Ways to optimize metric properties of protein structure descriptors

Røgen, Peter

Publication date:
2012

Document Version
Publisher's PDF, also known as Version of record

Ways to optimize metric properties of protein structure descriptors

P. Røgen [1]
[1] *Technical University of Denmark, Department of Mathematics, Kgs. Lyngby, Denmark*

Background and aim: The number of known protein and RNA 3d-structures grows very fast. As the number of structure pairs grows faster than computer power in time - it gets harder and harder to use pair-wise similarity measures to study the known structures. An alternative is to use structural descriptors whose main calculation time is linear in the number of structures. The aim of this work is to optimize the Euclidean metric used in the descriptor space to locally be close to RMSD without destroying the descriptors superior ability to give large distances between folds.

Method: Relatively low dimensional geometric descriptor vectors are sufficient to recognize protein (1, 2) and RNA (3) folds. In these works a descriptor vector v_i is (pre)-calculated for each chain molecule M_i and the standard Euclidean distance $\|v_i - v_j\|$ is used as a similarity measure (in fact a pseudo metric) between the original chain molecules M_i and M_j. The descriptor space is Euclidean making all-against-all comparisons, automatic classification (1) and clustering (4) very fast and efficient. The goal of this work is to maintain the benefits of a Euclidean descriptor vector space and to choose the optimal Euclidean metric $\|v_i - v_j\|_Q = \sqrt{(v_i - v_j)^t Q (v_i - v_j)}$ in it. Here Q is a positive semi definite matrix.

For small structural deformations of one chain, most similarity measures are highly correlated with RMSD. Therefore Problem 1 is: If protein structure M_i is a smaller deformation of structure M_j we want to have an isometric representation locally, i.e., $\|v_i - v_j\|_Q \cong RMSD(M_i, M_j)$.

Problem 2: As Problem 1 and if proteins M_i and M_j belong to different folds we want fold separation $\|v_i - v_j\|_Q > RMSD(M_i, M_j)$.

Problem 3: As Problem 2 but also optimizing the automatic classification procedure presented in (1).

Results: We present an optimal way to formulate Problems 1-3 as linear semi definite optimization problems. On protein data our results are: We can improve the local metric properties of the descriptor space while maintaining the ability to automatically classify close to the 96% of all chains reported in (1). Furthermore, the dimension of the resulting descriptor space drops such that less information is needed to be stored.

1. Røgen, P.; Fain, B. Automatic classification of protein structure by using Gauss integrals, *P. Natio. Acad. Sci.*, **2003**, 100(1), 119-124,
2. Røgen, P.; Karlsson, P. W. Parabolic section and distance excess of a space curve applied to protein structure classification. *Geon. Dedicata* **2008**, 134, 91-107.
3. Kirillova, S.; Tosatto, S. C.; Carugo, O. FRASS: the web-server for RNA structural comparison. *BMC Bioinformatics* **2010,** 11, 327.
4. Harder T.; Borg, M.; Boomsma, W.; Røgen, P.; Hamelryck, T. Clustering very large amounts of protein structures using Gauss Integrals, *In preparation*