Proc. German Conference on Bioinformatics (GCB'98), Köln (Cologne), Oct. 1998.

Similarity Search in 3D Protein Databases

Gabi Kastenmüller, Hans-Peter Kriegel, Thomas Seidl
Institute for Computer Science, University of Munich, Oettingenstr. 67, D-80538 München, Germany
contact: kriegel@dbs.informatik.uni-muenchen.de, phone: ++49-89-2178-2191, fax: ++49-89-2178-2192

1 Introduction

We introduce a new approach for similarity search in 3-D protein databases. By using histograms, we define adaptable similarity models that address the 3-D shape as well as chemical properties of proteins. Quadratic forms are employed as similarity distance functions for which efficient query processing algorithms are available [Sei 97]. Experimental examples illustrate the applicability of the method.

2 Shape Histograms of 3-D Objects

For a histogram-based similarity model, we have to decompose the 3-D space into appropriate cells. These cells were used as bins of the histograms. In our primary approach, we distinguish three techniques for the decomposition of the 3-D space. Figure 1 illustrates these models in a 2-D projection:

- **Shell Model:** Decomposition into concentric shells around the center point.
- **Sector Model:** Decomposition into sectors that emerge from the center point of the model.
- **Net-Shaped Model:** Decomposition obtained from a combination of shell and sector model.

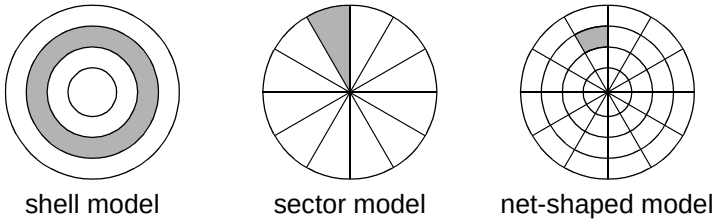

Figure 1: Shape histograms based on shells, sectors, and combinations of them.

We assume our 3-D objects to be given as sets of points [SK 95]. The histograms are determined by counting the number of points within each cell. In figure 2, we illustrate various shape histograms for the example protein, 1SER-B. In the middle, the space decomposition is indicated schematically, and on the right hand side, the corresponding shape histograms are depicted. The top histogram is

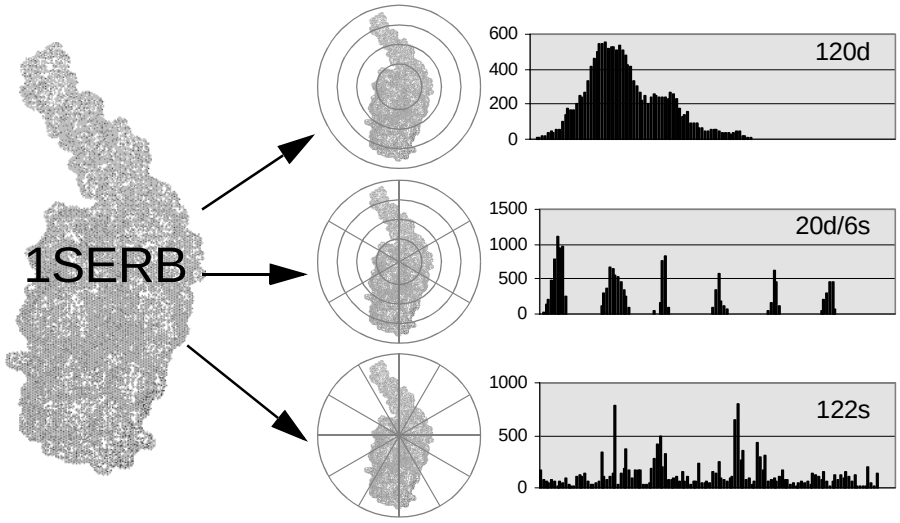

Figure 2: Several 3-D shape histograms of the example protein 1SERB.

purely based on shell bins (denoted as 120d), and the bottom histogram is defined by 122 sector bins (denoted as 122s). The histogram in the middle follows the combined model, it is defined by 20 shell bins and 6 sector bins (20d/6s). In this example, all the different histograms have approximately the same dimension of 120.

3 Combined Histogram Model

On top of the geometric information, a lot of thematic information belongs to spatial objects. Particularly in protein databases, the chemical structure and physical-chemical properties are important. Examples include atom types, residue types, partial charge, hydrophobicity, electrostatic potential among others. A general approach to manage thematic information along with spatial properties is provided by combined histograms. Figure 3 demonstrates the basic principle. Assume a spatial histogram structure as presented in the foregoing, and a thematic histogram structure to be given. A combined histogram structure is immediately obtained as the Cartesian product of the original structures. [Kas 98].

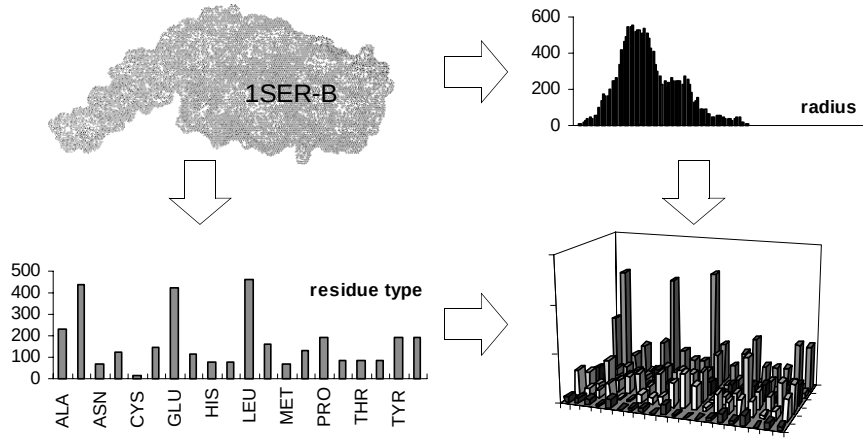

Figure 3: Example for a combined thematic and shape histogram for a molecule

4 Similarity Distance Functions

The shape histogram approach is a typical example of a feature-based similarity model for complex objects. Therefore, common definitions of similarity distance function also apply to histograms when considered as feature vectors. The standard example is the Euclidean distance of vectors which, however, owns severe shortcomings. In particular, the individual components of the feature vectors are assumed to be independent from each other, and no relationships of the components such as substitutability and compensability may be regarded.

By using a quadratic form distance function as an adaptable similarity function, the problems of the Euclidean distance may be overcome. The neighborhood of bins in general and of shells or sectors in particular may be represented as similarity weights in the similarity matrix. For similarity search based on quadratic form distance functions, efficient query processing algorithms are available even for high-dimensional histograms [Sei 97].

5 Experimental Evaluation

For our experiments, we used the example protein 1SER-B from above, a Seryl-tRNA Synthetase molecule from the PDB [Ber+ 77]. From the protein classification system CATH [Ore+ 97], we obtained information on structurally similar molecules, and found the family of Seryl-tRNA Synthetases which we illustrate in figure 4. Obviously, from a visual inspection, the strong shape similarity of the molecules can be observed.

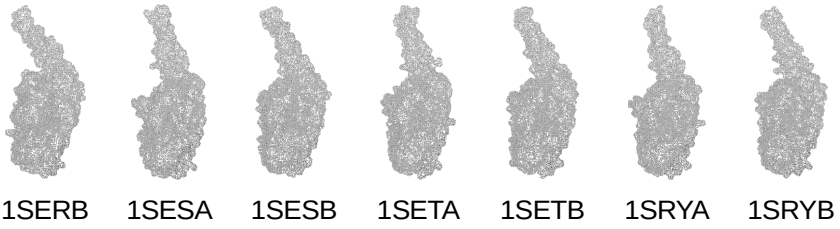

Figure 4: Family of similar molecules: Seryl-tRNA Synthetases.

For the shape histogram structures 120d, 6d/20s, and 122s, we performed similarity queries on our 3-D protein database of 5,000 molecules from the PDB. The query processing took only a few seconds of runtime. Figure 5 demonstrates the resulting rankings for the identity matrix. The diagrams depict the respective similarity distances in ascending order. For all of the three histogram structures, the seven members of the Seryl family rank on the top seven positions among the 5,000 molecules of the database. In particular, the similarity distance noticeable increases for the first non-Seryl protein in the ranking order. To provide an illustration of this effect, these most similar non-Seryl molecules are depicted on the right of each ranking diagram.

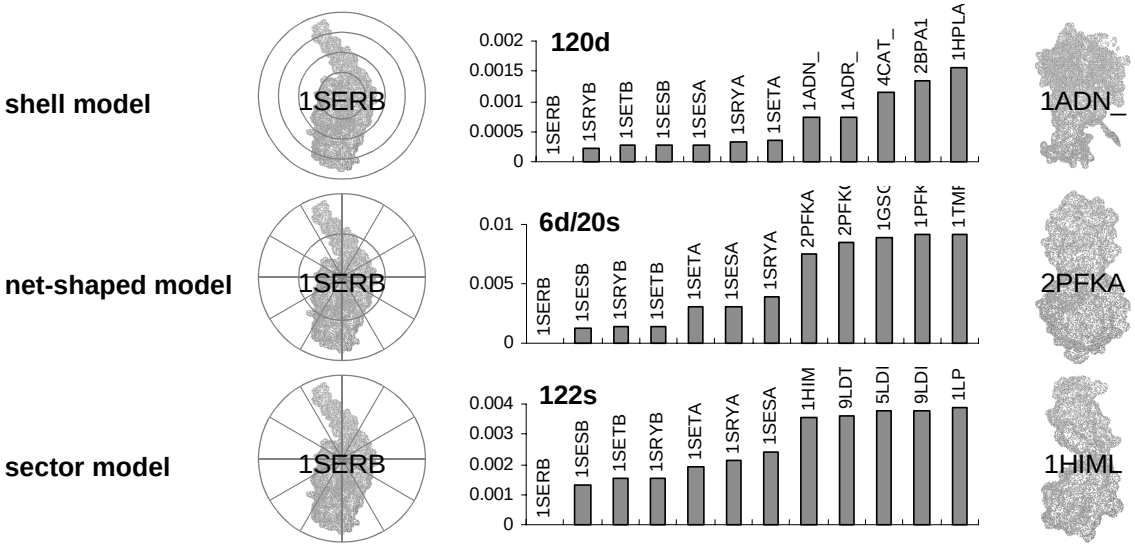

Figure 5: Similarity ranking for various histogram structures. On the left, the query protein 1SERB is shown. The diagrams depict the similarity distances of the 12 nearest neighbors in ascending order.

References

[Ber+ 77] Bernstein F. C., Koetzle T. F., Williams G. J., Meyer E. F., Brice M. D., Rodgers J. R., Kennard O., Shimanovichi T., Tasumi M.: *'The Protein Data Bank: a Computer-based Archival File for Macromolecular Structures'*, Journal of Molecular Biology, Vol. 112, 1977, pp. 535-542.

[Kas 98] Kastenmüller G.: *'Shape-oriented Similarity Search in 3D Protein Database Systems'*, Diploma Thesis, Institute for Computer Science, University of Munich, 1998. (in German)

[Ore+ 97] Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., Thornton, J.M.: *'CATH – A Hierarchic Classification of Protein Domain Structures'*, Structure, Vol. 5, No. 8, 1997, pp 1093-1108.

[Sei 97] Seidl T.: *'Adaptable Similarity Search in 3-D Spatial Database Systems'*, Ph.D. Thesis, Faculty for Mathematics and Computer Science, University of Munich, 1997. Herbert Utz Publishers, Munich.

[SK 95] Seidl T., Kriegel H.-P.: *'Solvent Accessible Surface Representation in a Database System for Protein Docking'*, Proc. 3rd Int. Conf. on Intelligent Systems for Molecular Biology (ISMB), 1995, pp.350-358