

# Ways to optimize metric properties of protein structure descriptors

DTU Mathematics  
Department of Mathematics

Peter Røgen, Technical University of Denmark Department of Mathematics  
Peter.Roegen@mat.dtu.dk

In the future pair-wise similarity measures gets harder to use as the number of known protein or RNA structure **pairs** grows faster than computer power in time.

Alternatively the calculation time of structural descriptors<sup>1,2</sup> is linear in the number of structures and offers today fast database scans<sup>3</sup> and clustering<sup>4</sup>.

We optimize the descriptor pseudo-metric to locally be close to RMSD without destroying its superior ability to separate folds.

**Optimizing local metric:** For similar molecular structures  $S_i \sim S_j$  with descriptor vectors  $v_i$  and  $v_j$  we ideally want  $\|v_i - v_j\| = RMSD(S_i, S_j)$  and search the linear transformation of the descriptor space that gets us as close to this goal as possible.

Equivalently, minimize

$$\sum_{S_i \sim S_j} \left| \left( (v_i - v_j)^t Q (v_i - v_j) \right)^{\frac{1}{2}} - RMSD(S_i, S_j) \right|$$

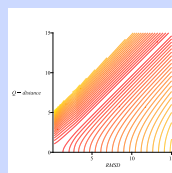
where the scalar product is given by a symmetric positive semi-definite matrix  $Q = Q^t \geq 0$ . Which is not solvable.

$$\min \sum \frac{|(v_i - v_j)^t Q (v_i - v_j) - RMSD^2(S_i, S_j)|}{f(RMSD)}$$

s.t.  $Q = Q^t \geq 0$   
and  $f(x)$  minimizes  $\int_{\Omega} (|x-y| - \frac{|x^2-y^2|}{f(x)})^2 dx dy$   
on  $\mathbb{R}^n$  around  $x=y$ .

Instead we solve an optimal  $Q$ -linearization of the problem.

**Results:** The object functions contour lines are close to the desired. **High correlation between RMSD and  $Q$ -distance** are obtained but also short inter-cluster distances.



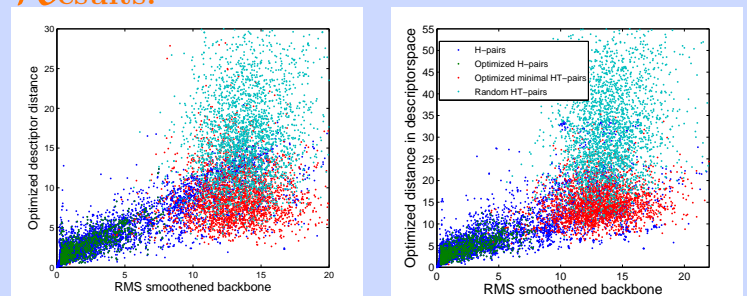
## Penalizing short inter-cluster distances.

**A:** For each CAHT2.4 homology class find the  $Q$ -nearest distinct topology class. We call this a HT-pair and add

$$\sum_{HT-pairs} \max \left( 0, \frac{RMSD(S_i, S_j) - (v_i - v_j)^t Q (v_i - v_j)}{f(RMSD(S_i, S_j))} \right)$$

to the objective function and **B:** Find a new optimal  $Q$ . **A:** and **B:** are repeated until no new HT-pairs are found.

## Results:

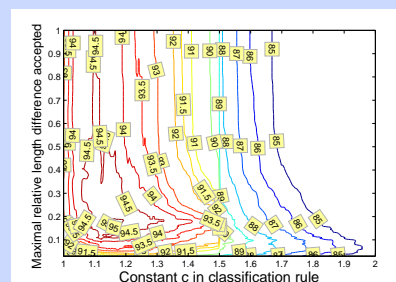


Left (see legend to the right): After optimizing distances between similar structures short minimal HT  $Q$ -distances are found. Right, penalizing these gives **better inter cluster  $Q$ -distances**.

## Optimizing automatic classification.

CATH domains are automatically classified<sup>2</sup> as follows: For each domain D find the  $Q$ -nearest and next-nearest clusters at  $Q$ -distances  $d_1 \leq d_2$ . If  $d_1 * c < d_2$  for a given constant  $c > 1$  domain D is classified as the nearest cluster else D is unclassified or a new cluster. Starting from the previous optimal  $Q$  un- and mis-classified structures are  $Q$ -linearly penalized in the objective function. A new optimal  $Q$  is found and this is repeated until no new un- and mis-classified structures are found.

## Results:



The **high automatic classification success rate** is shown as function of the classification constant  $c$  and the maximal accepted relative domain length difference.

**Descriptor dimension is reduced** in all cases from 48 to app. 10, 15 and 20 respectively.

(1) P. Røgen & P.W. Karlsson, Parabolic section and distance excess .. applied to protein structure classification, Geometriae Dedicata, 134(1), 91-107, 2008.  
(2) P. Røgen & B. Fain, Automatic classification of protein structures by using Gauss Integrals, PNAS, 100, 119-124, 2003.  
(3) S. Kirillova, S.C. Tosatto & O. Carugo, FRASS: the web-server for RNA structural comparison. BMC Bioinformatics, 11:327, 2010.  
(4) T. Harder et.al. Fast large-scale clustering of protein structures using Gauss integrals. In preparation.