ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ & ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Εθνικόν και Καποδιστριακόν Πανεπιστήμιον Αθηνών
ΙΔΡΥΘΕΝ ΤΟ 1837

Algorithms in Structural Biology

Homework 2 - Part 2

MSc Data Science and Information Technologies

Andrinopoulou Christina (DS2200013)

1 Part II

At this part of the assignment, we examine only one part of the SARS-COV-2 Spike glycoprotein. In particular, we examine the residues from 401 to 450 of the E chain. From these residues, we want only the Ca atoms. We will use this part of the complex in order to create the Cayley-Menger matrix and then we will try to simulate the results that we would take from NMR and we implement an algorithm that computes the 3D coordinates of a structure based on the corresponding distances.

The jupyter notebook "part2.ipynb" contains the implementation for this part of the exercise. However, the comments on the results are in this report.

1.1 Construction of the Cayley-Menger matrix

The Cayley-Menger matrix (or border matrix) is a matrix that contains the distances between atoms (in a way). The construction of the matrix is based on another matrix, the distance matrix. The distance matrix M is squared, symmetric with real entries matrix and the diagonal of the matrix is equal to zero. These properties come from the fact that the matrix contains the distances between atoms, in other words, the distances between points in the 3D space. As a result, the distance between a point and its self is zero, the distance between the point i and the point j is equal to the distance between point j and i, and of course, the distance cannot be negative. Every term of the distance matrix is calculated based on the formula below:

$$M_{ij} = \frac{1}{2} dist(p_i, p_j)^2$$

where p_i and p_j are the coordinates of the atom i and j.

The construction of the Cayley-Menger matrix is based on the M matrix:

$$\begin{pmatrix} 0 & 1 & 1 & ... & 1 \\ 1 & & & & \\ ... & & & M & \\ 1 & & & & \end{pmatrix}$$

1.2 Computation of the Cayley-Menger rank

The next step is to calculate the rank of the Cayley-Menger matrix. Because of the *Cayley-Menger Theorem*, we expect that the rank of the border matrix will be 5. This theorem claims that M embeds in $\mathbb{R}^d$ if and only if the rank of the border matrix is equal to $d+1$. In our case, we focus on 3D embedding, so the rank of the border matrix has to be equal to 5.

We utilize the corresponding function in the jupyter notebook for the calculation of the rank and we can confirm that the rank is 5, as we expected.

1.3 Perturbation of the Cayley-Menger matrix and rank

For this part of the assignment we perturb the entries of the Cayley-Menger matrix that we constructed before. In this way, we try to simulate the results from NMR spectroscopy that returns approximate distances and no the real ones.

For the perturbation of the border matrix, we change every entry, apart from the entries of the first row and the first column. We change the values of the table in a random way, by adding or subtracting a percentage of the initial value. The selection between the addition or the subtraction is completely random. The changes of the matrix entries do not damage the properties of the matrix (positive values, symmetry, zero diagonal). More details are included in the jupyter notebook.

In this case, we perturb the border matrix by 2% and 4% and then we calculate the rank of the matrix. For both cases, the rank is greater than 5. This means that the distance matrix that is included in the Cayley-Menger matrix cannot be embedded in the 3D space and this comes from the perturbation of the matrix that we did before.

1.4 From distances to coordinates

The final task is to translate the distances into coordinates. In general, if we know all the distances between a set of points we can extract the coordinates of these points. For finding the coordinates of the Ca atoms, in this case, we utilize the following embedding algorithm.

Algorithm 1: Embedding via SVD

Result: 3D coordinates of atoms
Pick the first atom of the structure as origin;
Determine the Gram matrix G, given the Cayley-Menger matrix;
Execute SVD on Gram matrix: $G = U\Sigma V^T$;
Force $rank(G) = 3$ and get the corresponding Σ';
Return the 3D coordinates that are calculated based on the formula: $P = \sqrt{\Sigma'}V^T$;

First of all, the function picks the first point of the structure as the origin and this point will be utilized for the calculation of the Gram matrix in the next step.

The Gram matrix is a $n \times n$ matrix, where n is the number of the atoms that we examined. The value of the term G_{ij} is given by the formula:

$$G_{ij} = \frac{d_{i0}^2 - dij^2 + d_{j0}^2}{2}$$

where d_{ij} the distance between the atoms i and j. The other two distances of the previous formula are the distances from the origin to i and to j. It is obvious that the term $\frac{d_{ij}^2}{2}$ is the term B_{ij} of the Cayley-Menger matrix. Therefore, the construction of the Gram matrix is based on the Cayley-Menger matrix in our implementation and as a result, if the Cayley-Menger matrix is perturbed, then the Gram matrix is influenced by this perturbation.

After the construction of the Gram matrix, we apply the Singular Value Decomposition (SVD) on this matrix and we utilize the Σ and the V^T matrices. First of all, if the distances between the matrices were the real ones, we would expect that the rank of the matrix would be 3 and as a result, the number of the non-zero entries in the Σ matrix would be only 3. However, after the perturbation, the distances are not the real ones and this means that there is no 3D embedding of the structure. Therefore, the rank is larger than 3, as we have already mentioned in the previous section, and it is obvious that the number of the non-zero values in the Σ matrix is more than three. However, we want to get a 3D embedding, so we force the rank to be equal to 3 and that's why we create a Σ' diagonal matrix that contains only the three larger values of Σ.

Finally for finding the 3D coordinates of the structure, the formula $P = \sqrt{\Sigma'}V^T$ is utilized.

1.5 Comparison with cRMSD

As we mentioned before, we perturbed the Cayley-Menger matrix in order to simulate the results from the NMR spectroscopy. However, in this case, we know the ground truth, so we can easily compare the real structure to the structures that are constructed assuming some noise. For this purpose, we use the cRMSD as a tool for comparison. The cRMSD, actually, will calculate the distance between the real structure and a structure that we could get by a method, such as the NMR, which definitely introduces some noise to the data. The implementation of the cRMSD is from scratch and is abstracted from the first assignment of the same course.

The cRMSD distance between the original structure and the perturbed structure by 2% is approximately equal to 1.5 and the cRMSD between the original structure and the perturbed structure by 4% is slightly larger and it is approximately equal to 1.7. We say that the distance is "approximately" equal to some value because the perturbation of the Cayley-Menger matrix is random and as a consequence, the structure is slightly different from one execution of the code to the other. Therefore, the final cRMSD changes from one execution to the other, but of course not to a great degree. Another worth-mentioning point is that a small increase in the percentage of the perturbation increases the cRMSD, as well. So, the introduction of a greater amount of noise to the initial data affects the procedure and this may lead to a higher (cRMSD) distance from the ground truth.