Отчёт о научно-исследовательской работе

Панченко С.

1 7-ой семестр, краткий обзор результатов

- По итогам работы за семестр были изучены несколько статей, посвященных исследованию разнообразных структур белков.
- Также было проведено исследование в области непараметрического оценивания плотности распределения случайной величины (ядерного сглаживания) и получены соответствующие результаты на искусственных данных. (TODO: обобщить код для многомерного распределения, протестировать на реальных данных, организовать результаты в виде депозитория на github)

Ссылки на релевантные статьи:

- https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0094334
- https://www.ncbi.nlm.nih.gov/pubmed/30414728
- https://www.ncbi.nlm.nih.gov/pubmed/30036062

2 TODO

Probabilistic metric space to model orientation dependent states (for fast protein and loop modelling)

The project goal is to rank the set of reference proteins, generated from the same set of amino-acids, according to the statistical similarity of the observed protein structure to the structures of reference proteins from the PDB.

Briefly, one has to assess does an observed protein looks like a typical reference protein or not, according to some criterion, which are not constructed yet. We have to construct these criterions and to propose the assessment method.

The basic element of the methodology is the probability distribution of the mutual orientations between two amino-acids an amino-acid a small molecule. Here the first case follows.

Denote by $\mathcal{A}^2 \ni (a,b)$ the alphabet all unsorted pairs of amino-acids, 20 amino-acids make 210 pairs. Denote by $p^{a,b}(\omega,r)$ the (joint, conditional, marginal: see Problems) probability distribution of the angular orientation ω given the distance r. Denote by $d = D_K L(p,p')$ the distance between two distributions, 210 pairs of amino-acids make 21945 distances. Let $d \sim \mathcal{N}(d_{\text{ref}}, \Sigma_{\text{obs}})$. So this distance vector produce the metric space to rank proteins according to the quadratic form

$$E = (\mathbf{d}_{\text{obs}} - \mathbf{d}_{\text{ref}})^{\mathsf{T}})\Sigma_{\text{ref}}^{-1}().$$

how far an observed protein from the nearest stable reference subset from the PDB. The reference subset is extracted using the collaborative filtering (topic modeling). The cartesian product is formed by the sets A^2 and proteins, described by the vector p.

Urgent to check: The distribution $p(\omega,r)$ must tend to uniform when r tends to inf. In fact after $r > r^*$.

Problems to be resolved: the distance r is not a random variable analysis of the energy distribution function

Analysis: This problem statement answers to the following questions How to avoid unnecessary aggregation making the protein description? How to include pairs of couples in the model? How to ensure instability of the reference protein description?

Both approaches consider the great variability of proteins.