 1918 influenza virus was seriously mishandled by several scientific laboratories. Apparently, the virus was shipped to scientists at UC San Francisco without using the appropriate shipping procedures. The news report has hypothesized that perhaps the package was damaged en route or potentially mishandled onsite at UCSF. In immediate reaction to the newspaper's report all related parties at UCSF have been arrested for the illegal dissemination of a biological agent to the public. Several of the arrested parties are researchers without US citizenship (but with appropriate visas) and some members of congress are calling for immediate deportation or even reclassification of their status to 'Enemy Combatants' and trying them as terrorists. In other related news, the virus strain from patients in San Francisco has been fully sequenced and, just today, released to the public.

Discussion:

1. What is this case about?

2. What do we need to know?

3. Where do we find the information we need?

Questions to Answer:

Say you're an enthusiastic bioinformatician suffering through a miserable (but not deadly) bout of the new flu. You don't know much about the 1918 flu virus but you're darn unhappy with how sick you feel. How do you figure out whom to blame? Evolution? Or thoughtless, potentially criminal, scientists?

1. The best site for cutting edge research information is **PubMed (a searchable site of research article citations)**.

- Go to the site and type **'1918 influenza'** into the search bar.

A) How many total references are returned?

B) Say you wanted to compile information on the 1918 influenza virus, how would you deal with the information from PubMed? Are the search terms '1918 Influenza' too broad or too narrow? How would you select the few papers you would actually like to read? Is PubMed the only resource you would use? Very briefly describe what your 'plan of attack' would be to get the information that you need.

2. Through your research you determine that the 1918 flu is an RNA virus: **Influenza A of type H1N1**. Since you have the sequence data from the San Francisco virus you decide to examine some of the sequence to see if it is similar to the 1918 flu sequence. The sequence you have is

>SF strain sequence
TTCATGCAGCAAAAGCAGGTTTAGTACCATGGACAACCAAACAAAGACAATG
ACTATCACTTTTCTCATCCTCCTGTTCACAGTAGTGAAAGGGGACCAAATAT
GTATCGGATACCATGCCAACAATTCCACAGAAAAAGTCGACACAATCTTGGA
GCGAAACGTCACCGTGACTCATGCCAAGGACATTCTTGAGAAAACGCATAA
TGGGAAGTTGTGCAGATTAAGCGGGATCCCTCCATTGGAACTGGGGGATTG
CAGCATTGCAGGTTGGCTCCTT
GGAAATCCGGAATGTGACCGGCTCTTAAGTGTACCTGAATGGTCCTATATA
GTGGAAAAGGAAAACCCGGCGAATGGTCTGTGCTACCCAGGCAGTTTCAAT
GATTATGAGGAACTGAAACATCTCCTCACCAGTGTGACACACTTTGAGAAAG
TTAAGATTCTGCCCAGAGATCAATGGACCCAGCACACAACAACTGGTGGTT

CTCGGGCCTGTGCAGTATCTGGCAACCCGTCATTCTTCAGAAACATGGTTT
GGCTTACAGAGAAGGGGTCAAACTACCCAATTGCTAAAAGATCATACAACAA
CACAAGCGGGAAGCAAATGCTGGTAATTTGGGGGATACATCATCCCAATGA
CGATACGGAACAAAGGACACTGTACCAAAATGTGGGAACATATGTTTCCGT
GGGAACATCAACACTGAATAAGAGGTCAATCCCTGAAATAGCAACAAGGCC
CAAAGTCAATGGACAAGGAGGGAGAATGGAATTCTCTTGGACTCTATTGGA
GACATGGGATGTCATAAATTTTGAGAGCACTGGTAATTTAATTGCACCAGAA
TACGGATTCAAAATATCGAAGAGAGGAAGCTCAGGAATTATGAAGACAGAG
AAAACGCTTGAAAATTGTGAAACCAAATGTCAGACCCCCTTGGGGGCAATAA
ATACAACACTGCCTTTTCACAACATTCACCCATTGACAATAGGTGAGTGCCC
CAAATATGTAAAGTCAGATAGATTGGTTAAGGCGACAGGGCTAAGAAATGTC
CCTCAGATTGAATCAAGGGGATTGTTTGGAGCAATAGCCGGGTTTATAGAA
GGCGGATGGCAAGGAATGGTTGATGGCTGGTATGGGTATCATCACAGCAAT
GATCAAGGATCAGGATATGCAGCAGACAAAGAATCCACTCAAAAGGCAATT
GATGGGATAACTAACAAAGTAAATTCTGTGATTGAAAAGATGAACACTCAGT
TTGAGGCTGTTGGGAAAGAGTTCAACAATCTAGAGAGAAGACTGGAAAACT
TAAATAAAAAGATGGAAGATGGATTTCTTGATGTATGGACATATAATGCCGA
ACTCCTAGTTTTAATGGAAAATGAGAGGACACTTGATTTCCATGATTCTAATG
TGAAGAATCTGTACGATAAGGTCAGAATGCAGTTGAGAGACAATGCTAAGG
AAATAGGGAATGGATGCTTTGAGTTTTATCATAAATGTGATGATGAATGCAT
GAATAGTGTCAGGAATGGGACATATGATTATCCCAAATATGAAGAAGAGTCC
AAGCTGAACAGAAACGAAATCAAAGGAGTGAAATTGAGCAATATGGGGGTT
TATCAAATACTTGCTATATACGCTACAGTTGCAGGCTCCTTGTCACTGGCAA
TCATGATAGCTGGGATTTCTTTCTGGATGTGTTCTAATGGGTCTCTGCAATG
CAGAATTTGCATATGACTGTAAGT

A) This sequence is supposed to code for a virus protein. This is an RNA sequence. How many reading frames does it have?

B)If it were a DNA sequence, how many reading frames would it have?

C) The start codon for translation is **ATG**. In the first reading frame (in the 5'3' direction) locate and label on your worksheet the first start codon. (Remember sequences are normally written 5'3')

D) Using this **translation table** find the first stop codon in the first reading frame and mark it on your worksheet. *(Because of how RNA is sequenced remember that the 'T's in the above sequence correspond to 'U's in the translation table.)*

E) How many bp (base pairs) long would this gene be?

<!--[if !supportLists]--**F)** How many amino acids is this? Is this a reasonable length for a gene?

G) Label the first start codons in the second and third reading frames in the 5'3' direction. <!--[endif]-->

H) In the second reading frame there are 1707 bp from the first Met to the stop codon. In the third reading frame there are 54 bp. Which reading frame likely contains the gene?

I) How many amino acids long is the actual gene?

As you can tell, this can be a time consuming process. Fortunately we have computer programs to help us do this quickly.

3. Go to the **Translate Tool** on the ExPASy web site. Cut and paste the above sequence into the box and click on **'translate sequence'**. Make sure to **only** cut and paste the sequence (and not the '>SF strain sequence').

***Note**: The ExPASy (Expert Protein Analysis System) web server is an extensive and valuable resource put together and maintained by the Swiss Institute of Bioinformatics. A large variety of databases and analysis tools related to proteins are available from them. ExPASy started in 1993 was the very first web server in the field of life sciences.*

A) Which reading frame contains the gene?

B) What are the first 6 amino acids of the protein?

C) All the amino acids except the start codons (Met) are abbreviated by single letters. What do the single letters of the first 6 amino acids stand for? *(If you don't know, google 'amino acid single letter codes')*

D) What is the single letter abbreviation of Met?

Click on the frame that contains the gene (i.e. click on the **'5'3' Frame...'** that matches the frame of the gene). This brings up a page with the sequence of the translation. Click on the Met that starts the gene to bring up a page titled **'Virtual:...'** (an informational page about the sequence).

Click on the **'FASTA format'** link at the bottom to bring up the FASTA formatted sequence of the protein (*note: FASTA format is simply a standard format that people use to write out sequences. Click here to get a full description*). Open up **Notepad**. Cut and paste the full sequence (including the **'>virt|...'** line) into **Notepad** and save it for later.

4. Now that we have both the nucleotide and amino acid sequence of the SF strain viral protein, let's try to find out some functional and evolutionary information about the sequence. Let's compare the sequence we have to other known sequences. We can do this with a **BLAST** search on the NCBI web site.

***Note**: NCBI (National Center for Biotechnology Information) was established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease*

Go to NCBI's **BLAST server**. Under **'Basic BLAST'** click on **'nucleotide blast'**. Cut and paste the above nucleotide sequence into the search window **'Enter Query Sequence'**. Then go to the section labeled **'Database'**. Using the drop down list chose the database **'Nucleotide collection (nr/nt)'**. Then click **'BLAST'** at the very bottom of the page. A new page will come up that will eventually have your results on them (it might take a moment).

A) Look at the top hit, what is the E-value of the hit?

B) What does an E-value mean? (hint: check out this page).

C) Scroll down to the section labeled (in green)**'Alignments'** What is the 'gb number' of the top hit?

D) Is the top hit a virus sequence?

E) If so, what kind of virus is it (name and type)? (Remember that the 1918 virus is an **Influenza A virus of type H1N1**).

F) Is this the same type as the 1918 virus?

G) What animal was the sequence isolated from? Where was it isolated?

H) What is the description (name) of the gene?

5. What we would really like to know is whether or not the SF strain is most closely evolutionarily related to the 1918 flu. So let's spend some time looking at the protein sequence of the gene. Go back to the FASTA amino acid sequence that you saved to **Notepad**. Go again to the **BLAST server** and this time click on **'protein blast'**. Cut and paste the amino acid FASTA sequence (the **'>virt|...'** sequence) into the search page and click on **'BLAST'** and wait for your results.

Scroll down to the **'Alignments'** section and look at the top five hits.

A) What are the names and types of the 5 top hits?

B) What are their gb numbers?

Select the top five hits (by clicking on the little box next to the gi number for each hit). Then click on **'Get selected sequences'**. This brings up a page listing the general info on the sequences. What we want are the actual amino acid sequences. Go to the drop down list next to **'Display'** and choose **'FASTA'**. After the new page loads go to the **'Send to'** drop down list and choose **'Text'**. This will give you a page that just lists the FASTA formatted sequences. Cut and paste all the sequences into the **Notepad** document where you have the other amino acid sequence. Make sure to include all of the **'>...'** lines when you cut and paste.

As you can imagine, just staring at lists of sequences is not very illuminating. Let's look at these sequences in a more visual way by aligning the sequences and then looking at the alignment.

Go to the **ClustalW sequence alignment site**. Under the heading **'OUTPUT FORMAT'** choose **'gcg MSF'** from the drop-down list (don't worry about entering your email or any of the other boxes). Now cut and paste all the amino acid sequences from the **Notepad** file into the query box. Make sure to include the lines that start with **'>...'**! It turns out that we also have the same protein from the 1918 virus. Add this to the sequences in the query box by cutting and pasting the sequence below into the box after the other sequences (again, make sure to include the **'>...'** line). You should have 7 sequences in the query box.

>gi|4325018|gb|H1N1_1918 protein
MEARLLVLLCAFAATNADTICIGYHANNSTDTVDTVLEKNVTVTHSVNLLEDSHN
GKLCKLKGIAPLQLGKCNIAGWLLGNPECDLLLTASSWSYIVETSNSENGTCYP
GDFIDYEELREQLSSVSSFEKFEIFPKTSSWPNHETTKGVTAACSYAGASSFYR
NLLWLTKKGSSYPKLSKSYVNNKGKEVLVLWGVHHPPTGTDQQSLYQNADAY
VSVGSSKYNRRFTPEIAARPKVRDQAGRMNYYWTLLEPGDTITFEATGNLIAPW
YAFALNRGSGSGIITSDAPVHDCNTKCQTPHGAINSSLPFQNIHPVTIGECPKYV
RSTKLRMATGLRNIPSIQSRGLFGAI
AGFIEGGWTGMIDGWYGYHHQNEQGSGYAADQKSTQNAIDGITNKVNSVIEK
MNTQ

After you have done this, click on **'Run'** and wait for your results. Scroll down to the alignment and click on **'Show Colors'**. The colors help you see when sequences are different from each other.

C) In the first line of sequences, how many times is there an amino acid difference between the 1918 sequence and the SF strain sequence? (remember, the SF sequence will be labeled something like **'virt|...'**)

D) Based on this alignment, do you think that the SF strain protein is most like the 1918 protein?

Now scroll back up to the top of the file. Click on the filename link next to **'Alignment file'**. The file will open up in a new window. Save this window for question six. The window will be referred to as the 'alignment file'.

6. An even better way to tell if things are closely related to each other evolutionarily is to look at a phylogenetic tree.

***Note**: Phylogenetics is the scientific field that examines the evolutionary relatedness of different species. Molecular phylogenetics uses either nucleotide or amino acid sequences to determine the evolutionary relationships. A phylogenetic tree is the visual representation of the determined evolutionary relationships between the sequences under inspection. An example of a phylogenetic tree is **this** (from **Taubenberger et al., Nature: 437, 889-893, 2005**). A well-built phylogenetic tree is a much more reliable indicator of the evolutionary relatedness of two sequences than a simple pair-wise comparison (because phylogenetic trees are built from multiple sequences which allows for a stronger evolutionary signal versus noise).*

Let's use the ClustalW site to calculate a phylogenetic tree. Go back to the **ClustalW sequence alignment site.** Cut and paste the alignment file from the previous question into the query box. This time look at the section labeled **'Phylogenetic Tree'** above the query box. Go to the drop down menu under **'Tree Type'** and choose **'phylip'**. Now click **'Run'** and wait for your results.

When the results come up, click on the link next to the **'Phylip tree file'** heading. This will bring up the text file of a tree. Obviously it's pretty hard to know what it looks like by just staring at the file. Let's visualize it. Go to the **Indiana University Biology Department Phylodendron site**. Click on the little button next to **'phenogram'**. Cut and paste the text of the tree file into the query box and then hit **'Submit'**. Look at the tree and answer the following questions.

A) Is the 1918 sequence closest to the SF strain sequence?

B) What is the label of the sequence closest to the SF strain sequence?

C) Look back at your sequences, what is the name, type, animal and location of that sequence?

D) After all this analysis, do you think that the pandemic flu is the 1918 flu virus?

E) Should the scientists be prosecuted or let go?

F) Why might you still be unsure?

7. The reconstitution of the 1918 flu virus has been very controversial. It is one of many ethical debates facing the scientific community today

A) What are the ethical concerns about this research?

B) What are arguments for the virus reconstitution?

C) If the decision were up to you, would you have reconstituted the virus?

D) Given that it has been reconstituted do you trust the scientific community adequately and ethically care for the virus?

E) What regulations/rules would you like to see put in place?

8. In preparation for this assignment I had you read over some Wikipedia pages.

A) What are some concerns about using a site such as Wikipedia? (hint: check out **Wikipedia**)

B) Why might a site like Wikipedia be good to use for research purposes?