Lab – PDB parser – Diameter - Centroid

PART1. Diameter of a protein

Write a program that computes the diameter of a protein, i.e. the maximum Euclidean distance between two atoms of a protein. The program consists of the following steps carried out sequentially.

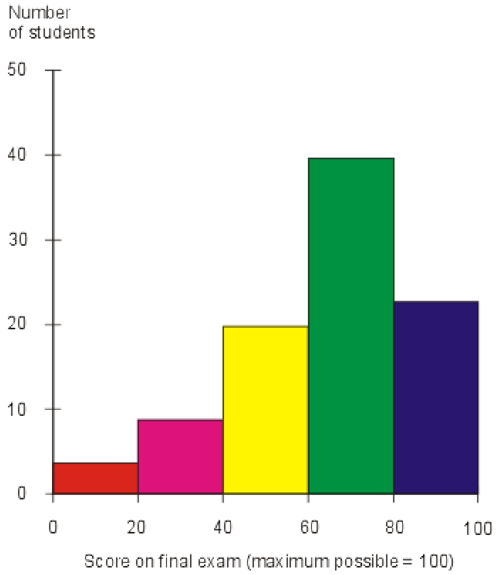
1. It takes in input the name of a protein with extension .pdb
2. Reads the PDB file and stores it as an array *PDB_data* with an element of the array corresponding to a line of the PDB file starting with **ATOM**
3. It extracts from the array *PDB_data* the coordinates of all atoms of the protein and creates three arrays x, y and z with such coordinates.
   In the PDB files the atom coordinates are contained in the lines starting with ATOM. (For the pdb file format, see the documentation at the PDB website or at t-square, resource pdb.pdf)
4. It computes and prints the maximum Euclidean distance over all pairs of atoms. Recall that in three dimensional space the Euclidean distance *d* of two points P1(x_1, y_1, z_1), and P2 (x_2, y_2, z_2) is defined as:

$$d \;=\; \sqrt[2]{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$$

PART2. Plot the results

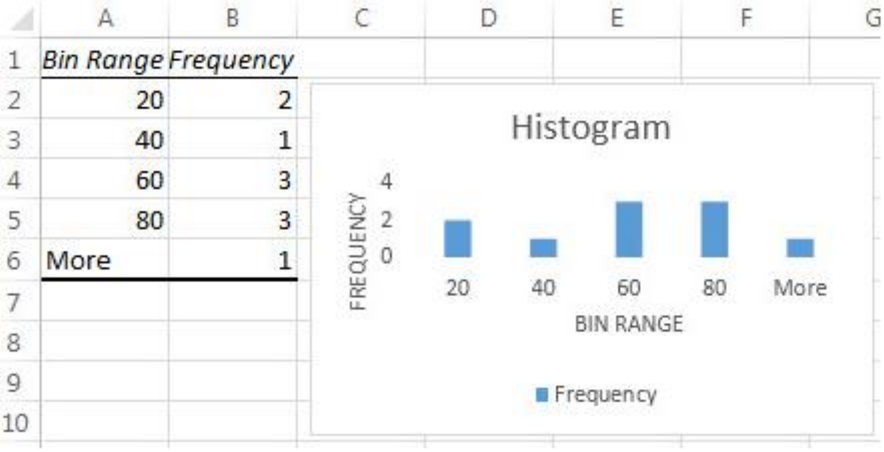
Build a histogram of the distances of all pairs of atoms of a protein.

In statistics a histogram is graphically displayed by vertical rectangles representing frequency data. The figure below is an example of a histogram (grades in a final exam).



In your program, you need to output the appropriate data for making a histogram in excel. The picture is like the above except that on the x axis you have range of distances and on the y axis number of atom pairs with distance in a range.

An example of histogram plot in excel (you don’t need to follow exactly as this example):



Bin Range	Frequency
20	2
40	1
60	3
80	3
More	1

The output data could be implemented by an array of elements where each element corresponds to an interval of distances and contains the frequency (or counter) of atom distances that fall in that interval. Assume, for instance, that the max distance between two atoms in the proteins is 200 Angstroms. If you divide the distance range (0 -200) into 20 intervals each of length 10, then the histogram array *h* will consist of 20 elements: h[0] contains the number of atom distances d_i with $0 \leq d_i < 10$, h[1] the number of distances $10 \leq d_i < 20$, .., h[19] the number of atom distances in the range $190 \leq d_i < 200$.

(Hint: to map a distance value *d* into an index i of the histogram array divide *d* by the interval length and take the floor of the results. Recall that in mathematics the floor of a real number x is defined as the largest integer not greater than x. For example if d=49, then d/10=4.9 and floor(4.9) = 4. Thus the index of the array is 4.)

Perl does not have an explicit floor function. However, it is very simple to create a floor function, since the int() function simply removes the decimal value and returns the integer portion of a number. Thus you can write i = int(x) where x = d/10.

PART3. Centroid

Write a program that computes the centroid of the set of atoms of a protein defined as follows.

Given a set of atoms with coordinates (x_1, y_1, z_1), (x_2, y_2, z_2), …, (x_n, y_n, z_n), the centroid or baricenter is the point C with coordinates (x_c, y_c, z_c) given by:

$$\mathrm{x_c} = (\textstyle\sum_{i=1}^{n} x_i)/n \quad \mathrm{y_c} = (\textstyle\sum_{i=1}^{n} y_i)/n \quad \mathrm{z_c} = (\textstyle\sum_{i=1}^{n} z_i)/n$$

The program consists of the following steps carried out sequentially.

1. It takes in input the name of a protein with extension .pdb
2. It reads the PDB file and stores it as an array *PDB_data* with an element of the array corresponding to a line of the PDB file starting with **ATOM**
3. It extracts from the array *PDB_data* the coordinates of all atoms of the protein and creates three arrays x, y and z with such coordinates.

In the PDB files the atom coordinates are contained in the lines starting with ATOM. (For the pdb file format, see the documentation at the PDB website or at t-square, resource pdb.pdf)

4. It computes and prints the coordinates of the centroid.
5. It computes the distance of every atom from the centroid and determines the minimum and maximum distances.

Submit two separate program files (with extension .pl) containing the two programs and the excel file with the plot. Name your files Diameter_your_last_name.pl (for PART1 and PART 2) and Centroid_your_last_name.pl (Part3). The programs read in the name of a protein from the argument on the command line as follows.

```
Diamater_your_last_name.pl  protein_name.pdb

Centroid_your_last_name.pl protein_name.pdb
```