Lab II: Genes and Genomes

Week 2

Project 3: Working with Whole Genome Databases

Procedure 3A: Working with Complete Viral Genomes

Adapted from pages 96-98 from Bioinformatics for Dummies (2003).

Viruses can be considered “parasites” because they cannot survive outside of a host. As we observed with HIV in Lab I, they depend on the host cell’s organelles and energy supply to replicate and propagate. All viruses contain nucleic acid, either DNA or RNA (but not both), and a protein coat called a capsid that surrounds the genetic material. Some viruses also have an envelope outside the capsid, which facilitates entry into the host cell. Viral genomes come in a variety of biochemical forms (RNA/DNA, circular/linear, single/double stranded) and size (a few kilobases to half a million). Viruses are generally classified by the organisms they infect - animals, plants, or bacteria.

We will use the NCBI Viral Genome resource to find out more about viruses.


2. On the black menu bar at the top of the form, click Genome.

3. Click the Viruses link (on the right side of the form, under the Related Resources section).

4. Scroll down the Viral Reference Genomes page until you reach the table of available viral genome sequences grouped by class (Deltavirus, Retroid viruses, and so on).

5. Type HIV1 in the Search window, and then click the Find button.

Your browser returns a global picture of the HIV-1 genome, which indicates the identity and respective positions of all the genes.

Under Lineage, HIV-1 is listed as a Retroid Virus (aka retrovirus). Hopefully you remember that the genetic material of retroviruses is in the form of RNA (Figure 4), which means that the virus must first convert its RNA into DNA and then integrate the DNA copy into the host’s genetic material. Retroviruses accomplish this using the enzyme reverse transcriptase.
6. Click the Coding Regions link, on the left side of the page (blue background).

The Protein List page appears. As we learned in Lab I, the three main genes of interest for HIV-1 are gag, pol, and env. The gag gene encodes for proteins that mediate viral particle assembly and budding from the plasma membrane. The pol gene encodes for the protein products involved in transcription - reverse transcriptase, protease, and ribonuclease. The env gene encodes for the envelope proteins that bind to the host T cell and facilitate entry into the cell. We will look at some of these sequences later in more detail.

For more information on HIV replication, visit the following websites:

http://www.stanford.edu/group/nolan/tutorials/tutorials.html
http://www.thebody.com/niaid/hiv_lifecycle/virpage.html
http://www.pbs.org/wgbh/nova/aids/action.html

Procedure 3A Formal Response Questions

a) How many base pairs are there in the HIV-1 viral genome?

b) Look at the segment of base pairs between 4630 and 5400. How many gene products have a portion of their sequence in this segment? What are these gene products?
Procedure 3B: Working with Complete Bacterial Genomes

Adapted from pages 99-101 from Bioinformatics for Dummies (2003).


2. Click Genome on the black menu bar near the top of the form.

3. Click the Microbial link, near the top of the Related Resources section (right column).

4. Click the Escherichia coli O157:H7 link.

5. Click in the circle.

Click the Coding Regions link (in the left margin) to obtain an individual sequence downloading form, similar to the one we accessed for the HIV-1 genome.

6. Click one of the individual rectangles.

This returns the pre-computed results of a comprehensive BLAST similarity search between this precise protein sequence and a large list of species. We will conduct BLAST searches in Labs III and V.

Procedure 3C: More Bacterial Genomics at TIGR

Adapted from pages 102-104 from Bioinformatics for Dummies (2003).

1. Access the TIGR homepage at www.tigr.org/tdb/.

2. Click the Streptococcus pneumoniae link beneath the Comprehensive Microbial Resource (CMR) link.

Your browser displays a simple-looking genome page that contains all you’d ever want to know about Streptococcus pneumoniae and its genome. This page includes links from which you can freely download sequences and functional annotations.

3. Choose chromosome Streptococcus pneumoniae TIGR4 from the pull-down menu in the lower-right corner of the page.

As soon as you release your mouse button, your browser takes you to a genome viewer. Spend a few minutes exploring this genome browser.
Procedure 3C Formal Response Question

a) Give one real-life scenario in which you would use NCBI Entrez and TIGR to explore the genome of a previously sequenced microbe.

Project 4: Exploring the Human Genome

Procedure 4A: Getting Started on the Ensemble Site

Adapted from pages 106-107 from Bioinformatics for Dummies (2003).

1. Access the Project Ensembl home page at www.ensembl.org/.

2. Click the Ensembl tour link, under the Help and Documentation section.

3. Scroll down to the bottom of the page and click the right-pointing arrow on the control panel you see. (The arrow is the one right next to the little Home icon.)

4. When you’re done with the tour, go back to the Ensembl home page by clicking the Home icon.

5. Click the Documentation button on the lower-right side of the Ensembl home page (in the Help and Documentation section).

6. On the Documentation page, click the Sitemap link near the bottom-left corner.

Procedure 4B: Zooming In and Out Along Human Chromosomes

Adapted from pages 107-110 from Bioinformatics for Dummies (2003).


2. Click the Human button under the Ensembl Species heading in the right column.

3. Click chromosome 12, then click right in the middle of the white band annotated p13.31 (near the top of the chromosome).

4. In the Overview box, click on 7.10 Mb directly under the black and white band. NOTE: If 7.10 Mb is not within the range on your screen, move along the chromosome by clicking on the blue DNA (contigs) band.
This zooms you onto the CD4 gene location. You should be able to spot CD4 among other Ensembl genes in the overview.

5. Scroll down to the detailed view and click on the CD4 gene name under Ensembl trans.

You should see the Ensembl Gene Report for CD4. This report gives you a wealth of information on the CD4 gene, including predicted transcripts, homology matches, and disease matches. Explore the links at your leisure.

Procedure 4B Formal Response Questions

a) What is the location (in base pairs) of the CD4 gene in humans?

b) Which organisms have homology matches with the human CD4 gene? What does this tell you about the nucleotide sequences for these genes?

c) On which chromosome is the CD4 gene located for each homology match? What does this tell you about the genomes of eukaryotes?

Procedure 4C: Finding Genes with Coding SNPs

Adapted from pages 110-113 from Bioinformatics for Dummies (2003).

We first looked at single nucleotide polymorphisms (SNPs) when we linked to the variation sequences in GenBank entry NM_000616 for the mRNA gene for CD4. We can also find out which genes on a specific chromosome have been linked to a disease by identifying coding SNPs – nucleotide changes that may alter the sequence and possibly the shape and/or function of the corresponding protein.

- NOTE: This online resource is updated frequently and may not match up exactly with the procedure. If you notice differences, try to adapt the procedure to the new format on your screen. If you are unable to obtain the results, notify the instructor.

1. Access the Project Ensembl home page at www.ensembl.org/.

2. Click on the EnsMart button under the Ensemble Species heading.

3. Click on the Start button next to the Using EnsMart heading.

4. On the Start page, keep the default Homo sapiens and Ensembl Genes settings and click the Next button.
5. On the Filter page, make the following changes:

   a.) In the Region box: choose 12 from the Chromosome name pull-down menu.
   b.) In the Gene box: choose Disease Genes from the pull-down menu.
   c.) In the Expression box: leave the default parameters.
   d.) In the Multi Species Comparisons box: leave the default parameters.
   e.) In the Protein box: leave the default parameters.
   f.) In the SNP box: select the Entries Associated with SNPs of Type check box. Select the Coding Only radio button.

6. Click Next.

7. On the Output page, make the following changes:

   On the Features tab:
   a.) In the Region box: select the Chromosome Name check box and the Band check box.
   b.) In the Gene box: select the Ensembl Gene ID check box and the Disease OMIM ID check box.
   c.) In the Expression box: nothing should be selected.
   d.) In the Multi Species Comparisons box: nothing should be selected.
   e.) In the Protein box: nothing should be selected.

   On the SNPs tab:
   a.) In the Gene Associated SNPs box: select the Allele check box, the CDS location (bp) check box, and the Location in gene (coding etc) check box.
   b.) Leave everything else with the default parameters.

8. Click the blue Export button at the top-right or bottom-right of the page.

   Although you may have to wait several minutes for the results page to load, Ensembl returns a table that contains every disease gene you asked for, along with the coding SNP nucleotide changes and their precise locations the coding region of each affected protein. Each Chromosome Name, Ensembl Gene ID, and Band ID is an active hyperlink.

9. Scroll down until you find the Ensembl ID ENSG00000010610 (located on band p13.31) and click one of the links.

   You should recognize the Ensembl Gene Report for CD4 that you accessed previously using the Contig View page of chromosome 12.

Procedure 4C Formal Response Questions
a) Why is a mutation in a coding SNP segment more serious than a mutation in a non-coding SNP segment?

b) Choose a disease gene from one of the SNP coding regions in the EnsMart results list. Write a paragraph on the disease, including the mutation that causes the disease as well as the implications.

c) Sometimes SNP polymorphisms are not harmful and can even result in resistance or partial resistance to a disease, as is the case with the C-C chemokine receptor 5 on human T cells. Use any of the bioinformatics skills you’ve learned so far to find out more about SNP mutations in this gene. Write a one page double-spaced paper on this topic and be sure to include the chromosomal location of the mutation, as well as the effect of the mutation on individuals with the SNP polymorphism. Remember to include your references.

Procedure 4D: More Human Genome Browsing Tools

Adapted from pages 113-115 from Bioinformatics for Dummies (2003).

There are several other browsing resources for the human, the mouse, and other animal genomes. The choice of which resource to use depends on your particular needs. Here, we will find the genomic location of the GenBank entry NM_000616 (human CD4 mRNA gene again) using the Genome Bioinformatics Laboratory at the University of California in Santa Cruz.

1. Access the UCSC Genome Bioinformatics home page at http://genome.cse.ucsc.edu/.

2. Choose Human from the Organism pull-down menu at the top left, and then click the Browser link (just below the Organism window).

3. Type NM_000616 in the Position text box, and then click the Submit button.

The Genome Browser displays the genome region for the CD4 gene. You can explore this resource on your own.

Procedure 4D Formal Response Question

a) Compare the genome browser from the UCSC Genome Bioinformatics webpage with the Ensembl genome browser. What features do the browsers have in common? What are the differences between the two? Which do you find easier to use and why?
Project 5: Comparing Genomes Using Vista Genome Browser

The Vista Genome Browser is a Java program that allows you to view the results of whole genome comparisons. It displays a Vista graph of the whole genome alignments interactively, allowing you to easily scroll, zoom in, and adjust display parameters. The Vista curve is calculated as a windowed-average identity score for the alignment. A variable sized window (Calc Window) is slid across the alignment and a score is calculated at each base in the coordinate sequence. That is, if the Calc Window is 100 base pairs, then the score for every point X is the percentage of exact matches between the two alignments in a 100bp-wide window centered on that point X. Due to resolution constraints when visualizing large alignments, it is often necessary to condense information about a hundred or more base-pairs into one display pixel. This is done by only graphing the maximal score of all the base pairs covered by that pixel. The Vista Genome Browser is a useful resource for comparing genomes of different species in terms of percent of conservation but is not specific to the actual nucleotides that are conserved. We will examine nucleotide conservation between genomes at a later date using more sophisticated software.

Procedure 5A: Using Vista Genome Browser


2. Scroll down to Using Vista Genome Browser and read this section to familiarize yourself with the program.

3. At the top of the page, click on the Launch Vista Genome Browser button.

4. Type CD4 in the Gene Position box and click on the Reload button.

5. Click on the Contig Details button to see which two genomes you are comparing.

6. Click Get Conserved Regions.

   Your browser returns a table of conserved regions between the CD4 gene for the human genome and the mouse genome, along with percent identity and the type of region (exon, non-coding, etc.).

7. On the Vista Plot page, click on the UCSC Browser button to access the UCSC Genome Browser page for the human CD4 gene.

   This page is similar to the one you found in the last section of this lab, as it should be, since both records are for the human CD4 gene.
Procedure 5A Formal Response Questions

a) What do the heights and colored regions of the peaks in the Vista Plot tell you about the similarity of the two genomes? Which regions would you expect to show the most similarity between the two genomes and why?

b) What are the limitations in using Vista Browser?

Refer to Chapter 3 in Bioinformatics for Dummies (2003) for additional information on genomic databases.

* A large portion of this laboratory was adapted from and text citations throughout this lab are for: Claverie, J. and C. Notredame. Bioinformatics for Dummies. Wiley Publishing, Inc., 2003.