Creation of an objective function for the RNA folding problem

guillaume.postic@univ-evry.fr

For a given ribonucleotide chain, the RNA folding problem consists in finding the native fold among the astronomically large number of possible conformations. The native fold being the one with the lowest Gibbs free energy, the objective function should be an estimator of this energy.

For this practical course, you will implement 3 Python scripts to:

- train the objective function, using interatomic distance distributions that are computed from a dataset of known (*i.e.*, experimentally determined) 3D structures;
- plot the scoring profiles, *i.e.* the score (or estimated Gibbs free energy) as a function of the interatomic distance;
- use the objective function to evaluate predicted structures from the RNA-Puzzles dataset.

1. Training

The training script should accomplish the following tasks:

- compute interatomic distances from the given dataset of PDB files;
  - only C3' atoms are taken into account;
  - 10 distance distributions, for the 10 base pairs (AA, AU, AC, AG, UU, UC, UG, CC, CG, GG);
  - only "intrachain" distances are considered;
  - only consider residues separated by at least 3 positions on the sequence (*i.e.* residues i and $i+4$, i and $i+5$, etc.);

- compute the observed frequencies: 10 × 20 distances intervals (0 to 20 Å);
- compute the reference frequency (= the "XX" pair);
- compute the log-ratio of the two frequencies;

The observed probability (*i.e.* frequency) of observing two residues i and j separated by a distance bin r is calculated as follows:

$$f_{i,j}^{\mathrm{OBS}}(r) = N_{i,j}(r) \;/\; N_{i,j}$$

200 frequencies

where $N_{i,j}(r)$ is the count of i and j within the distance bin r, and $N_{i,j}$ is the count of i and j for all distance bins.

The reference frequency is the same formula, except that the different residue types (A, U, C, G) are indistinct ("X"):

$$f_{X,X}^{REF}(r) = N_{X,X}(r) \;/\; N_{X,X}$$

20 frequencies
marginal freauency

Finally, the score (pseudo-energy) $\bar{u}_{i,j}(r)$ is computed as follows:

$$\bar{u}_{i,j}(r) = -\log\left(f_{i,j}^{\mathrm{OBS}}(r) \;/\; f_{i,j}^{\mathrm{REF}}(r) \right)$$

1 scoring table for one

10 scoring table, each table only have one column

The training script should, therefore, generate 10 files of 20 lines (1 line = 1 scoring value).

Note: the maximum scoring value will be arbitrarily set to 10.

The second script will plot the interaction profiles (with R, ggplot, etc.): the score as a function of the distance.

10

display the scoring profiles
20 points
10 files

2. Scoring

The third script will be partially similar to the first one, as it will compute all the distances for a given structure (same thresholds: 20 Å and i, $i+4$). For each distance, a scoring value will be computed, using a linear interpolation. By summing all these scores, the script will calculate the estimated Gibbs free energy of the evaluated RNA conformation.

Rating criteria

Your project will be rated based on the functionality of your scripts ("does it work?") and on the respect of the following good practices:

- non-redundancy of the code (by defining functions and/or classes);
- documentation (README, "--help" option, etc.)
- versioning with `git`: your code must be uploaded on a GitHub or GitLab repository.