ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ & ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Εθνικόν και Καποδιστριακόν
Πανεπιστήμιον Αθηνών
ΙΔΡΥΘΕΝ ΤΟ 1837

Algorithms in Structural Biology

Homework 1 - cRMSD and dRMSD

MSc Data Science and Information Technologies

Andrinopoulou Christina (DS2200013)

Introduction

Molecules have a 3-dimensional structure. The geometry of a molecule usually is called *conformation*. Many times we want to compare the structure between two molecules. In this exercise, we studied this problem using the well-known approaches of *coordinate Root Mean Square Deviation (cRMSD)* and *distance Root Mean Square Deviation (dRMSD)*.

The python notebook *cRMSD_dRMSD.ipynb* contains the implementation for this exercise and the corresponding results. However, we are going to discuss these results and the algorithms in this report.

We used as input the file *80 conformations.txt*, which contains the coordinates in 3D space of each atom, for 80 different conformations. We suppose that every pair of conformations that we are going to be compared has the same number of atoms and there is correspondence between the k-th atom of the first conformation and the k-th atom of the second conformation.

coordinate Root Mean Square Deviation (cRMSD)

The first approach that we studied for the comparison of the 3D structure of two molecules is the cRMSD algorithm. In this case, we should compare the structure between 2 different conformations and the goal is to reduce as much as possible the cRMSD distance, that we are going to discuss later on.

For this purpose, we should transform one of the two conformations in order to approach as much as possible the other conformation and as a result to reduce the cRMSD distance. The aforementioned transformation includes rigid transformations (that preserve the distances between the atoms), such as translations and rotations.

The algorithm below includes all the steps for this approach. We suppose that the first conformation is X and the second conformation is Y. The first step of the algorithm is to find the centroid of each conformation using the formulas: $x_c = \sum_{i=1}^{N} \frac{x_i}{n}$ and $y_c = \sum_{i=1}^{N} \frac{y_i}{n}$. After that, we subtract these centroids from each coordinate. This step moves the conformations to the origin of the space. The next step includes a well-known technique from linear algebra, the *Singular Value Decomposition (SVD)*. The SVD finds the best transformation for the conformation. The Q matrix is this transformation, so we apply this matrix to the X conformation and as a result, we rotate and translate this conformation so as to approaches the Y conformation and calculate the corresponding cRMSD distance.

For the calculation of the cRMSD distance, we measure the norm between each y_i atom and the transformed x_i atom and we raise the result in the power of two. We add all these distances, we divide the result with the number of atoms (n) and we calculate the squared root.

Algorithm 1: cRMSD

Result: cRMSD distance

$x_c = \sum_{i=1}^{N} \frac{x_i}{n}$;

$y_c = \sum_{i=1}^{N} \frac{y_i}{n}$;

$X = \{x - x_c : x \in X\}$;

$Y = \{y - y_c : y \in Y\}$;

SVD: $X^T Y = U \Sigma V^T$;

$Q = UV^T$;

if $detQ < 0$ **then**

  $Q = [U_1, U_2, -U_3]^T V^T$;

end

return $\sqrt{\frac{\sum_{i=1}^{n} ||Qx_i - y_i||^2}{n}}$;

In the implementation, we changed the calculation of cRMSD. Instead of using a for-loop for each atom and calculating the norm $||Qx_i - y_i||$, we utilized the matrix multiplication. In particular, we do the multiplication between the matrix Q and X and then we subtract from the result the matrix Y with one command. After that, we calculate the norm of the matrix.

The python notebook includes the class *cRMSD*, which contains functions for all the steps of the algorithm. So, the user has to create a cRMSD object and call the *compare* function, which activates the full pipeline.

For the first question of the exercise, we calculate the cRMSD distance between the first two conformations of the txt file. The cRMSD is equal to 0.6271694758794248. We also give the results of the SVD:

$$U = \begin{pmatrix} -0.614345 & 0.311999 & -0.724732 \\ -0.046984 & -0.931337 & -0.361115 \\ -0.787637 & -0.187799 & 0.586821 \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} 34676.241277 & 8650.358619 & 5259.297438 \end{pmatrix}$$

$$V = \begin{pmatrix} -0.613880 & 0.311817 & -0.725204 \\ -0.050610 & -0.932335 & -0.358035 \\ -0.787775 & -0.183088 & 0.588123 \end{pmatrix}$$

and the transformation matrix

$$Q = \begin{pmatrix} 1.000000 & -0.000315 & 0.000611 \\ 0.000318 & 0.999988 & -0.004851 \\ -0.000609 & 0.004851 & 0.999988 \end{pmatrix}$$

For the second question we should calculate the cRMSD between all $\binom{80}{2}$ pairs of conformations. For this reason, we have to call the function *compare_all* of the class *cRMSD*, which activates the pipeline for all the possible pairs of conformations of the txt file.

We calculate the mean and the median of the distances. The mean is 11.015199832626305 and the median is 10.853087486496054 . Also, we create a histogram of the distances.

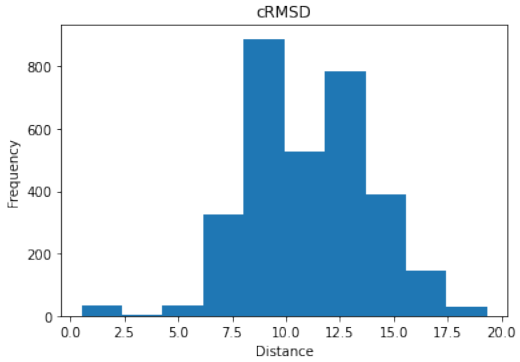

Figure 1: Histogram of cRMSD distances

distance Root Mean Square Deviation (dRMSD)

The last question of the exercise is similar to the second, but in this case, we should calculate the dRMSD distance, instead of the cRMSD. For this purpose, we created a second class, in the same notebook, which is called *dRMSD*.

In this case, each molecule is characterised by the distances between the atoms. So, we assume that we know the distances d_i and d'_i between point-pairs for each conformation. The calculation of dRMSD is based on the formula below:

$$dRMSD = \sqrt{\frac{1}{k}\sum_{i=1}^{k}(d_i - d'_i)^2}$$

where $k \leq \binom{n}{2}$.

The benefit of this approach is that we have not to find the optimal transformation in the space for one of the conformations, because in this case, we compare the distances between the atoms of the two conformations. However, it is time-consuming if $k = \binom{n}{2}$. Another drawback is that this approach does not take into consideration conformations that are mirrored in the right way.

First of all, we calculate the dRMSD between all $\binom{n}{2}$ distances within each conformation. The mean of the distances is 6.7950321975044385 and the median is 6.468606946347217 .

Furthermore, we create the corresponding histogram.

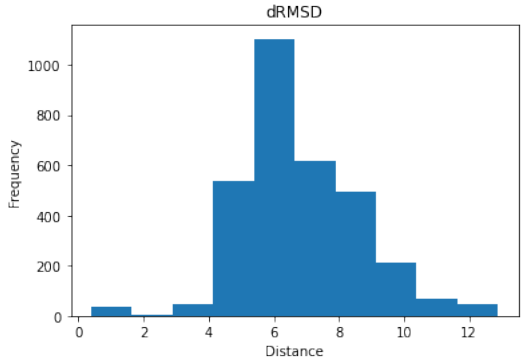

Figure 2: Histogram of dRMSD distances

After that, we choose only $k = 3n$ pairs of atoms, randomly. So, we compare the conformations, using only the distances of these randomly selected pairs of atoms. The mean of the dRMSD is 6.769383746351923 and the median is 6.453469919580371.

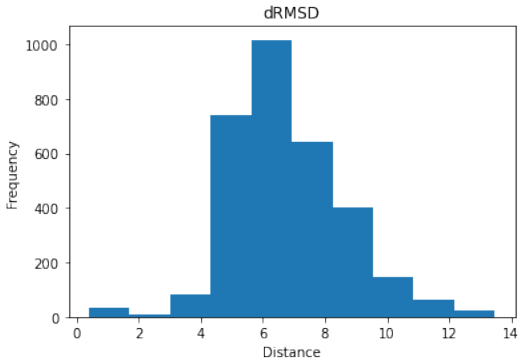

Figure 3: Histogram of dRMSD distances

We notice that the results of the random approach are similar to the results of the brute-force approach. However, in the random approach the execution time is less than the brute-force to a great degree.