CS 234: “Computational Methods for the Analysis of Biomolecular Data”

Homework 1

Posted: October 10th, 2022 Due: 11:59pm, October 19th, 2022

Directions

• Please submit your answers via email to Amir amohs002@ucr.edu as a PDF attachment (please use your name as filename)

• You are expected to work on this assignment on your own. Anything that you submit that comes from “external sources” (a friend, a web page, a book, etc.) must be acknowledged and will be graded accordingly.

Q1: Create a web page (20 points)

Create a simple web page with your name, your contact information and your research interests (if any). This page will be used to contain the progress reports about your final project. Please post updates on your project at least every two weeks.

Put email the link to your own CS 234 page to Stefano at stelo@cs.ucr.edu

Q2: Protein folding (20 points)

In this question we will explore protein folding by playing a fun game.

1. Go to https://fold.it/

2. Create an account using the following convention for the the username: use the first three letters of your first name, followed by ‘.’, followed by first three letters of your last name, followed by .cs234 For example, the username for Stefano Lonardi should be ste.lon.cs234

3. Once you have created and logged into your Foldit account, go to the CS 234 FoldIt group at https://fold.it/portal/node/2006408, and click "Request Membership"

4. Download and install the game “FoldIt” from https://fold.it/portal/ on your computer (Windows, Mac, and Linux)

5. Play the game, and try to complete the introductory puzzles (there are several, but most of them are really quick)

6. Feel free to attempt some of "Science Puzzles"

7. The CS 234 group page https://fold.it/portal/node/2006408 will show all the players in the class

Nothing need to be submitted for Q2. Just have fun playing this game and learn about protein folding.
Q3: Genome sizes and data storage (30 points)

The NIH’s National Center for Biotechnology Information (NCBI) provides a huge repository and a multitude of databases for biological information.

1. Go to NCBI’s [https://www.ncbi.nlm.nih.gov/home/genomes/](https://www.ncbi.nlm.nih.gov/home/genomes/) to find the approximate size in base pairs (i.e., nucleotides) of the genomes for the following organisms:
   - hepatitis delta virus
   - microbial *Escherichia coli* K12
   - bakers’ yeast *Saccharomyces cerevisiae*
   - worm *Caenorhabditis elegans* - haploid
   - human (*Homo sapiens*) - haploid
   - plant *Paris japonica*
   - marbled lungfish *Protopterus aethiopicus*

2. Find out the estimated total number of genes in each of the above organisms. Is the size of genome proportional to the total number of genes? Give at least one reason why this is or is not the case. Is it always true that the more complex the organism, the large genome it has? Give an example if your answer is no and explain why.

3. What is the minimum number of bytes required to store the genomes listed above? You can assume that each nucleotide can be represented by 2 bits (i.e., A:00, C:01, G:10, T:11), therefore a byte can represent four nucleotides. What if you want to store the human genome in its diploid rather than haploid form? Show your calculations!

4. What is the minimum number of bytes needed to store all human genomes on our planet? All such genomes can be represented as a single individual’s genome plus the variations, or polymorphisms, seen in all other human genomes. Assume that the human population is about 7.8 billion, and that polymorphic sites tend to be simple single nucleotide polymorphisms (SNPs, such as A in one genome and C in another) and occur about once every 1000 bases. You can assume that SNPs are distributed uniformly, but the reality is that they are not (see this article [http://book.bionumbers.org/how-genetically-similar-are-two-random-people/](http://book.bionumbers.org/how-genetically-similar-are-two-random-people/)). Show your calculations below.

5. How many 1 TB memory sticks (flash drives) would it take to store all the human genomes on the planet?

6. Some nucleotide sequence data might need more than two bits/base. Could you think of a reason why this would be the case?

Q4: Exploring a genetic disease on the web (30 points)

Adrenoleukodystrophy (ALD) is an inherited neurodegenerative disorder that effects about 1 in 15-20,000 males. It arises through a mutation in the ABCD1 (or ALD) gene, which is found on the X chromosome. People with adrenoleukodystrophy accumulate high levels of saturated, very
long chain fatty acids in their brain, which causes the loss of myelin on nerve fibers. The effective
detection of the ABCD1 mutation is key for diagnosing this disease.

We will concentrate on the gene ABCD1. Go to https://www.ncbi.nlm.nih.gov/search/, and search for “Homo sapiens ABCD1”. NCBI will show many hits from several databases. We want to take first a look at the protein, so click on “Protein”. You will get about several dozen hits, select the entry labelled Accession:NP_000024 (version 2, which is represented by “.2”)

Familiarize yourself with the data format returned by Genbank. Scan the entries, scroll down to look at the features and the actual primary sequence. Observe in the annotations that some amino acids can chemically modified (phosphorylation, glycosylation). Q 4.1: How many amino acids are in the protein produced by the Human ABCD1 gene?

Now jump to the corresponding entry in the nucleotide database by clicking on the REFSEQ link (on the fifth line of the Genbank file), accession NM_000033 (that’s the mRNA entry for that gene, this is one is version 4)

Again, familiarize yourself with the data format returned by Genbank for the transcript (or mRNA). Scan the annotations (exons, polyA sites, etc.) Q 4.2: How many exons are annotated in the Human ABCD1 mRNA? Q 4.3: How many nucleotides are in the Human ABCD1 mRNA? Q 4.4: Why there are no intron in this entry?

Now we want to take a look at the known SNPs for the ABCD1 gene. Go back to the NCBI page https://www.ncbi.nlm.nih.gov/search/, search for “Homo sapiens ABCD1” and select the dbVar database. You will get several hundred hits.

Observe that some of the SNPs are annotated “Pathogenic”, some are “Benign”, but the many have no known clinical significance.

Now we want to see how many copies of the ABCD1 gene we have in the human genome. Go to BLAST http://www.ncbi.nlm.nih.gov/BLAST/ and click “Human” under the Search query. Do not change the parameters, but type in the text gadget the ID of the nucleotide entry NM_000033. Select as database “T2T-CHM13v2.0” which is the most recent assembly of the human genome.

You request will be processed by BLAST. It may take a few seconds, but you will get eventually the alignment of ABCD1 transcript on the human genome. Check out the alignment produced by BLAST by clicking on the Description column.

Q 4.5: How many BLAST hits did you get? Q 4.6: On which chromosomes did you get hits? Q 4.7: Which chromosome did get the hit with the highest score?

As it turns out, over 90% of the last 1/3 of the ABCD1 gene matches exactly some sections on several other chromosomes. This suggests that the ABCD1 locus of the X chromosome has somehow duplicated to these other chromosomes. Because of the high similarities observed, it is thought that these ABCD1 copies arose a mere 5-10 million years ago, which is relatively recently on an evolutionary scale. These copies of ABCD1 are not thought to be functional, and have been observed previously.

Q 4.8: How do you explain the fact that this gene is fragmented in 10 pieces along the X chromosome (i.e., the BLAST alignment is broken into 10 ranges)?