

Shape Distributions and Protein Similarity*

Stefan Canzar¹ and Jan Remy²

¹ Université Henri Poincaré
LORIA, B.P. 239
54506 Vandœuvre-lès-Nancy, France
canzar@loria.fr

² Institut für Theoretische Informatik
ETH Zürich
CH-8092 Zürich
jremy@inf.ethz.ch

Abstract: In this paper we describe a similarity model that provides the objective basis for clustering proteins of similar structure. More specifically, we consider the following variant of the protein-protein similarity problem: We want to find proteins in a large database \mathcal{D} that are very similar to a given query protein in terms of geometric shape. We give experimental evidence, that the shape similarity model of Osada, Funkhouser, Chazelle and Dobkin [OFCD02] can be transferred to the context of protein structure comparison. This model is very simple and leads to algorithms that have attractive space requirements and running times. For example, it took 0.39 seconds to retrieve the eight members of the seryl family out of 26,600 domains. Furthermore, a very high agreement with one of the most popular classification schemes proved the significance of our simplified representation of complex proteins structure by a distribution of C_α - C_α distances.

1 Introduction

Understanding the rapidly increasing number of protein three-dimensional structure data deposited in the Brookhaven Protein Data Bank (PDB) [BWF⁺00] poses a major challenge in the post-genome-sequence era. One reliable method to assign function to gene products that have no experimentally inferable molecular (biophysical or biochemical) function is on the basis of sequence similarity to proteins of known function. Since structure is evolutionary better conserved than sequence, the structural similarity to one or more proteins of known structures infers an even more powerful clue to the structure-function relationship. Clearly, the classification of recurrent protein folds constitutes a major step towards the understanding of protein structure.

*This paper includes work done while the authors were at Technische Universität München, Institut für Informatik. Research was partially supported by the DFG project KN 309/1-1 "Information Mining".

The placement in categories must be done according to a similarity criterion or distance (metric) that reflects the degree of shape affinity for pairs of proteins. The most popular classification systems either use a totally automated approach (FSSP) [HS97], classify manually (SCOP) [MBHC95] or are based on a combination of both (CATH) [OMJ⁺97]. The three-dimensional structures are usually compared by structural alignment algorithms such as CE [SB98], DALI [HS93], and VAST [MGB95], which is, mainly because of its intrinsic complexity, a time-consuming task.

Problem Statement. We consider a special variant of the molecular similarity problem. Let \mathcal{D} be a database containing a collection of proteins. We want to find the proteins in \mathcal{D} that are similar to a given query protein Q . There is no common definition of what “similarity of proteins” exactly means. As motivated above, we restrict ourselves to the similarity of three-dimensional structure. This kind of similarity is very “human oriented”, since two objects - or in our case proteins - are usually said to be similar if a human observer thinks that they are. Thus, we have two criteria for performance: *i*) if $Q \in \mathcal{D}$ then Q should be recognized as the most similar and *ii*) molecules rated as very similar to Q should be also recognized by a human as being very similar. Note that the second criterion does not include the first. If the shape of Q is not very characteristic, it could be difficult for a human to recognize an identical structure. Since the database \mathcal{D} contains usually thousands of proteins (the PDB contains currently 32,823 structures) it is important that the comparison of a single pair of proteins is very fast. This usually requires some preprocessing of the database. It is desirable that the data structures produced during preprocessing have modest space consumptions.

Related Work. Geometric approaches to measure the similarity of proteins were extensively studied in various aspects. In order to give a representative selection, we like to mention geometric hashing [Wol90, NW91, FNW92, FNNW93, NLWN95], footprinting [BS97, BS99] and correlation techniques [KKSE⁺92, GJS97]. None of these algorithms has a running time that allows fast queries to a large database. Methods that do not depend on a structural alignment are based on graph theory [HPM⁺02], local feature profiles of C_α distance matrices [CKK04], C_α - C_α distances [CP02] or secondary structure matching [KH04]. Special algorithms for similarity search in protein database were considered by Kriegl and Seidl [KS98] and Ankerst, Kastenmüller, Kriegl and Seidl [AKKS99]. The first approach is based on parametric approximation of surface segments. In the second paper, proteins are described by density histograms that are robust under rotation.

Our Results. The concept of shape distributions was introduced by Osada, Funkhouser, Chazelle and Dobkin [OFCD02]. They evaluated their approach by comparing simple objects like cars, humans, phones or mugs. We have successfully transferred their similarity model to the protein similarity context. The main purpose of our work is to evaluate whether shape distributions are suitable means to compare the three-dimensional structure of proteins or molecules. Our experiments give evidence that the performance criteria mentioned above are satisfied: The protein in the database with the most similar shape distribution was always the query protein itself. Furthermore, top ranked proteins could be observed to be structural similar to the query protein. The ability to distinguish CATH homologous superfamilies with a success rate of 98% confirmed this subjective evaluation.

We claim that this algorithm has some advantages compared to previous methods. First, the comparison step is fast enough for database search, since we are able to make around 100,000 comparisons per second. Second, the algorithm is much more simple than most of the other approaches. Third, the space requirement of the data structure we generate in a preprocessing step is only linear in the number of proteins contained in the database. And fourth, our approach does not depend on any knowledge-based decisions, like the assignment of secondary-structure elements.

The remainder of the paper is organized as follows. In section 2 we review the concept of shape distributions. In section 3 we introduce the algorithm for similarity search. Finally, section 4 presents experimental results.

2 Shape Distributions

Osada, Funkhouser, Chazelle and Dobkin [OFCD02] introduced a simple model for shape similarity of objects. Let \mathcal{S} be a set of points on the surface of an object. A *shape function* $\xi(\mathcal{S})$ measures a geometric property that depends on \mathcal{S} . A typical example for a shape function is the Euclidean distance $d(a, b)$ for $\mathcal{S} = \{a, b\}$. Other types of shape functions include angles, areas or volumes.

If \mathcal{S} is chosen at random from all points on the surface of the object, then $\xi(\mathcal{S})$ is a random variable having some distribution $F(\xi(\mathcal{S}))$. Osada et. al. claim that this distribution, the *shape distribution*, is very characteristic for the shape of the object. Thus the shape matching problem can be reduced to the comparison of two probability distributions. The algorithmic side of shape distributions is very simple. For the sake of exposition, we assume that our shape function is the Euclidean distance of two points. As mentioned above, the distance of two random (surface) points is a random variable. The distribution of distances is reconstructed by choosing N pairs of surface points at random. Of course, for technical reasons, the distribution must be discretized into, say B many intervals. In essence, by counting the number of distances that fall into each interval, we obtain a histogram that consists of B bins that expresses the “probability” for a distance being within some interval. The similarity (or dissimilarity) of two objects can be computed by comparing their shape distribution, i.e., the histograms under an arbitrary metric. The most natural example is the Minkowski norm \mathcal{L}_N .

3 The Algorithm

In this section we give an overview of the algorithm. The input is a set \mathcal{D} of 3D protein structures. The atomic coordinates are taken from the Brookhaven Protein Data Bank (PDB) [BWF⁺00]. In our experiments we varied the definition of the point set \mathcal{S} (cf. Section 2) to contain either all atoms, exclusively atoms located on the molecular surface or all C_α atoms. We have chosen the Euclidean distance as a shape function $\xi(\mathcal{S})$, since it seems to provide the best results.

Preprocessing The preprocessing is identical for each protein in \mathcal{D} and only depends on the definition of \mathcal{S} . First we extract the coordinates of points in \mathcal{S} , which is a trivial step

in the case of \mathcal{S} being equal to the set of all C_α atoms. To derive the shape distribution from the surface of the protein we determine the atoms that can be touched by a solvent molecule of fixed size (e.g. 1.4 Å). This can be done with an algorithm of Sanner, Olsen and Spohner [SOS96] in $\mathcal{O}(n \log n)$ time. Simply speaking, this algorithm computes the surface atoms as an intermediate result. Second we calculate the distances of each pair of atoms in \mathcal{S} . This yields a histogram with B bins each counting the number of occurrences of certain distances. By a normalization of the resulting shape distribution one could simply add an invariance under scaling, e.g. consider the shape of proteins independent of their size. Second we store the (not normalized) histogram as a sequence of B integers. The preprocessing of a protein with n atoms requires optionally time $\mathcal{O}(n \log n)$ for the computation of the surface plus time $\mathcal{O}(n^2)$ for the approximation of the shape distribution. The overall complexity can be reduced to $\mathcal{O}(n \log n)$ if we consider only $\mathcal{O}(n \log n)$ random pairs in \mathcal{S} for the computation of the shape distribution.

Similarity Query Let Q denote the query protein. We compute the similarity measure between Q and each structure in \mathcal{D} by comparing their shape distributions. We experimented with similarity measures based on the Minkowski \mathcal{L}_N norms for $N = 1, 2, 10$.

It remains to discuss the complexity of the similarity query. The distance of two distributions f and g in the Minkowski norm is given by

$$D(f, g) = \left(\sum_{i=1}^B |f_i - g_i|^N \right)^{1/N} \quad (1)$$

In fact, this value is the distance of two points, f and g in the \mathbb{R}^B under the \mathcal{L}_N metric. Furthermore, the histograms of the proteins in \mathcal{D} may be modeled as a set of points in a high dimensional space with coordinates determined by the approximations of the shape distributions. Also, the shape distribution of the query protein defines a point in the \mathbb{R}^B . Hence the similarity problem for proteins can be transformed into proximity problem among a set of points. This transformation is very helpful, as there are algorithms for proximity problems that have desirable asymptotics.

We want to query the database for the most similar proteins, i.e., proteins with scores that are lesser than a given threshold. So we have to solve a proximity problem which is known as *range searching*. There are data structures that provide fast queries for orthogonal search regions in spaces, provided the dimension is small. In our case, the search region is circular and $d = B$ is usually very large. Unfortunately there are no fast data structures for circular queries in high dimensional spaces. However, Arya and Mount [AMN⁺94] proposed a data structure that allows queries with circular ranges if one is willing to accept some approximation. More precisely their data structure ensure that the following is true for all $\varepsilon > 0$. Let t denote the given threshold, i.e., diameter of the query range. Then points lying within distance $\varepsilon \cdot t$ around the boundary of the query range either may or may not be included in the output of the query. The running time of such a query is $\mathcal{O}((1/\varepsilon)^d + \log m)$ and it is also good in practice as Arya and Mount claim in their paper.

4 Experimental Results

We have implemented the algorithm described in section 3 in C++. The experiments were done on a system with a 1,60 GHz Pentium M-Processor. The three-dimensional coordinate data was taken from the Brookhaven Protein Data Bank (PDB) [BWF⁺00] and was dissected into domains according to CATH version v2.0. The resulting collection \mathcal{D} of protein structures (about 26,600 CATH domains) was preprocessed into shape distributions and finally stored on disk.

In contrast to [OFCD02], both the restriction to atoms on the molecular surface and the random sampling of \mathcal{S} means a loss in characteristics of shape distribution for the complex structure of proteins. In contrast, the difference in classification accuracy depending on whether \mathcal{S} contained all atoms or only the subset of C_α atoms was marginal. To shorten computation time we thus focused on the latter case which we will discuss now.

It turned out during the experiments, that using $B = 60$ bins for the representation of the shape distribution and the Minkowski \mathcal{L}_2 -norm for measuring the dissimilarity between pairs of distributions is a good choice.

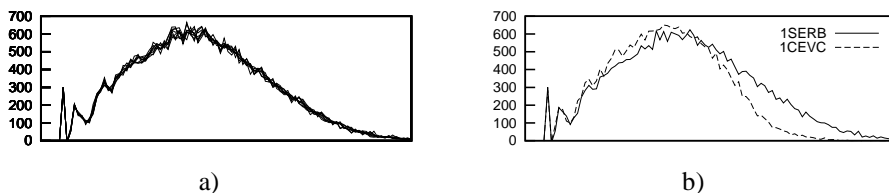


Figure 1: a) The superposed shape distributions of the eight seryl family members. b) With respect to query protein 1SERB domain 1CEVC ranked on position 125. Their distributions can be distinguished visually.

4.1 Basic Similarity Search

In order to demonstrate the general applicability of shape distributions to the characteristic representation of the three-dimensional structure of proteins, we report on experiments on a group of molecules that are known to be related. We tried to retrieve the eight members of the *seryl-tRNA synthetase* family (1SERA, 1SERB, 1SESA, 1SESB, 1SRYA, 1SRYB, 1SETA, 1SETB) out of roughly 26,600 domains contained in our database.

If the query molecule is 1SERB we obtained a ranking as depicted in Figure 2. The eight members of the seryl family rank on the top eight positions, followed by roughly 26,600 molecules. This ranking is conform with the shape of the molecules (Fig. 3). Furthermore, the shape distributions of the seryl family members are clearly distinguishable from those derived from higher ranked domains (Fig. 1). This kind of query could be the first step when searching for structural homologs of a given protein Q . Screening the whole

PDB by using shape distributions could result in a small number of structural homologs of Q (for example by range searching, as mentioned in section 3), which are further analyzed by rigid-body superposition (e.g. May and Johnson [MJ95]) to find the best possible alignment.

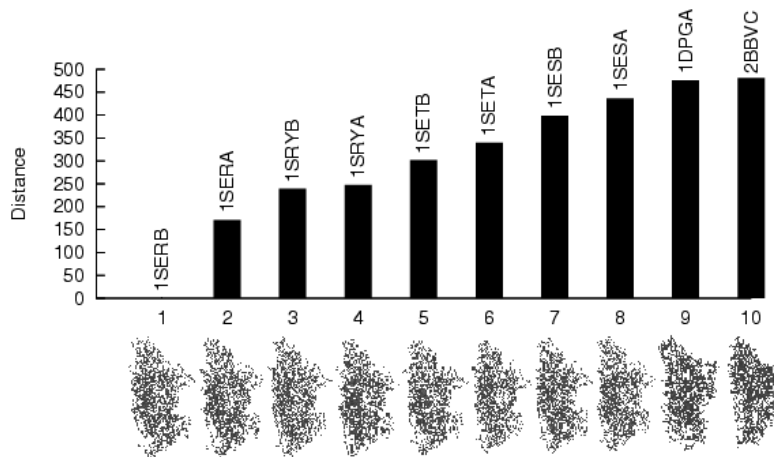


Figure 2: Similarity scores of the most similar molecules to 1SERB. The eight members of the seryl-tRNA-synthetase family rank on the top eight positions among 26,600 domains. The first non-seryl protein 1DPGA is classified by CATH to fall into the same class.

4.2 Classification by Structural Similarity

The placement of protein structures in categories heavily depends on the nature of the underlying similarity model. In order to investigate whether the transformation of protein structures into points in B -dimensional Euclidean space \mathbb{R}^B has a negative impact on the accuracy of classification, we performed an all-against-all comparison according to our distance measures on one of the most popular classification schemes, the CATH database [OMJ⁺97] (353,766,700 structural comparisons). CATH, as a hierarchical classification scheme, clusters protein structures in the PDB at four major levels, Class (C), Architecture (A), Topology (T) and Homologous superfamily (H). Based on our symmetric distance matrix (metric property of our distance measure) we determined the nearest neighbor N for every molecule in the database \mathcal{D} , ignoring the query structure Q itself, for which $d(Q, Q) = 0$ holds for all $Q \in \mathcal{D}$. When asking whether N and Q fall into the same CATH category on level l , $l = 1, 2, \dots, 7$, we considered all those domains, that were labeled identically by CATH on levels $1, \dots, l - 1$.

From domains sharing the first six CATH labels, C, A, T, H, S, and N, 71% have been assigned the correct label on level seven (I) (cf. Table 1). Ascending the hierarchy, this value increases up to 98% at H-level, where the last three labels were allowed to vary. We attach great importance to the high categorization accuracy particularly at this level,

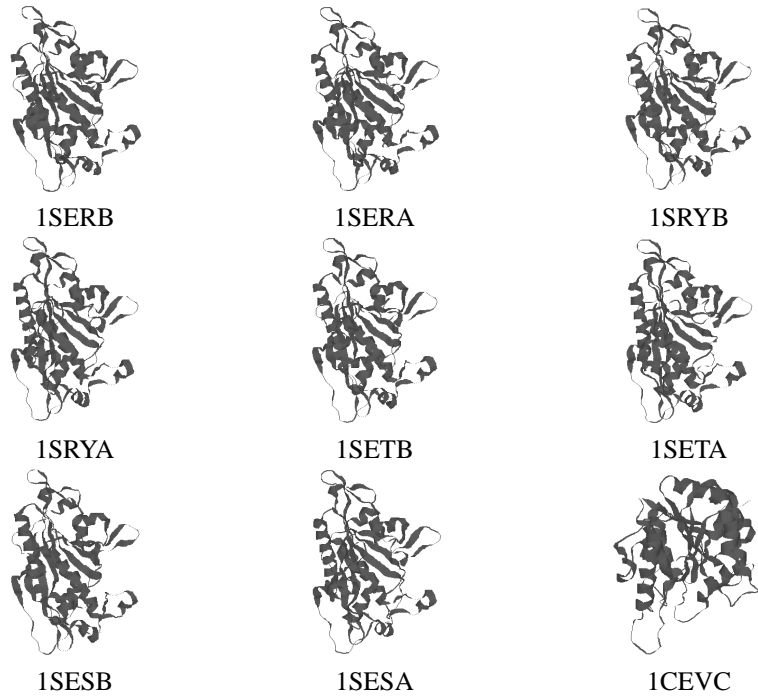


Figure 3: From left to right and top to bottom: the eight seryl family members, which have been rated as being most similar to 1SERB, clearly have similar structures. In contrast, domain 1CEVC (ranked on position 125 with respect to query protein 1SERB) apparently shows different three-dimensional arrangement.

as homologous superfamilies cluster proteins with highly similar structures and functions. Furthermore, we think that distinguishing different architectures with a success rate of 97% is a remarkable result, as label assignment at A-level is based on the human eye.

These features of our similarity measure are further illustrated by the cluster-analysis dendrogram shown in Figure 4. We randomly selected 36 domains from three different nodes on the T-level of the CATH hierarchy (12 domains from each node), where the first node can be described by labels $C=2$, $A=30$ and $T=30$, the second node by $C=3$, $A=10$ and $T=20$ and the third node by $C=1$, $A=10$ and $T=150$. Not only that there was a clear discrimination between these three groups, but one can also associate lower CATH levels with subclusters in the clustering tree. For example, removing the longest edge from the minimum spanning tree of a graph, whose vertices correspond to protein domain structures from the third group ($C=2$, $A=30$, $T=150$) and whose edges are weighted with the distances based on our similarity measure, results in two clusters, one containing the domains labeled $H=20$ and one with domains labeled $H=100$. Similarly this holds for domains sharing labels $C=1$, $A=10$, $T=150$, $H=20$, and $S=1$ and differing in $N=1$ or $N=3$.

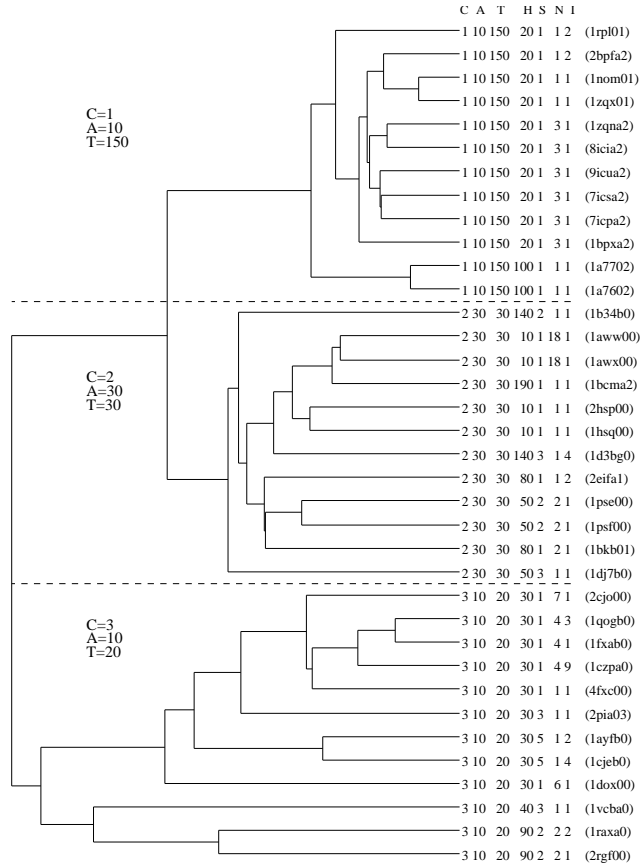


Figure 4: Cluster-analysis dendrogram of randomly selected CATH domains. The shape distributions of the protein domain structures have been clustered by an agglomerative hierarchical algorithm using the single linkage similarity criterion.

4.3 Running Time

Since the shape distributions can be computed in a preprocessing step, we can perform the queries to the database very fast. In section 3 we have mentioned that the query time is roughly $\mathcal{O}(\log m)$ if we assume that ε and B are constant. In practice the constants are too large – at least for “small” databases. Nevertheless, the query time is still attractive. In our implementation, which was not optimized for speed, a query to a database of size $m \approx 26,600$ took only 0.39 seconds, ignoring the time spent on input/output operations. The computation of an all-against-all distance matrix (353,766,700 comparisons) was finished after less than an hour.

CATH LABEL	CATH LEVEL	NEAREST NEIGHBOR AGREEMENT (%)
C	Class	97.1
A	Architecture	97.2
T	Topology	96.0
H	Homologous superfamily	98.0
S	Sequence families	96.8
N	Nearly-identical representatives	91.9
I	Identical representatives	71.5

Table 1: Nearest neighbor classification for CATH categories based on our similarity score. An agreement of $x\%$ on level l describes, that $x\%$ of domains sharing the first $l - 1$ CATH labels have been assigned the correct label on level l .

5 Concluding Remarks

We have given experimental evidence that the distribution of distances between C_α atoms provides a significant signature for the three-dimensional structure of proteins. By transferring the similarity model of Osada, Funkhouser, Chazelle and Dobkin [OFCD02] to the context of protein fold comparison, we were able to retrieve the eight members of the seryl family among 26,600 domains in 0.39 seconds of CPU time. But despite the simplified representation of protein structure, our approach exhibits a classification accuracy of 98% for CATH homologous superfamilies.

Several alternative methods based on a simplified representation of protein structure have been proposed recently. The one of Carugo and Pongor [CP02] considers C_α - C_α distances between residues separated by a variable number of amino acid residues and is thus conceptually related to our approach. Nevertheless, they represent each molecule by a set of 28 histograms that have to be compared by a contingency table analysis. As a consequence, the comparison of a pair of proteins is more expensive both in terms of computation time and space consumption. The similarity score of Choi, Kwon and Kim [CKK04] is based on profiles of representative local features (LFF) of C_α distance matrices. Compared to shape distributions, LFF profiles necessitate an considerable preprocessing step and yield an agreement with CATH categories that ranges from 53.3% on Homology level to 70% on Class level.

In short, no other approach combines comparable high classification accuracy with approximate efficiency both in terms of time and space, while at the same time being independent of any sequence information or human input. These features allow for a quick categorization of recently determined structures by scanning large databases like the PDB and thus help to keep our ordering of the protein fold space always up to date, as opposed to knowledge-based schemes like SCOP and CATH.

References

- [AKKS99] Mihael Ankerst, Gabi Kastenmüller, Hans-Peter Kriegel, and Thomas Seidl. Nearest Neighbor Classification in 3D Protein Databases. In *Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology*, pages 34–43, 1999.
- [AMN⁺94] Sunil Arya, David M. Mount, Nathan S. Netanyahu, Ruth Silverman, and Angela Y. Wu. An optimal algorithm for approximate nearest neighbor searching fixed dimensions. *Journal of the ACM*, 45(6):891–923, 1994.
- [BS97] Gill Barequet and Micha Sharir. Partial Surface and Volume Matching in Three Dimensions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(9):929–948, September 1997.
- [BS99] Gill Barequet and Micha Sharir. Partial Surface Matching by Using Directed Footprints. *Computational Geometry: Theory and Applications*, 12(1–2):45–62, February 1999.
- [BWF⁺00] H.M. Berman, J. Westbrook, Z. Feng, Gilliland G., T.N. Bhat, Weissig H., I.N. Shindyalov, and P.E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, 2000.
- [CKK04] In-Geol Choi, Jaimyoung Kwon, and Sung-Hou Kim. Local feature frequency profile: A method to measure structural similarity in proteins. In *Proceedings of the National Academy of Sciences*, volume 101, pages 3797–3802, March 2004.
- [CP02] O. Carugo and S. Pongor. Protein Fold Similarity Estimated by a Probabilistic Approach Based on C^α-C^α Distance Comparison. *Journal of Molecular Biology*, 315(4):887–898, January 2002.
- [FNNW93] Daniel Fischer, Raquel Norel, Ruth Nussinov, and Haim J. Wolfson. 3-D Docking of Protein Molecules. In *Proceedings of the 4th Annual Symposium on Combinatorial Pattern Matching*, Lecture Notes in Computer Science (684), pages 20–34. Springer, June 1993.
- [FNW92] Daniel Fischer, Ruth Nussinov, and Haim J. Wolfson. 3-D Substructure Matching in Protein Molecules. In *Proceedings of the 3rd Annual Symposium on Combinatorial Pattern Matching*, Lecture Notes in Computer Science (644), pages 136–150. Springer, April/May 1992.
- [GJS97] Henry A. Gabb, Richard M. Jackson, and Michael J.E. Sternberg. Modelling Protein Docking using Shape Complementary, Electrostatics and Biochemical Information. *Journal of Molecular Biology*, 272(1):106–120, September 1997.
- [HPM⁺02] A. Harrison, F. Pearl, R. Mott, J. Thornton, and C. Orengo. Quantifying the similarities within fold space. *Journal of Molecular Biology*, 323(5):909–26, November 2002.
- [HS93] L. Holm and C. Sander. Protein-structure comparison by alignment of distance matrices. *Journal of Molecular Biology*, 233:123–138, September 1993.
- [HS97] L. Holm and C. Sander. DALI/FSSP classification of three-dimensional protein folds. *Nucleic Acids Research*, 25(1):231–234, January 1997.
- [KH04] E. Krissinel and K. Henrick. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallographica Section D*, 60(12):2256–2268, Dec 2004.
- [KKSE⁺92] Ephraim Katchalski-Katzir, Isaac Shariv, Miriam Eisenstein, Asher A. Friesem, Claude Aflalo, and Ilya A. Vakser. Molecular surface recognition: Determination of geometric fit between proteins and their ligands by correlation techniques. In *Proceedings of the National Academy of Sciences*, volume 89, pages 2195–2199, March 1992.

- [KS98] Hans-Peter Kriegel and Thomas Seidel. Approximation-Based Similarity Search for 3-D Surface Segments. *GeoInformatica Journal*, 2(2):113–147, 1998.
- [MBHC95] A.G. Murzin, S.E. Brenner, T. Hubbard, and C. Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247:536–540, 1995.
- [MGB95] T. Madej, J. F. Gibrat, and S. H. Bryant. Threading a database of protein cores. *Proteins: Structure, Function and Genetics*, 23:356–369, 1995.
- [MJ95] A.C. May and M.S. Johnson. Improved genetic algorithm-based protein structure comparison: pairwise and multiple superpositions. *Protein Engineering*, 8(9):873–82, Sep 1995.
- [NLWN95] Raquel Norel, Shuo L. Lin, Haim J. Wolfson, and Ruth Nussinov. Molecular Surface Complementary at Protein-Protein Interfaces: The Critical Role Played by Surface Normals at Well Placed, Sparse, Points in Docking. *Journal of Molecular Biology*, 252(2):263–273, 1995.
- [NW91] Ruth Nussinov and Haim J. Wolfson. Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques. In *Proceedings of the National Academy of Sciences*, volume 88, pages 10495–10499, 1991.
- [OFCD02] R. Osada, T. Funkhouser, B. Chazelle, and D. Dobkin. Shape Distributions. *ACM Transaction on Graphics*, 21(4):807–832, October 2002.
- [OMJ⁺97] C.A. Orengo, A.D. Michie, S. Jones, D.T. Jones, M.B. Swindells, and Thornton J.M. CATH - a hierarchic classification of protein domains structures. *Structure*, 5(8):1093–108, August 1997.
- [SB98] I.N. Shindyalov and P.E. Bourne. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Engineering*, 12(9):739–747, 1998.
- [SOS96] Michel F. Sanner, Arthur J. Olson, and Jean-Claude Spehner. Reduced surface: An efficient way to compute molecular surfaces. *Biopolymers*, 38(3):305–320, March 1996.
- [Wol90] Haim J. Wolfson. Model-Based Object Recognition by Geometric Hashing. In *Proceedings of the 1st European Conference on Computer Vision*, Lecture Notes in Computer Science (427), pages 526–536. Springer, April 1990.