Proc. 7th Int. Conf. on Intelligent Systems for Molecular Biology (**ISMB'99**), Heidelberg, Germany.
This paper has been published by the American Association for Artificial Intelligence (http://www.aaai.org)

Nearest Neighbor Classification in 3D Protein Databases

Mihael Ankerst[1], Gabi Kastenmüller[2], Hans-Peter Kriegel[1], Thomas Seidl[1]

[1] University of Munich
Institute for Computer Science
Oettingenstr. 67, 80538 Munich, Germany
http://www.dbs.informatik.uni-muenchen.de

[2] Munich Information Center for Protein Sequences
Max Planck Institute for Biochemistry
Am Klopferspitz 18a, 82152 Martinsried, Germany
http://www.mips.biochem.mpg.de

Abstract

In molecular databases, structural classification is a basic task that can be successfully approached by nearest neighbor methods. The underlying similarity models consider spatial properties such as shape and extension as well as thematic attributes. We introduce 3D shape histograms as an intuitive and powerful approach to model similarity for solid objects such as molecules. Errors of measurement, sampling, and numerical rounding may result in small displacements of atomic coordinates. These effects may be handled by using quadratic form distance functions. An efficient processing of similarity queries based on quadratic forms is supported by a filter-refinement architecture. Experiments on our 3D protein database demonstrate the high classification accuracy of more than 90% and the good performance of the technique.

Keywords: 3D Protein Databases, Nearest Neighbor Classification, Geometric Similarity Search, Machine Learning

1 Introduction

One important task for modern database applications in the areas molecular biology, astronomy, mechanical engineering, medical imaging, and meteorology is efficient and accurate classification. In protein databases, each year a high number of new 3D structures of molecules is determined either by crystallography or by NMR techniques or by automated structure prediction. A basic question is to which class the new objects belong. The classes represent collections of objects that have characteristic properties in common and thus are similar, whereas different classes contain objects that have more or less strong dissimilarities. After classification, subsequent investigations may be guided by the class information: predictions of primary and secondary effects of drugs, docking prediction in general etc.

Typically, the classes are obtained from class information about objects that are already classified by exploiting domain and expert knowledge. Nearest neighbor classifiers, as a common approach, assign the class label of the most similar object in the previously classified database to the new object. As a basis for nearest neighbor classification, an appropriate similarity model has to be provided. In all of the mentioned applications, the geometric shape of the objects is an important similarity criterion. Along with the geometry, also thematic attributes such as physical and chemical properties have an influence on the similarity of objects.

As for most real-world applications, the available data about 3D protein structures inherently suffer from errors, including errors of measurement, calibration, sampling errors, numerical rounding errors, displacements of reference frames, and small shifts as well as rotations of the entire object or of local details of the shapes. Although no full invariance against rotations is required if the objects are already provided in a standardized orientation, these errors have to be taken into account. In this paper, we introduce a flexible similarity model that considers these problems of local inaccuracies and may be adapted by the users to their specific requirements or individual preferences.

The remainder of this introduction surveys related work from molecular biology, machine learning, and geometric similarity search. In Section 2, we introduce the components of our similarity model: 3D shape histograms for object representation, and a flexible similarity distance function. Due to the large and increasing size of current databases, the performance of query processing is an important task and, therefore, we present an efficient multistep system architecture in Section 3. The experimental results concerning the effectiveness and efficiency of our technique are presented in Section 4, and Section 5 concludes the paper.

1.1 Classification in Protein Databases

Structural classification schemata for protein databases are already available. In many systems, classifying new objects when inserting them into the database requires supervision by experts that are very experienced and have a deep knowledge of the domain of molecular biology. What is desired, is an efficient classification algorithm that may act as a fast filter for further investigation and that may be restricted e.g. to geometric aspects.

A sophisticated classification is available from the FSSP database (*Families of Structurally Similar Proteins*) generated by the DALI system (Holm and Sander 1994) (Holm and Sander 1998). The similarity of two proteins is based on their secondary structure. The evaluation of a pair of proteins is very expensive, and query processing for a single molecule

against the entire database currently takes an overnight run on a workstation. In (Holm and Sander 1995), one day of computer time for an all-against-all comparison of 385 representative proteins is reported.

Another classification schema is provided by CATH (Orengo et al. 1997), a hierarchical classification of protein domain structures which clusters proteins at four major levels, class (C), architecture (A), topology (T) and homologous superfamily (H). There are four class labels that are derived from secondary structure content: (1) *Mainly Alpha*, (2) *Mainly Beta*, (3) *Alpha Beta*, and (4) *Few Secondary Structures*. These four class labels are assigned for more than 90% of the protein structures automatically. The architecture label which describes the gross orientation of secondary structures, independent of connectivities, is currently assigned manually. The assignments of structures to topological families and homologous superfamilies are made by sequence and structure comparisons.

The superposition algorithm LOCK (Singh and Brutlag 1997) is also based on a hierarchy of structural representations, from the secondary structure level to the atomic level. For the method, a set of seven scoring functions is used to compare pairs of vectors from different proteins. The algorithm includes three steps, a local secondary structure superposition, an atomic superposition, and a core superposition. The system is able to detect structural similarities at the same level as the FSSP/DALI approach. It takes approximately 18 minutes to compare myoglobin to a representative set of 796 proteins from the PDB on an SGI workstation. Approximately 62% of the execution time was spent on the secondary and atomic superposition stages.

The VAST algorithm (*Vector Alignment Search Tool*) (Madej, Gibrat, and Bryant 1995) places great emphasis on the definition of the threshold of significant structural similarity. By a threading method, similarity to known domain folds is recognized. Ten proteins were selected for a blind prediction.

The SCOP system (*Structural Classification of Proteins*) (Murzin et al. 1995) is another structural classification that includes geometric features. It is unique in distinguishing between evolutionary relationships and those that arise from the physics and chemistry of proteins.

(Nakashima et al. 1986) analyzed the folding types of an entirety of 135 proteins in terms of the amino acid composition. The classification into five classes representing the folding types alpha, beta, alpha/beta, alpha+beta and irregular types gave an accuracy of 70%.

(Zhang et al. 1998) proposed a classification criterion relying on statistical analysis of the secondary structure contents of proteins. Each protein of a set of 200 proteins randomly selected from SCOP was assigned to one of three folding classes (mainly-alpha, mainly-beta, alpha-beta). A classification accuracy of 94% could be reached by this method using the three class labels.

In (Preißner, Goeda, and Frömmel 1998), an approach to model structural similarity of proteins using a concept of Dictionary of Interfaces in Proteins (DIP) is introduced. Equivalent molecular contact surfaces (patches) between neighboring secondary structures are compared, and the authors found that nearly all equivalent patches larger than 10 atoms have an identical geometry. Thus, geometry is of high significance for similarity search in protein databases, and it seems to be promising to focus on geometry primarily.

The problem we address in this paper includes the fast selection of candidate structures from a large database. The obtained set of candidates may be used as an input for more sophisticated methods of detecting structural similarity including dynamic programming algorithms or iterative RMSD minimization techniques.

1.2 Nearest-Neighbor Classification

A lot of research has been performed in the area of classification algorithms; surveys are presented in (Weiss and Kulikowski 1991), (Michie, Spiegelhalter, and Taylor 1994), and (Mitchell 1997). All the methods require that a training set of objects is given for which both the attribute values and the correct classes are known a priori. Based on this knowledge of previously classified objects, a classifier predicts the unknown class of a new object. The quality of a classifier is typically measured by the classification accuracy, i.e. by the percentage of objects for which the class label is correctly predicted.

Many methods of classification generate a description for the members of each class, for example by using bounding boxes, and assign a class to an object if the object matches the description of the class. Nearest neighbor classifiers, on the other hand, refrain from discovering a possibly complex description of the classes. As their name indicates, they retrieve the nearest neighbor p of a query object q and return the class label of p in order to predict the class label of q. Obviously, the definition of an appropriate distance function is crucial for the effectiveness of nearest neighbor classification. In a more general form, called k-nearest neighbor classification, k nearest neighbors of the query object q are used to determine the class of q. Thus, the effectiveness depends on the number k as well as on the weighting of the k neighbors. Both, appropriate similarity models as well as efficient algorithms for similarity search are required for successful nearest neighbor classification.

1.3 Geometric Similarity Search

Considerable work on shape similarity search in spatial database systems has been performed in recent years. As a common technique, the spatial objects are transformed into high-dimensional feature vectors, and similarity is measured in terms of vicinity in the feature space. The points in the feature space are managed by a multidimensional index. Many of the approaches only deal with two-dimensional objects

such as digital images or polygonal data and do not support 3D shapes.

Let us first survey previous 2D approaches from the literature. In (Gary and Mehrotra 1993), a shape is represented by an ordered set of surface points, and fixed-sized subsets of this representation are extracted as shape features. This approach supports invariance with respect to translation, rotation and scaling, and is able to deal with partially occluded objects. The method includes a linearization of polygon boundaries and, therefore, are hard to extend to 3D objects. In (Jagadish 1991), shapes are approximated by rectangular coverings. The rectangles of a single object are sorted by size, and the largest ones are used for the similarity retrieval. The method of (Korn et al. 1996) is based on mathematical morphology and uses the max morphological distance and max granulometric distance of shapes. It has been applied to 2D tumor shapes in medical image databases. A 2D technique that is related to our 3D shape histograms is the Section Coding technique (Berchtold, Keim, and Kriegel 1997). For each polygon, the circumscribing circle is decomposed into a given number of sectors, and for each sector, the area of the polygon inside of this sector divided by the total area of the polygon is determined. Similarity is defined in terms of the Euclidean distance of the resulting feature vectors. The similarity model in (Ankerst, Kriegel, and Seidl 1998) handles 2D shapes in pixel images and provides a solution for the problem of small displacements.

The QBIC (Querying By Image Content) system (Faloutsos et al. 1994) (Hafner et al. 1995) contains a component for 2D shape retrieval where shapes are given as sets of points. The method is based on algebraic moment invariants and is also applicable to 3D objects (Taubin and Cooper 1991). As an important advantage, the invariance of the feature vectors with respect to rigid transformations (translations and rotations) is inherently given. However, the adjustability of the method to specific applications is restricted. From the available moment invariants, appropriate ones have to be selected, and their weighting factors may be modified. Whereas the moment invariants are abstract quantities, the shape histograms presented in this paper are more intuitive and may be graphically visualized, thus providing an impression of the exact geometry. The approximation-based similarity model presented in (Kriegel and Seidl 1998) handles surface segments of 3D objects as required for docking prediction but does not include the global shape of 3D solids.

1.4 Invariance Properties of Similarity Models

All the mentioned similarity models incorporate invariance against translation of the objects, some of them also include invariance against scaling which is not desired in the context of molecular databases. With respect to invariance against rotations, two approaches can be observed. Some of the similarity models inherently support rotational invariance, e.g. by means of the Fourier transform or the algebraic moment invariants. Most of the techniques, however, include a preprocessing step that rotates the objects to a normalized orientation, e.g. by a Principal Axis Transform. For some applications, rotational invariance may be not required, e.g. if the objects are already stored in a standardized orientation.

An important kind of invariance has not been considered in previous work, the robustness of similarity models against errors of measurement, calibration, sampling errors, errors of classification of object components, numerical rounding errors, and small displacements such as shifts or slight rotations of geometric details. In our model, these problems are addressed and may be considered by the user by specifying and adapting a similarity matrix for histogram bins. A preliminary version of the histogram model was sketched in (Kastenmüller, Kriegel, and Seidl 1998).

2 A 3D Shape Similarity Model

In this section, we introduce our 3D shape similarity model by defining the two major ingredients: First, the shape histograms as an intuitive and discrete representation of complex spatial objects. Second, an adaptable similarity distance function for the shape histograms that may take small shifts and rotations into account by using quadratic forms.

2.1 Shape Histograms

The definition of an appropriate distance function is crucial for the effectiveness of any nearest neighbor classifier. A common approach for similarity models is based on the paradigm of feature vectors. A *feature transform* maps a complex object onto a feature vector in a multidimensional space. The similarity of two objects is then defined as the vicinity of their feature vectors in the feature space.

We follow this approach by introducing 3D shape histograms as intuitive feature vectors. In general, histograms are based on a partitioning of the space in which the objects reside, i.e. a complete and disjoint decomposition into cells which correspond to the bins of the histograms. The space may be geometric (2D, 3D), thematic (e.g. physical or chemical properties), or temporal (modeling the behavior of objects).

We suggest three techniques for decomposing the 3D space: A shell model, a sector model, and a spiderweb model as the combination of the former two. In a preprocessing step, a 3D solid is moved to the origin. Thus, the models are aligned to the center of mass of the solid.

Shell Model. The 3D is decomposed into concentric shells around the center point. This representation is particularly independent from rotations of the objects, i.e. any rotation of an object around the center point of the model results in the same histogram. The radii of the shells are determined from the extensions of the objects in the database. The outermost

shell is left unbound in order to cover objects that exceed the size of the largest known object.

Sector Model. The 3D is decomposed into sectors that emerge from the center point of the model. This approach is closely related to the 2D Section Coding method (Berchtold, Keim, and Kriegel 1997). However, the definition and computation of 3D sector histograms is more sophisticated, and we define the sectors as follows: Distribute the desired number of points uniformly on the surface of a sphere. For this purpose, we use the vertices of regular polyhedrons and their recursive refinements. Once the points are distributed, the Voronoi diagram of the points immediately defines an appropriate decomposition of the space. Since the points are regularly distributed on the sphere, the Voronoi cells meet at the center point of the model. For the computation of sector-based shape histograms, we need not to materialize the complex Voronoi diagram but simply apply a nearest neighbor search in 3D since typical number of sectors are not very large.

Combined Model. The combined model represents more detailed information than models purely based on shells or sectors. A combination of two fine-grained 3D decompositions may result in a high dimensionality. However, the resolution of the space decomposition is a parameter of the model, and the number of dimensions is easily adapted to the particular application.

In Figure 1, we illustrate various shape histograms for an example protein, the seryl-tRNA synthetase 1SER-B. The various space decompositions are indicated schematically, and the corresponding shape histograms are depicted. The top histogram is purely based on shell bins, the middle histogram is a combination of shell bins and sector bins, and the bottom histogram consists of sector bins only. In the example, the histograms are defined in a way that they all have the same dimension of approximately 120. Note that the histograms are not built from volume elements but from uniformly distributed surface points taken from the molecular surfaces.

2.2 Similarity Distance Function

Although the Euclidean distance is a very common distance function for high-dimensional feature vectors such as shape histograms, it exhibits severe limitations with respect to similarity measurement. In particular, the individual components of the feature vectors which correspond to the dimensions of the feature space are assumed to be independent from each other, and no relationships of the components such as substitutability and compensability may be regarded.

An approach to overcome these limitations has been investigated for color histograms in the QBIC project (Query by Image Content) at IBM Almaden (Faloutsos et al. 1994) (Hafner et al. 1995). The authors suggest to use quadratic

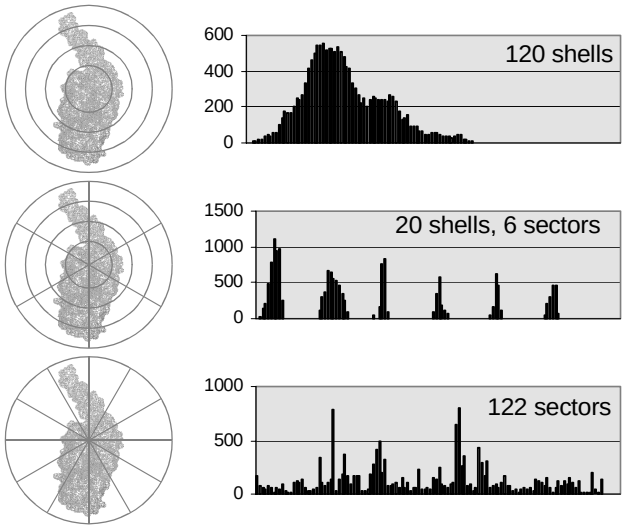

Figure 1: Several 3-D shape histograms of the example protein 1SER-B. From top to bottom, the number of shells decreases and the number of sectors increases.

form distance functions which have been successfully applied to several multimedia database applications (Seidl 1997) (Seidl and Kriegel 1997) (Ankerst, Kriegel, and Seidl 1998) (Kriegel and Seidl 1998). A quadratic form distance function is defined in terms of a similarity matrix A where the components a_{ij} represent the similarity of the components i and j in the underlying vector space.

$$d_A^2(x, y) = (x - y) \cdot A \cdot (x - y)^T = \sum_{i=1}^{N} \sum_{j=1}^{N} a_{ij}(x_i - y_i)(x_j - y_j)$$

In our case, cross-dependencies of dimensions result from the consideration of the neighborhood of cells. These dependencies are represented by similarity weights in the similarity matrix A. Following the approach of (Hafner et al. 1995), we compute the similarity weights by the formula $a_{ij} = e^{-\sigma \cdot d(i,j)}$ where $d(i, j)$ denotes the distance of the cells that corresponds to the bins i and j. The cell distance is calculated from the difference of the shell radii and the angle between the sectors. The parameter σ controls the global shape of the similarity matrix. We observed good results for σ between 1.0 and 10. The classic Euclidean distance is a special case of the quadratic form distance which is achieved by using the identity matrix Id as similarity matrix.

2.3 Invariance Properties of the Models

From structure determination by crystallographic or NMR methods, the geometric structure of a proteins can be located anywhere in the 3D, and the orientation may vary arbitrarily. For defining meaningful and applicable similarity models, invariance against translation and rotation has to be provided.

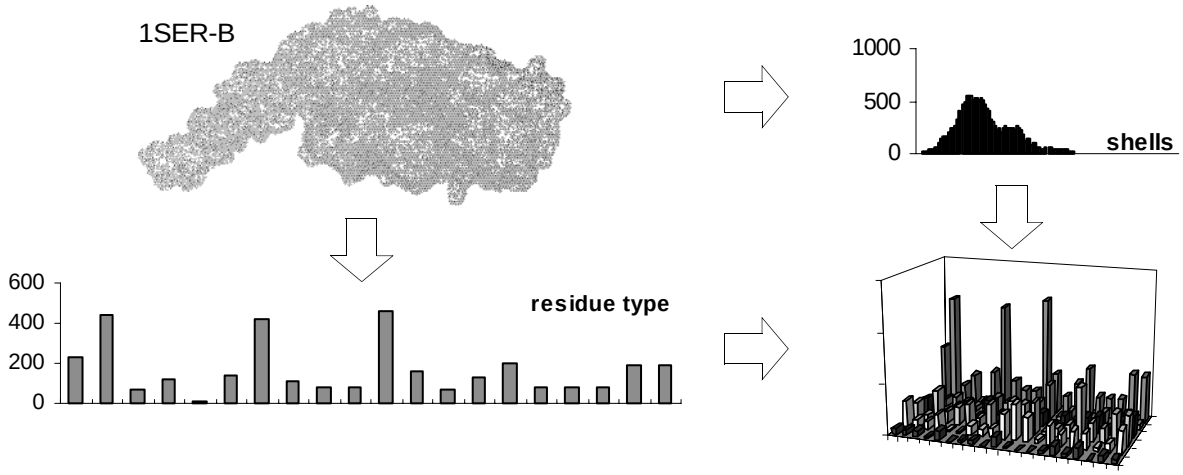

Figure 2: Example for a combined thematic and shape histogram for a molecule

In our system, we perform a preprocessed normalization step and, thereby, achieve invariance against translation and rotation. For each object, we map the center of mass onto the origin and perform a Principal Axis Transform. As a result, all the covariances of the transformed coordinates vanish. Although this method in general leads to a unique orientation of the objects, this does not hold for the exceptional case of an object with at least two variances having the same value. In order to overcome this problem, the objects may be rotated artificially by certain angles as suggested in (Berchtold, Keim, and Kriegel 1997). We almost never observed such cases in our experiments and, therefore, assume a unique orientation of the objects.

The similarity models themselves have inherent invariance properties. Obviously, the sector model is invariant against scaling, whereas the shell model trivially has rotational invariance. Often, no full invariance is desired, instead just small displacement, shifts or rotations of geometric details occur in the data, for example caused by errors of measurement, sampling or numerical rounding errors. This variation of invariance precision which is highly application- and user-dependent is supported by the user-defined similarity matrix modeling the appropriate similarity weight for each pair of bins.

2.4 Extensibility of Histogram Models

What we have discussed so far is a very flexible similarity model for 3D objects. However, the distance function of the similarity model is based just on the spatial attributes of the objects. Frequently on top of the geometric information, a lot of thematic information is used to describe spatial objects. Particularly in protein databases, the chemical structure and physical 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 2 demonstrates the basic principle. Given a spatial histogram structure as presented above, and an additional thematic histogram structure, a combined histogram structure is immediately obtained as the Cartesian product of the original structures.

Having integrated thematic information into the object representation e.g. by the Cartesian product approach, the question remains how to combine the distance functions d_{shape} and d_{them} to an overall distance function $d_{overall}$. A straightforward way is to use the standard Euclidean distance

$$d_{overall}(p, q) = \sqrt{d^2_{shape}(p_s, q_s) + d^2_{them}(p_t, q_t)}$$

where p_s, q_s denote the projection of p, q to their shape components, and p_t, q_t denote the projection of p, q to the thematic components.

Again, an even more powerful model is provided by quadratic forms. They support to consider cross-dependencies of thematic attributes such as the similar hydrophobicity of different residue types and, in case of the combined histograms, they are able to handle relationships between thematic attributes and spatial attributes by using appropriate similarity matrices.

As an example, the *hydrophobic effect* is reflected by combined histogram structures (Kastenmüller 1998). This phenomenon means that in a polar environment, hydrophobic residues tend to be located in the core of a protein whereas hydrophilic residues often are found at the surface of the molecules. This effect can be observed from a representation combining shells and residue types. The abscissa axis in Figure 3 indicates the shells from the center with increasing radii, and the residue types are colored by their hydrophobicity. In order to reveal the hydrophobic effect, we distinguish only between three classes of amino acids: hydrophobic residues (dark gray), neutral residues (white), and hydrophilic

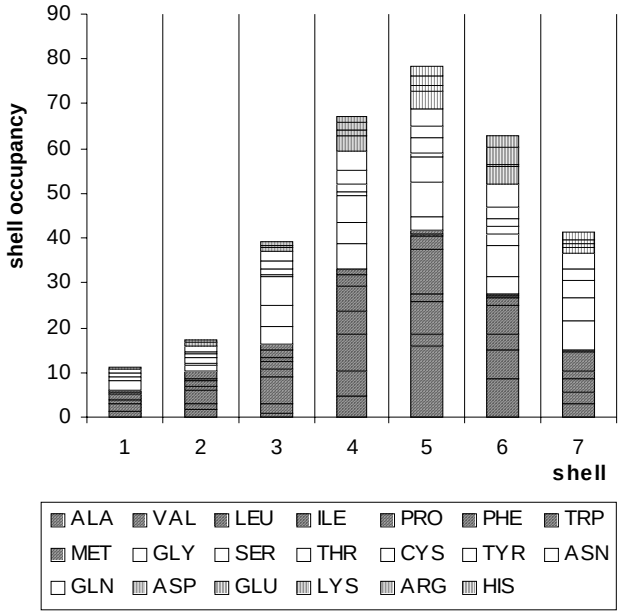

Figure 3: Hydrophobic effect for the yeast hydrolase (carboxylic esterase) 1TCA. The hydrophobic residues are marked by dark gray, the neutral residues by white, and the hydrophilic residues by light gray according to the ordering of (Wampler 1996).

residues (light gray). As expected, the inner shells (1 and 2) are dominated by hydrophobic residues whereas the majority of the hydrophilic residues is located at the protein surface, i.e. in the shells 4 to 7.

3 Efficient Query Processing

Due to the enormous and still increasing size of modern databases that contain tens and hundreds of thousands of molecules, mechanical parts or medical images, the task of efficient query processing becomes more and more important. In the case of quadratic form distance functions, the evaluation time of a single database object increases quadratically with the dimension. For calculating the distance for a single pair of objects, we measured average CPU times of 0.23 milliseconds for 21-D histograms, 1.1 milliseconds for 112-D and 6.2 milliseconds in 256-D space (Seidl 1997). In such a context, linearly scanning the entire database is prohibitive, since the overall runtime results from this single evaluation time multiplied with the number of objects in the database plus the disk I/O time. Reading a 4K disk page takes approximately 10 milliseconds and transfers at least 4 objects into main memory (in case of 256-D histograms), i.e. the average I/O time is 2.5 milliseconds per object. Thus, the overall runtime is dominated by the CPU cost and even caching the database in main memory would not help.

In order to achieve a good performance, our system architecture follows the paradigm of multi-step query processing: An index-based filter step produces a set of candidates, and a subsequent refinement step performs the expensive exact evaluation of the candidates.

3.1 Optimal Multi-step *k*-Nearest Neighbor Search

Whereas the refinement step in a multi-step query processor has to ensure the correctness, i.e. no false hits may be reported as final answers, the filter step is primarily responsible for the completeness, i.e. no actual result may be missing from the final answers and, therefore, from the set of candidates. Figure 4 illustrates the architecture of our multi-step similarity query processor that fulfills this property (Seidl and Kriegel 1998). The algorithm is proven to be optimal, i.e. it produces only the minimum number of candidates. Thus, expensive evaluations of unnecessary candidates are avoided.

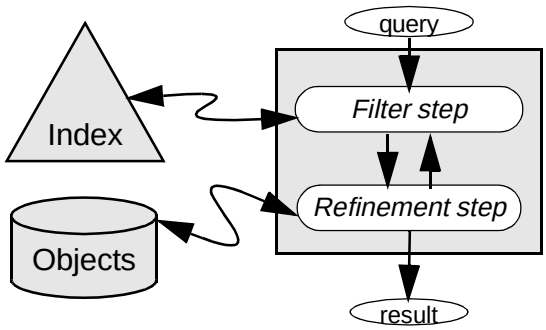

Figure 4: Multi-step similarity query processing.

Based on a multidimensional index structure, the filter step performs an incremental ranking that reports the objects ordered by their increasing filter distance to the query object using an algorithm derived from (Hjaltason and Samet 1995). The number of accessed index pages is minimal and the termination is controlled by the refinement step in order to guarantee the minimum number of candidates (Seidl and Kriegel 1998). Only for the exact evaluation in the refinement step, the exact object representation is retrieved from the object server. In order to guarantee no false dismissals caused by the filter step, the filter distance function d_f has to be a lower bound of the exact object distance function d_o that is evaluated in the refinement step. That is, for all database objects p and all query objects q, the inequality $d_f(p, q) \leq d_o(p, q)$ has to be fulfilled.

3.2 Reduction of Dimensionality for Quadratic Forms

A common approach to manage objects in high-dimensional spaces is to apply techniques to reduce the dimensionality. The objects in the reduced space are then typically managed by any multidimensional index structure (Gaede and Günther 1998). The typical use of common linear reduction

techniques such as the Principal Components Analysis (PCA) or Karhunen-Loève Transform (KLT), the Discrete Fourier or Cosine Transform (DFT, DCT), the Similarity Matrix Decomposition (Hafner et al. 1995) or the Feature Subselection (Faloutsos et al. 1994) includes a clipping of the high-dimensional vectors such that the Euclidean distance in the reduced space is always a lower bound of the Euclidean distance in the high-dimensional space.

The question arises whether these approved techniques are applicable to general quadratic form distance functions. Fortunately, the answer is positive; an algorithm to reduce the similarity matrix from a high-dimensional space down to a low-dimensional space according to a given reduction technique was developed in the context of multimedia databases for color histograms (Seidl and Kriegel 1997) and shapes in 2D images (Ankerst, Kriegel, and Seidl 1998). The method guarantees three important properties: First, the reduced distance function is a lower bound of the given high-dimensional distance function. Obviously, this criterion is a necessary design goal in order to meet the requirements of multi-step similarity query processing. Second, the reduced distance function again is a quadratic form and, therefore, the complexity of the query model is not increased while decreasing the dimension of the space. Third, the reduced distance function is the greatest of all lower-bounding distance functions in the reduced space. As an important implication of this property, the selectivity in the filter step is optimal: In the reduced space, no lower-bounding distance function is able to produce a smaller set of candidates than the resulting quadratic form.

3.3 Ellipsoid Queries on Multidimensional Index Structures

The task remains to efficiently support *k*-nearest neighbor search and incremental ranking for quadratic form distance functions in low-dimensional spaces. Due to the geometric shape of the query range, a quadratic form-based similarity query is called an *ellipsoid query* (Seidl 1997). An efficient algorithm for ellipsoid query processing on multidimensional index structures was developed in the context of approximation-based similarity search for 3-D surface segments (Kriegel and Seidl 1998). The method is designed for index structures that use a hierarchical directory based on rectilinear bounding boxes such as the R-tree, R+-tree, R*-tree, X-tree, and Quadtrees among others; surveys are provided e.g. in (Samet 1990) or (Gaede and Günther 1998). The technique is based on measuring the minimum quadratic form distance of a query point to the hyperrectangles in the directory. A parallel version of the X-tree is introduced in (Berchtold et al. 1997). Recently, an improvement by using conservative approximations has been suggested in (Ankerst et al. 1998).

An important property of the method is its flexibility with respect to the similarity matrix. The matrix does not have to be available at index creation time and, therefore, may be considered as a query parameter. Thus, the users may specify and adapt the similarity weights in the matrix even at query time according to their individual preferences or to the specific requirements of the application. In any case, the same precomputed index may be used. This property is the major advantage compared to previous solutions that were developed in the context of color histogram indexing in the QBIC project (Faloutsos et al. 1994) (Hafner et al. 1995) where the index depends on a specific similarity matrix that has to be given in advance.

4 Experimental Evaluation

We implemented the algorithms in C++ and ran the experiments on our HP C160 workstations under HP-UX 10.20. For single queries, we also implemented a HTML/Java interface that supports query specification and visualization of the results. The atomic coordinates of the 3D protein structures are taken from the Brookhaven Protein Data Bank (PDB) (Abola et al. 1997). For the computation of shape histograms, we use a representation of the molecules by surface points as it is required for several interesting problems such as the molecular docking prediction (Seidl and Kriegel 1995). The reduced feature vectors for the filter step are managed by an X-tree (Berchtold, Keim, and Kriegel 1996).

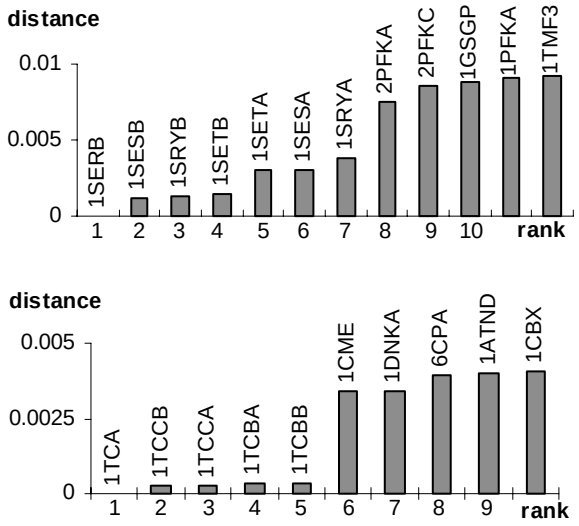

Figure 5: Similarity ranking for the seryl-tRNA synthetase 1SER-B (top) and the yeast hydrolase 1TCA (bottom) for histograms of 6 shells and 20 sectors. The diagrams depict the top nearest neighbors and their similarity distances to the query protein.

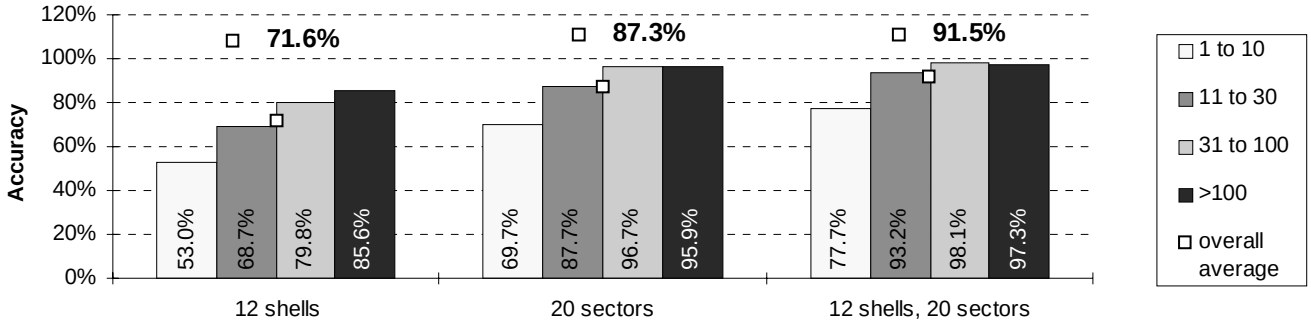

Figure 6: Classification accuracy of various histogram models (12 shells, 20 sectors, and the combination) for 3,422 proteins.

4.1 Basic Similarity Search

In order to illustrate the applicability of the similarity model, we first demonstrate the retrieval of the members of a known family. As a typical example, we chose the seven seryl-tRNA synthetase molecules from our database that are classified by CATH (Orengo et al. 1997) to the same family. The diagrams in Figure 5 present the result using shape histograms for 6 shells and 20 sectors. 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 2PFK-A, the first non-seryl protein in the ranking order. The same effect can be observed for the hydrolase family 1TCA where a distance gap occurs after the five members of the family.

4.2 Classification by Shape Similarity

For the classification experiments, we restricted our database to the proteins (domains) that are also contained in the FSSP database (Holm and Sander 1994) and took care that for every class, at least two molecules are available. From this preprocessing, we obtained 3,422 proteins assigned to 281 classes. The classes contain between 2 and 185 molecules, and we grouped the proteins by the cardinality of their class to provide more information than simply the average accuracy. Table 1 shows the four groups and the number of proteins that belong to the corresponding classes.

cardinality of classes	number of proteins
1 to 10	838
11 to 30	973
31 to 100	727
more than 100	884

Table 1: Number of proteins grouped by the cardinality of their classes

In order to measure the classification accuracy, we performed *leave-one-out* experiments for the various histogram models. This means that for each molecule in the database, the nearest neighbor classification was determined after removing that element from the database. Technically, we always used the same database and selected the second nearest neighbor since the query object itself is reported to be its own nearest neighbor. The class label of the selected object is assigned to the query object. The overall classification accuracy is computed as the percentage of the correctly predicted class labels among all 3,422 database objects.

Figure 6 demonstrates the results for histograms based on 12 shells, 20 sectors, and the combination of them. Obviously, the more fine-grained spiderweb model yields the best classification accuracy of 91.5 percent in the overall average, but even for the coarse sector histograms, a noticeable accuracy of 87.3 percent is achieved. The average overall runtime for a single nearest neighbor query (see Figure 7) reflects the larger dimension of the combined model. It ranges from 0.05s for 12 shells over 0.2s for 20 sectors up to 1.42s for the combination.

Figure 8 illustrates the effect of simply increasing the dimension of the model without combining orthogonal space

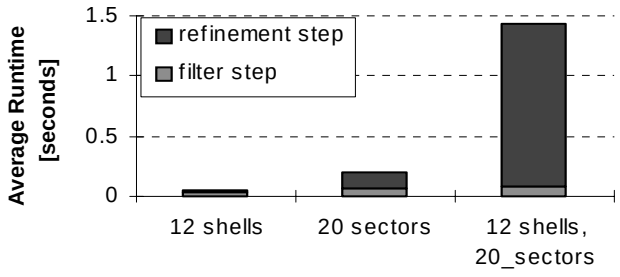

Figure 7: Average runtime of a single nearest neighbor query for histograms with 12 shells, 20 sectors, and their combination for a database of 3,422 proteins.

partitionings. Again, we observed the expected result that more information yields better accuracy. For ten times higher histogram dimensions, the overall average accuracy increases from 71.6 to 88.1 for the shell model, and from 87.3 to 91.6 for the sector model. Thus, we achieve the same accuracy for a fine-grained 122D sector model as we obtained from the 12 x 20 (240D) combined model.

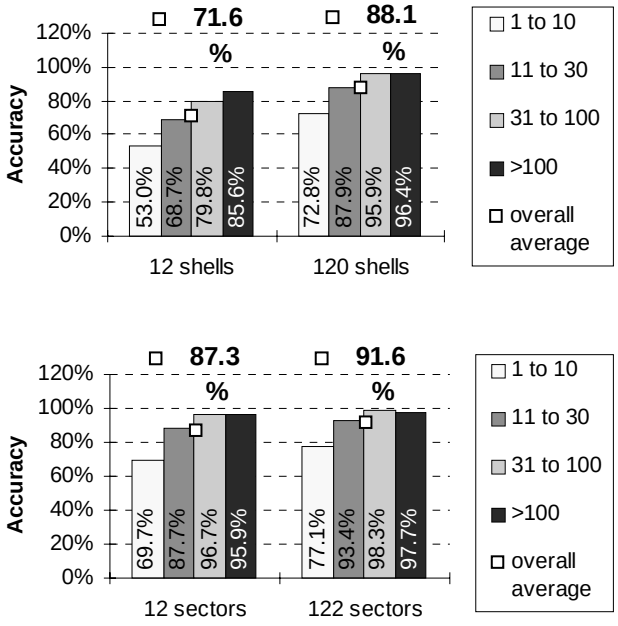

Figure 8: The accuracy increases with the granularity of the space partitioning for both, shell histograms (top diagram) and sector histograms (bottom diagram). Again, the larger classes (cf. legend) yield the better accuracy.

5 Conclusions

We presented a new intuitive and flexible model for classification of 3D solids such as proteins based on shape similarity. The geometry of 3D objects is mapped to shape histograms, thus performing a specific feature transform. The histogram model naturally is extensible to thematic attributes such as physical and chemical properties. In order to account for errors of measurement, sampling, numerical rounding etc., quadratic form distance functions are used that are able to cope with small displacements and rotations. For efficient query processing, a filter-refinement architecture is used that supports similarity query processing based on high-dimensional feature vectors and quadratic form distance functions.

The experiments demonstrate the high classification accuracy of our shape histogram model. Automatic class label assignment is performed with an accuracy of 90% for a large variety of 281 different class labels. This result competes with CATH and other automatic classification techniques that also yield accuracies of more than 90% but work with significantly smaller numbers of some few class labels and use more complex similarity models.

On top of the high accuracy, the experiments reveal the good performance of the underlying query processor. The high efficiency supports online query processing in contrast to batch query processing of competing systems where the query result is sent back by e-mail.

In our future work, we plan to optimize the space partitioning and the geometry of the cells which form the histogram bins. Both, the number as well as the geometry of the cells affect the effectiveness and also the efficiency of similarity search and classification. Furthermore, we plan to extend our classification system by a visualization tool for shape histograms in order to provide an explanation component for the user. This is an important issue since any notion of similarity is subjective in a high degree, and the users want to have as much feedback as possible concerning the decision of the system depending on their queries and input parameters. Furthermore, the confidence of the users in an automatic classification increases with the reproducibility of the decision by the user which can be enhanced by visualization methods.

References

Abola E. E., Sussman J. L., Prilusky J., Manning N. O. 1997. Protein Data Bank Archives of Three-Dimensional Macromolecular Structures. In Carter C. W., Sweet R. M. (eds.) *Methods in Enzymology* 277:556-571. Academic Press.

Ankerst M., Braunmüller B., Kriegel H.-P., Seidl T. 1998. Improving Adaptable Similarity Query Processing by Using Approximations. In *Proceedings of the 24th International Conference on Very Large Data Bases (VLDB)*, 206-217. San Francisco, Calif.: Morgan Kaufmann.

Ankerst M., Kriegel H.-P., Seidl T. 1998. A Multistep Approach for Shape Similarity Search in Image Databases. *IEEE Transactions on Knowledge and Data Engineering* 10(6):996-1004.

Berchtold S., Böhm C., Braunmüller B., Keim D., Kriegel H.-P. 1997. Fast Parallel Similarity Search in Multimedia Databases. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 1-12. ACM Press. (Best Paper Award)

Berchtold S., Keim D., Kriegel H.-P. 1996. *The X-tree:* An Index Structure for High-Dimensional Data. In *Proceedings of the 22nd International Conference on Very Large Data Bases (VLDB)*, 28-39. Morgan Kaufmann.

Berchtold S., Keim D.A., Kriegel H.-P. 1997. Section Coding: A Method for Similarity Search in CAD Databases. In *Proceedings of the German Conference on Databases for Office Automation, Technology, and Science (BTW)*, 152-171, Series Informatik Aktuell. Springer Verlag. (in German)

Faloutsos C., Barber R., Flickner M., Hafner J., Niblack W., Petkovic D., Equitz W. 1994. Efficient and Effective Querying by Image Content. *Journal of Intelligent Information Systems* 3:231-262.

Gaede V., Günther O. 1998. Multidimensional Access Methods. *ACM Computing Surveys* 30(2):170-231.

Gary J. E., Mehrotra R. 1993. Similar Shape Retrieval Using a Structural Feature Index. *Information Systems* 18(7):525-537.

Hafner J., Sawhney H. S., Equitz W., Flickner M., Niblack W. 1995. Efficient Color Histogram Indexing for Quadratic Form Distance Functions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 17(7):729-736.

Hjaltason G. R., Samet H. 1995. Ranking in Spatial Databases. In *Proceedings of the 4th International Symposium on Large Spatial Databases*, Lecture Notes in Computer Science 951:83-95. Springer Verlag.

Holm L., Sander C. 1994. The FSSP Database of Structurally Aligned Protein Fold Families. *Nucleic Acids Research* 22:3600-3609.

Holm L., Sander C. 1995. 3-D Lookup: Fast Protein Structure Database Searches at 90% Reliability. In *Proceedings of the 3rd International Conference on Intelligent Systems for Molecular Biology (ISMB)*, 179-187. Menlo Park, Calif.: AAAI Press.

Holm L., Sander C. 1998. Touring Protein Fold Space with Dali/FSSP. *Nucleic Acids Research* 26:316-319.

Jagadish H. V. 1991. A Retrieval Technique for Similar Shapes. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 208-217. ACM Press.

Kastenmüller G. 1998. *Shape-oriented Similarity Search in 3D Protein Database Systems.* Master thesis, Institute for Computer Science, University of Munich. (in German)

Kastenmüller G., Kriegel H.-P., Seidl T. 1998. Similarity Search in 3D Protein Databases. In *Proceedings of the German Conference on Bioinformatics (GCB)*. University of Cologne.

Korn F., Sidiropoulos N., Faloutsos C., Siegel E., Protopapas Z. 1996. Fast Nearest Neighbor Search in Medical Image Databases. In *Proceedings of the 22nd International Conference on Very Large Data Bases (VLDB)*, 215-226. Morgan Kaufmann.

Kriegel H.-P., Seidl T. 1998. Approximation-Based Similarity Search for 3-D Surface Segments. *GeoInformatica Journal* 2(2):113-147. Kluwer Academic Publishers.

Madej T., Gibrat J-F., Bryant S.H. 1995. Threading a database of protein cores. *Proteins* 23(3):356-369.

Michie D., Spiegelhalter D.J., Taylor C.C. 1994. *Machine Learning, Neural and Statistical Classification*. Series in Artificial Intelligence. New York: Ellis Horwood.

Mitchell T.M. 1997. *Machine Learning*. New York: McCraw-Hill.

Murzin A. G., Brenner S. E., Hubbard T., Chothia C. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology* 247:536-540.

Nakashima H., Nishikawa K., Ooi T. 1986. The Folding Type of a Protein is Relevant to the Amino Acid Composition. *Journal of Biochemistry* (Tokyo) 99(1):153-162

Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., Thornton, J.M. 1997. CATH – A Hierarchic Classification of Protein Domain Structures. *Structure* 5(8):1093-1108.

Preißner R., Goeda A., Frömmel C. 1998. Dictionary of Interfaces in Proteins (DIP), Data Bank of Complementary Molecular Surface Patches. *Journal of Molecular Biology* 280:535-550.

Samet H. 1990. *The Design and Analysis of Spatial Data Structures.* Reading, Mass.: Addison-Wesley.

Seidl T. 1997. *Adaptable Similarity Search in 3-D Spatial Database Systems*. Ph.D. thesis, Faculty for Mathematics and Computer Science, University of Munich. Munich, Germany: Herbert Utz Publishers. http://utzverlag.com, ISBN: 3-89675-327-4.

Seidl T., Kriegel H.-P. 1995. Solvent Accessible Surface Representation in a Database System for Protein Docking. In *Proceedings of the 3rd International Conference on Intelligent Systems for Molecular Biology (ISMB)* 350-358. Menlo Park, Calif.: AAAI Press.

Seidl T., Kriegel H.-P. 1997. Efficient User-Adaptable Similarity Search in Large Multimedia Databases. In *Proceedings of the 23rd International Conference on Very Large Data Bases (VLDB)* 506-515. Morgan Kaufmann.

Seidl T., Kriegel H.-P. 1998. Optimal Multi-Step k-Nearest Neighbor Search. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 154-165. ACM Press.

Singh A. P., Brutlag D. L. 1997. Hierarchical Protein Structure Superposition using both Secondary Structure and Atomic Representations. *Proceedings of the 5th International Conference on Intelligent Systems for Molecular Biology (ISMB)*, 284-293. Menlo Park, Calif.: AAAI Press.

Taubin G., Cooper D. B. 1991. Recognition and Positioning of Rigid Objects Using Algebraic Moment Invariants. *Geometric Methods in Computer Vision* 1570:175-186. International Society for Optical Engineering (SPIE).

Wampler J. 1996. *Tutorial on Protein Structure*. Dept. of Molecular Biology, University of Georgia, http://selene.biochem.uga.edu.

Weiss S.M., Kulikowski C.A. 1991. *Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems*. San Mateo, Calif.: Morgan Kaufmann.

Zhang C., Zhang R. 1998. A New Criterion to Classify Globular Proteins based on their Secondary Structure Contents. *Bioinformatics* 14(10):857-865.