Assignment 2

Many of the assignments in this course will introduce you to topics in computational biology. You do not need to know anything about biology to do these assignments other than what is contained in the description itself. The objective of each assignment is for you to acquire certain particular skills or knowledge, and the choice of topic is independent of that objective. Sometimes the topics will be related to computational problems in biology, chemistry, or physics, and sometimes not.

This particular assignment is an exercise in extracting information from files that are too big for mere mortals to process manually. The real power of computers is that they can do simple things extremely quickly, meaning millions, perhaps billions of times per second, much faster than people can. There is a kind of file called a PDB file that contains structural information about proteins, nucleic acids, and other macromolecules. A macromolecule is just a big molecule. Macro means big. PDB is an acronym for the Protein Data Bank. PDB files can be downloaded from the Protein Data Bank at [http://www.rcsb.org/pdb/home/home.do](http://www.rcsb.org/pdb/home/home.do).

A PDB file contains information obtained experimentally, usually by either X-ray crystallography, NMR spectroscopy, or cryo-electron microscopy [1]. (You do not need to know this to do the assignment, but it is important for those who intend to pursue a bioinformatics concentration.) These files completely characterize the molecule, providing for example,

- the three-dimensional positions of every single atom in the file,
- where the bonds are,
- which amino acids it contains if it is a protein (or nucleotides if DNA or RNA),

and much more. The information is not necessarily exact. Associated with some of this information are confidence values that indicate how accurate it is. A PDB file is a plain text file; you can view its contents in any text editor, such as gedit or nedit, or with commands such as cat, more, and less. Each line in a PDB file begins with a word that characterizes what type of line it is. For example, some lines start with the word REMARK, which means they are comments about the file itself. Some lines start with SOURCE, and they have information about the source of the data in the file. Some lines start with words such as MODEL, CONECT, ATOM, and HETATM. Each has a different meaning in the file. Take a look at some of the PDB files in the directory /data/biocs/b/student.accounts/cs132/data/pdb_files before you read any further, so see what they contain.

Proteins are chains of amino acids. Amino acids are organic compounds that carry out many important bodily functions, such as giving cells their structure. They are also instrumental in the transport and the storage of nutrients, and in the functioning of organs, glands, tendons and arteries. Amino acids have names such as alanine, glycine, tyrosine, and tryptophan. They are also known more succinctly by unique three-letter codes. The table below lists the twenty standard amino acids with their three-letter codes.


[2] Amino acids are the building blocks of proteins, which may contain many thousands of them.
Each line that starts with the word ATOM represents a unique atom in the protein. So do the lines that start with HETATM, but these are atoms in water molecules surrounding the particular protein when it was crystallized, and we want to ignore them for now. Lines that start with ATOM contain the three-letter code for the amino acid of which that atom is a part. For example, an atom line for an atom in a phenylalanine molecule looks like this:

ATOM 3814 N PHE J 24 -17.763 -7.816 -12.014 1.00 0.00 N

The three-letter code is always in uppercase.

Suppose for a moment that you had access to several PDB files representing various proteins and you needed to know how many atoms within that protein belonged to a particular type of amino acid, i.e., how much of that protein was made up of a given amino acid. You could open the file and start counting the lines by hand. This would take forever. Instead, you could use your knowledge of UNIX to solve the problem in a few seconds. There are commands in UNIX that you have learned in this class so far, that you can use to determine how many atoms are in any PDB file, and even more, how many atoms of a specific type are in a given file. These commands are relatively easy to use, assuming you have a little ingenuity. You will have to read the man pages for them. In this assignment, your task is to figure out how to use these UNIX commands to count the number of a specific type of atom found in particular PDB files located in the directory /data/biocs/b/student.accounts/cs132/data/pdb_files.

To be specific, for each of the files listed below, which are contained in the pdb_files directory, you are to determine how many atoms of the types listed next to the file name are in that file:

1A4P.pdb  tyrosine, glycine
1A36.pdb  serine
1BZR.pdb  valine, leucine
1AI0.pdb  glutamate
1018.pdb  alanine, cysteine
Create a file that contains five lines, one for each file named above, and on each line show the number of atoms of the given amino acids in that file. It is up to you to determine how to create this file, but it must be plain text. I do care how you count the atoms, and for this reason, in addition, at the bottom of the file, you are to write the UNIX commands that you used to obtain this information. You must use UNIX commands to count these lines.

**Submitting Your Solution**

Name your file in the form `hwk2_username` where `username` is your exact username, case-sensitive, and place it in the directory

```
data/biocs/b/student.accounts/cs132/projects/project2
```

For example, my homework file would be named `hwk2_sweiss`. Make sure it has the permissions `rw-------`, meaning only you can read it or modify it.