1 Exercise 2 c-RMSD and d-RMSD

Given are 10 conformations of a molecule in file "10 conformations.txt" on eclass:Assignments with n = 369 atoms on the backbone (hence in correspondence). The file starts with 2 lines containing 10 and n; the rest uses tabs to define 3 columns containnig n triplets x y z per conformation i.e. 2 + 10n rows total:

10
369
2.816 -11.005 10.087
4.43 -10.545 10.011
...

Implement c-RMSD and d-RMSD in Matlab, Mathematica, Maple or other system offering linear algebra (SVD); submit your code. If your system provides either of these functions, it is OK to just use it.

1. Compute the c-RMSD distances between all $\binom{10}{2}$ pairs of conformations. Use them to find the L1-centroid conformation i.e. the one that minimizes the sum of distances to the other 9 conformations.

2. Repeat (1) for d-RMSD using (a) all k = $\binom{n}{2}$ distances within each conformation, or (b) a random subset of k = 3n distances.

3. Do they all 3 approaches yield the same centroid? How do they compare in terms of speed?

2 Solution

2.1 Question 1

At first i created a data frame with the 10_conformations.txt. The function `read_conformations()` outputs the following data frame head:

Index	X	Y	Z	Molecule
0	2.816	-11.005	10.087	Mol_1
1	4.430	-10.545	10.011	Mol_1
2	3.476	-10.324	9.659	Mol_1
3	3.551	-10.478	8.633	Mol_1
4	3.109	-8.958	10.026	Mol_1

Table 1:

At this data frame there is an extra column that indicates the name of the moelcule as Mol_i.

After that i applied the algorithm for the Coordinate Root Mean Square Deviation :

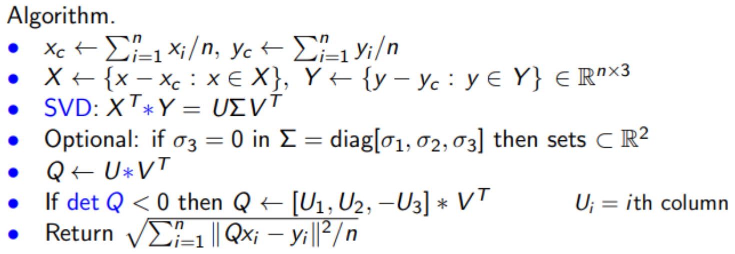
1) Find the centroid of each conformation

2) Move the conformations to the origin of the space: Subtract the centroids from each coordinate

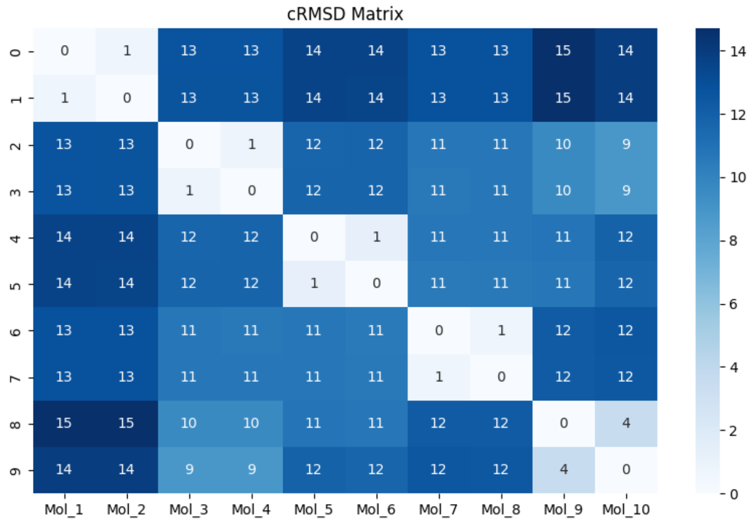
3) Singular Value Decomposition (SVD): best transformation Q for the conformation

4) Apply transformation to a conformation

5) Calculate the corresponding cRMSD distances as we can see on matrix 2.1



```
Algorithm.
• $x_c \leftarrow \sum_{i=1}^{n} x_i/n,\ y_c \leftarrow \sum_{i=1}^{n} y_i/n$
• $X \leftarrow \{x - x_c : x \in X\},\ \ Y \leftarrow \{y - y_c : y \in Y\} \in \mathbb{R}^{n\times 3}$
• SVD: $X^T * Y = U\Sigma V^T$
• Optional: if $\sigma_3 = 0$ in $\Sigma = \mathrm{diag}[\sigma_1, \sigma_2, \sigma_3]$ then sets $\subset \mathbb{R}^2$
• $Q \leftarrow U * V^T$
• If det $Q < 0$ then $Q \leftarrow [U_1, U_2, -U_3] * V^T$          $U_i =$ $i$th column
• Return $\sqrt{\sum_{i=1}^{n} \|Qx_i - y_i\|^2/n}$
```



	Mol_1	Mol_2	Mol_3	Mol_4	Mol_5	Mol_6	Mol_7	Mol_8	Mol_9	Mol_10
0	0	1	13	13	14	14	13	13	15	14
1	1	0	13	13	14	14	13	13	15	14
2	13	13	0	1	12	12	11	11	10	9
3	13	13	1	0	12	12	11	11	10	9
4	14	14	12	12	0	1	11	11	11	12
5	14	14	12	12	1	0	11	11	11	12
6	13	13	11	11	11	11	0	1	12	12
7	13	13	11	11	11	11	1	0	12	12
8	15	15	10	10	11	11	12	12	0	4
9	14	14	9	9	12	12	12	12	4	0

The `crmsd()` function follows exactly every step of the algorithm given the 2 matrices X, Y.

Using the `start_time` command, as i applied it for every possible pair of molecule, I eventually ended up with a matrix with all the cRMSD distances. The cRMSD matrix took 0.05213141441345215 sec to be created.

Using calculations that can be found in the codex file, I conclude that the molecule that has the least distance from the others is the 4th molecule and its average distance from the others is 8.935966389127664 .

2.2 Question 2-3

Similarly in order to find the Distance Root Mean Square Deviation i just created a function that applies the dRMSD's formula :

$$dRMSD = \sqrt{\frac{1}{k}\sum i = 1k(d_i - d_i')^2}, k \leq \binom{n}{2}$$

Where d_i are the distances between the atom i and the rest of one molecule and d_i' are the distances of an other.

The function `drmsd()` finds the distances between the atoms in a molecule and stores them in a symmetric matrix.

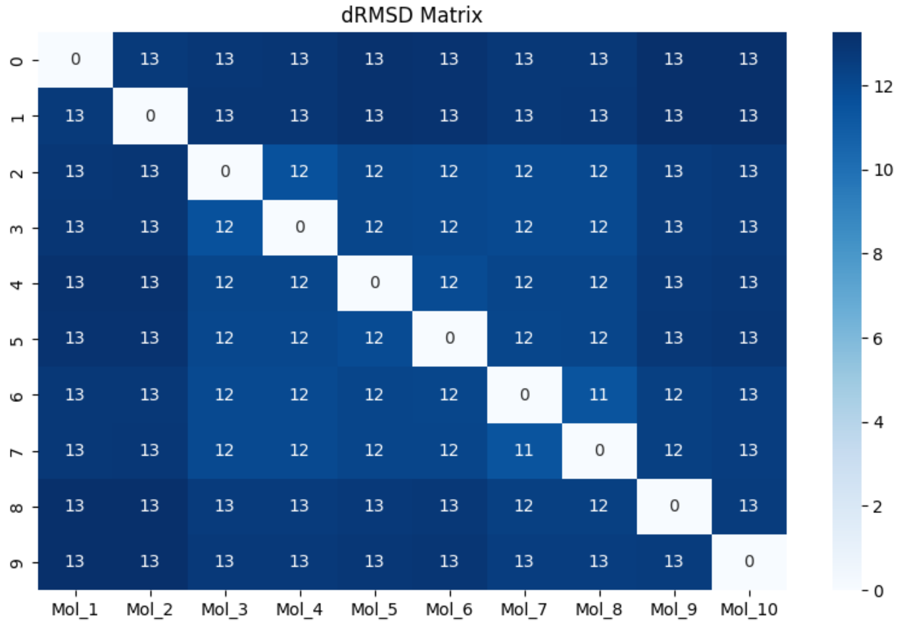
For two conformation matrices it calculates the distance of their coordinates and after that it is applied the dRMSD formula.

I applied the above equation between all the conformations and ended up with the following dRMSD matrix 2.2.

The time to find the dRMSD matrix was clearly longer, 84.52876996994019 sec. I also concluded that the molecule that has the smallest distance from the others is the 7th molecule (so they have don't the same centroid) with the average dRMSD value equals to 11.031283337213589.

Also it is true that

$$\frac{cRMSD}{\sqrt{n}} \leq dRMSD \leq 2cRMSD$$



	Mol_1	Mol_2	Mol_3	Mol_4	Mol_5	Mol_6	Mol_7	Mol_8	Mol_9	Mol_10
0	0	13	13	13	13	13	13	13	13	13
1	13	0	13	13	13	13	13	13	13	13
2	13	13	0	12	12	12	12	12	13	13
3	13	13	12	0	12	12	12	12	13	13
4	13	13	12	12	0	12	12	12	13	13
5	13	13	12	12	12	0	12	12	13	13
6	13	13	12	12	12	12	0	11	12	13
7	13	13	12	12	12	12	11	0	12	13
8	13	13	13	13	13	13	12	12	0	13
9	13	13	13	13	13	13	13	13	13	0