Algorithms in Structural Bioinformatics

I.Z. Emiris and E. Chrysina, Assignment 1

Announced March 30, 2023, Deadline: 23/04/23, midnight

Assignments must be submitted on `e-class:Assignments` as a single PDF or ZIP file whose name starts with your last name. All questions must be answered in a single PDF file; further files (e.g. code) may be included in the ZIPfile (along with the PDF).

1. RNA folding

Find all optimal secondary structures of the RNA sequence $AAUACUCCGUUGCAGCAU$ with the following crude energy minimization algorithm. Starting from the slides' algorithm, use the following initialisation:

$$j + 5 > i \implies E(i, j) = 100, \quad i > j,$$

and bond energy: $-4, 0, 4$, for Watson-Crick bonds, GU, and all other possible pairs respectively.

Implement your algorithm in Matlab, R, Python or other convenient system; submit your code. Print the filled-in table E. Draw (by hand) all optimal folds, show the bonds, and each corresponding backtrack path.

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?

3. Distances

Consider 50 Ca atoms starting at A102 of the main protease of SARS-COV-2 given in complex with a peptide-like inhibitor (PDB id: 6LU7). Construct the 51×51 Cayley-Menger matrix B.

1. Compute rank(B); explain why the obtained value is correct.
2. Perturb entries of B by 5% (maintaining symmetry, positive entries, 0's, 1's), then explain the new value of rank(B). Compute Gram matrix G, apply SVD: $G = U\Sigma U^T$. Let S be the diagonal matrix containing the 3 largest singular values of G. Get the 3D coordinates as $\sqrt{S}U^T$, and report the c-RMSD against the original structure.