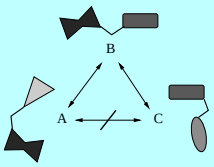
Automatic Classification of Protein Structures Using Gauss Integrals

Peter Røgen, Department of Mathematics & Quantum Protein Center (QUP),
Technical University of Denmark, Peter.Roegen@mat.dtu.dk.
Boris Fain, Department of Structural Biology, Stanford University.

Introduction: Structure of biological molecules is a very important clue to understanding and manipulating biological function.



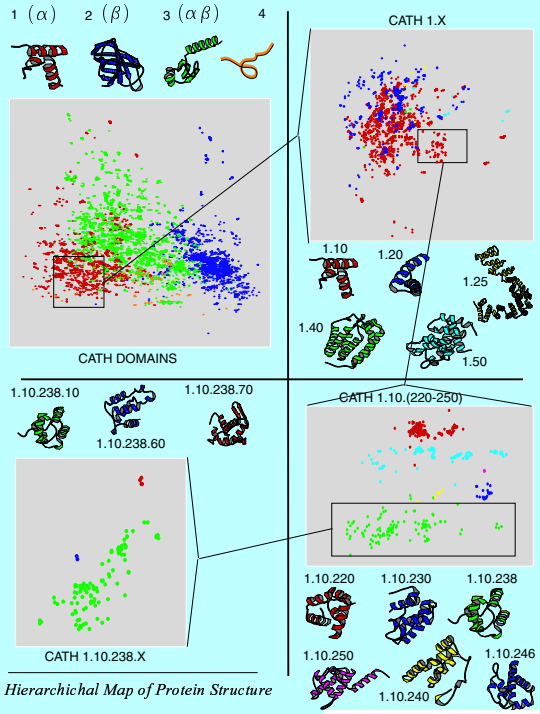
Current subset-matching structural measures fail to satisfy the metric conditions, in particular, the triangle inequality.

Method: For each protein structure we compute 30 topological invariants (Gauss Integrals) of the polygonal curve connecting the C_α atoms.

Once each polypeptide is mapped onto a point in $\mathbb{R}^{30}$, we use the usual euclidean metric

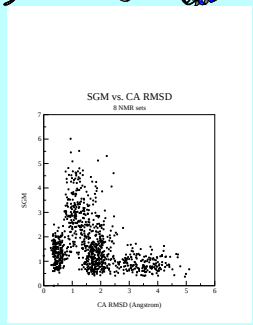
$$d(x,y) = \sqrt{\sum_{i=1}^{30}(x_i - y_i)^2}$$

to compare chains. We call this (pseudo) metric SGM, the Scaled Gauss Metric.



Hierarchichal Map of Protein Structure

SGM vs. RMSD: SGM and Carbon alpha RMSD are poorly correlated (R=0.2) on sets of NRM structures where alignment is unambiguous.

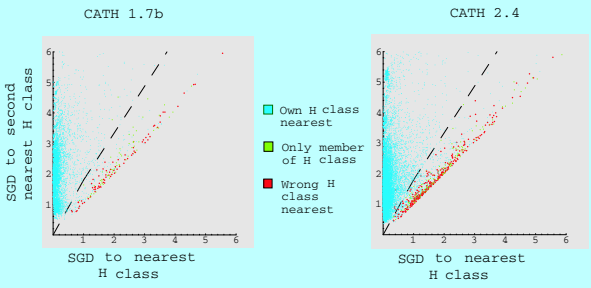


Separation: The figure "*Hierarchical Map of Protein Structure*" shows a clear separation of the classes of the CATH protein structure classification system even in 2D-projections of $\mathbb{R}^{30}$. The upper left corner contains 9955 CATH DOMAINS coloured according to their Class membership. CATH 1.X enlarge CATH class 1 and its five constituent architectures. Next, topologies 1.10.(220 through 250) are displayed in larger detail and finally the clearly separated H-categories of topology 1.10.238 are shown.

The algorithm: Given a chain $\mathcal{C}$ we compute the 30 invariants and find the nearest and the second nearest clusters $\mathcal{N}_1$ and $\mathcal{N}_2$ at distances D_1 and D_2.

If $D2 \geq 1.75D1$ then $\mathcal{C}$ lies inside a populated region of $\mathcal{N}_1$'s cluster and joins it.

If $D2 < 1.75D_1$ then $\mathcal{C}$ is fairly close to two different clusters or is far away from both. $\mathcal{C}$'s classification is declared *unknown*.



This figure illustrates classification into the first 4 CATH categories (C.A.T.H). The plots display D_1 vs. D_2 for all chains in CATH1.7b (left) and CATH2.4 (right). The decision boundary is given by one adjustable parameter, $D2 = 1.75D1$, and is shown as a dashed line. The sub-region with the red and green points is the *unknown* region in the classification.

Beyond 96% success

- 95.51% (19996/20937) of CATH 2.4 domains are assigned appropriate C, A, T, and H designation.
- 0.82% or 171 domains are correctly classified *unknown* as each of them is a solitary member of there class. All new folds are flagged *unknown*.
- 3.65% or 765 of the chains cannot be assigned an H designation.
- 0.02% or 5 are assigned wrong T or H designation.

Speed: This test involves all against all comparison of 20937 CATH 2.4 domains and takes 2 CPU-hours.

Conclusion:

- The algorithm itself is a useful tool that will speed up the process of classification and save expert human judgment for the most interesting cases.
- More importantly, it is our hope that the very simplicity of this algorithm demonstrates the power of treating protein structure via our measures.
- In the future we hope to see the development of many very fast and comprehensive algorithms using Gauss Integrals.