SHREC'07 - Protein Retrieval Challenge

Maja Temerinac, Marco Reisert and Hans Burkhardt
Albert-Ludwig University Freiburg
Computer Science Laboratory
79110 Freiburg, Germany
temerina@informatik.uni-freiburg.de

Abstract

The SHREC - 3D Shape Retrieval Contest aims to evaluate the effectiveness of 3D-shape retrieval algorithms on various types of data. In this particular track the structure of three dimensional proteins is under consideration. The Protein Database (PDB) offers over 30000 protein structures. To cope with such a huge amount of data automatic classification and search tools get more and more important in biomolecular research. Feature based approaches are known to be the tool to provide a fast content based retrieval. In this contest we want to evaluate such methods. Each protein is attached with a fingerprint which is relying solely on coordinates of the atom sequence of the protein, no further information is used. Four different methods methods are evaluated.

1 Motivation

Proteins are linear sequences of amino acids which fold into three dimensional structures. Throughout evolution the amino acid composition can change, but the three dimensional structure of the protein stays conserved. The three dimensional structure of a protein is closely linked to its function. So, by finding similar three dimensional protein structures, their function and evolutionary linkage can be determined.

Molecular biologists are often interested in getting a survey of the objects in a biomolecular database making classification one of their basic tasks: To which of the recognized classes in the database does a new molecule belong? Several classification schemata such as SCOP [1], CATH and DALI/FSSP are available in the Internet. When a new object is inserted into the database the supervision by experts that are very experienced and have a deep knowledge in the domain of molecular biology is necessary in most cases. An efficient classification algorithm is desired that can speed up the classification process by acting as a fast filter for further investigation.

While SCOP and CATH require classification by human experts, a fully automatic classification is available from the FSSP database (Families of Structurally Similar Proteins), generated by the DALI (Distance matrix ALIgnment) system. The evaluation of a protein query by the DALI method is very expensive; comparing a single molecule against the entire FSSP database currently takes an overnight run.

2 The task

The task of this competition is to classify protein domains to one of the SCOP folds. The participants were able to train their feature extraction algorithms on the provided data set. One day before the end of the competition, the participants were provided with a set of 30 unknown protein domains. The query files had contained all atoms of the protein domain and their 3D coordinates. The task was then to assign the query protein domains to SCOP folds. Since the entire SCOP database is divided into more than 970 folds, we limited the task to assigning the unknown protein domains to one of the 27 folds provided in the data set.

We provided a dataset, which consists of 685 protein domains divided into 27 folds according to their SCOP classification. We have chosen this dataset because it is rather difficult, it does not contain any close-by related structures. Thus, the performance differences between the competitors become more apparent. For each protein only the atom positions are allowed to be used for retrieval. No additional information like chemical properties or others, e.g. temperature were allowed to be used. The 3D coordinates were provided in the common *pdb* file format.

2.1 SCOP

In the SCOP[1] (Structural Classification of Proteins) database published in 1995 all proteins of known structure are ordered according to their evolutionary and structural relationship. The protein domains are hierarchically grouped into families, superfamilies, folds and classes The basic unit in SCOP is a protein domain. The domain is either a monomer or a part of a protein and it should reflect a structure that did not change throughout evolution. Since this definition is very hard to measure by an algorithm, SCOP solely relies on visual inspection by experts.

Each domain can be addressed either by an unique integer (sunid) or by a concise classification string (sccs). For example, the protein with the PDB identity 1dlr has the sunid 34906 and the sccs 'c.71.1.1', where 'c' stands for the class, '71' the fold, '1' the superfamily and the last '1' for the family. In the 'dir.des.scop.txt' file the domains sunid, sccs and English names for proteins, families, superfamilies, folds and classes are listed. Also the sequence number where the domain in the chain starts and ends is contained in this file.

A family consists of proteins which either have residue identities over 30% or have similar structure or functions. Globins and Triosephosphate isomerase (TIM) are examples of protein families. A superfamily consists of proteins with lower than 30% sequential identity and a probable common evolutionary origin. Examples for superfamilies are Actin-crosslinking proteins. A fold contains proteins having same major secondary structures in same arrangement with the same topological connections. The most interesting members of a fold are those with low sequential similarity where there exists an evolutionary link to the other proteins of the fold. A class contains folds with similar secondary structure and is the most general way of defining a protein structure.

3 Participants

In this track we had two groups participating:

- B. Li, Y. Fang, K. Ramani, D. Kihara (Purdue University, USA)
- P. Daras, V. Tsatsaias (ITI, Greece)

The group from ITI participated with two different methods:

- a three dimensional shape-structure comparison method (Trace) [5]
- a graph based method (Graph) (not yet published)

Each group submitted a ranked list of the unknown 30 protein structures (See Figure 2) together with the distance of each query to each protein from the 633 training set computed by their method. The submitted ranked lists are available at *http://lmb.informatik.uni-freiburg.de/events/shrec07/results.html*. The SCOP classifiaction [1] was considered as the ground truth. Only the ATOM section of the PDB [2] files was provided.

We also compared the results to the classification achieved by our method (LMB, Germany)[3]. Since we organized the track, our results are out of competition. But we want to emphasize here that our features were not tuned on the 30 test proteins, we only used the training set for parameter tuning.

4 Methods

Li et al. focus on the topology of each protein: they use STRIDE [4] to detect the secondary structure, including the hydrogen bond. Then, they compute the beta sheets (beta strands connected with hydrogen bond) and the order. For main class a, b, c, d, g, and folds of a and g, they used the length and percentage of alpha helix and beta strand to classify. For each fold in each class b, c, d, they used the orders to classify.

P. Daras and V. Tsatsaias submitted two ranked lists computed with two different methods. The first method (Trace) is described in the paper [5]. The second method (Graph) is called '3D Protein Classification Using Toplogical and Geometrical Information'. The 3D objects are firstly segmented to their molecular structure. Then, descriptors are extracted for each segment using spherical harmonics algorithms, and graphs are constructed for the molecules. Next, a sub-graph matching procedure is utilized in order to provide final similarity distances between the graphs.

Our method (LMB) was proposed in [3]. The basic idea of our approach is to obtain invariant fingerprint of the 3D structure. Therefore, we use a group integration approach. Practically the features can be seen as joint histograms over spatial distances, sequential distances and 2 angle-like quantities.

5 Evaluation

The ranked lists were evaluated by the following simple method: The next neighbor in the ranked list, meaning the protein domain with the least distance to the query protein is considered and the query protein

[1] http://scop.mrc-lmb.cam.ac.uk/scop/

is assigned to its class. One point is scored for the correct SCOP class only, two points for the correct SCOP fold and zero points if neither of them is correct. The maximal amount of points is 60, when the fold for each query protein is correctly classified.

As can be seen in Figure 1, from the three submitted methods , the team from Purdue performed (total score 45) best even though using simple features. The two methods submitted by team ITI misclassified half of the query proteins and their best method Graph scored 29 points. However, even better classification could be achieved by the LMB team, total score 52.

The query set (Figure 2) was chosen randomly from the 27 scop folds. Some proteins consisted of only one domain, others (e.g. Protein2, Protein8, Protein23) of several domains which were however all belonging to the same fold. Also, the size of the protein domains ranged from 31 amino acids (Protein11) to 364 amino acids (Protein16).

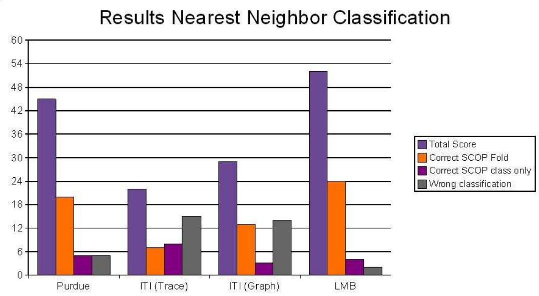

Figure 1. Comparison of the methods in terms of nearest neighbor classification.

Group	Wrong	Correct Class	Correct Fold	score
Purdue	5	5	20	**45**
ITI(Trace)	15	8	7	**22**
ITI(Graph)	14	3	13	**29**
LMB	2	4	24	**52**

Table 1. The results of the nearest neighbor classification according to the ranked lists submitted by each group.

6 Discussion and Conclusion

It is astonishing that the very simple method from Purdue is working so well in comparison to the methods proposed by ITI. It seems that properties and frequencies of secondary structure elements are a very important information for the presented task. The method (Trace) by ITI covers more the overall 3D structure, the tertiary structure, and is less sensitive for secondary and primary structure elements. Another issue is that (Trace) uses a normalization approach (by the center of mass) to obtain invariance against translations of the 3D structure. This approach can be very unstable when only partial structures are matching. Though the (Graph) method works a little bit better but is still worse. It seems that statistical features are the better alternative than trying to establish one-to-one correspondences by a matching approach. The good performance of our method (LMB) could be explained by the fact that it describes all structural levels uniformly. Primary and secondary structure elements are described by cooccurences of small sequential and spatial distances, the tertiary structure is contained in occurrences of larger distances.

In conclusion, we can say that this competition has shown that statistics that rely upon low-level features as primary and secondary structure are much more important for the protein retrieval task, than features of the tertiary structure, that is the overall shape of the protein.

References

[1] 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*, J. Mol. Biol. 247, pp. 536-540, 1995.

[2] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhata, H. Weissig, I.N. Shindyalov and P.E. Bourne, *The Protein Data Bank*, Nucleic Acids Research, Vol. 28, pp. 235-242, 2000.

[3] M. Temerinac, M. Reisert and H. Burkhardt, *Invariant Features for Searching in Protein Fold Databases*, International Journal on Computer Mathematics , 'Special Issue on Bioinformatics', to appear 2007.

[4] D. Frishman, P. Argos *Knowledge-Based Protein Secondary Structure Assignment*, Proteins: Structure, Function, and Genetics 23:566-579, 1995

[5] P. Daras, D. Zarpalas, A. Axenopoulos, D. Tzovaras and M.G. Strintzis, *Three-Dimensional Shape-Structure Comparison Method for Protein Classification*, IEEE/ACM transactions on Computational Biology and Bioinformatics, Vol. 3, No. 3, pp. 193-207, July 2006.

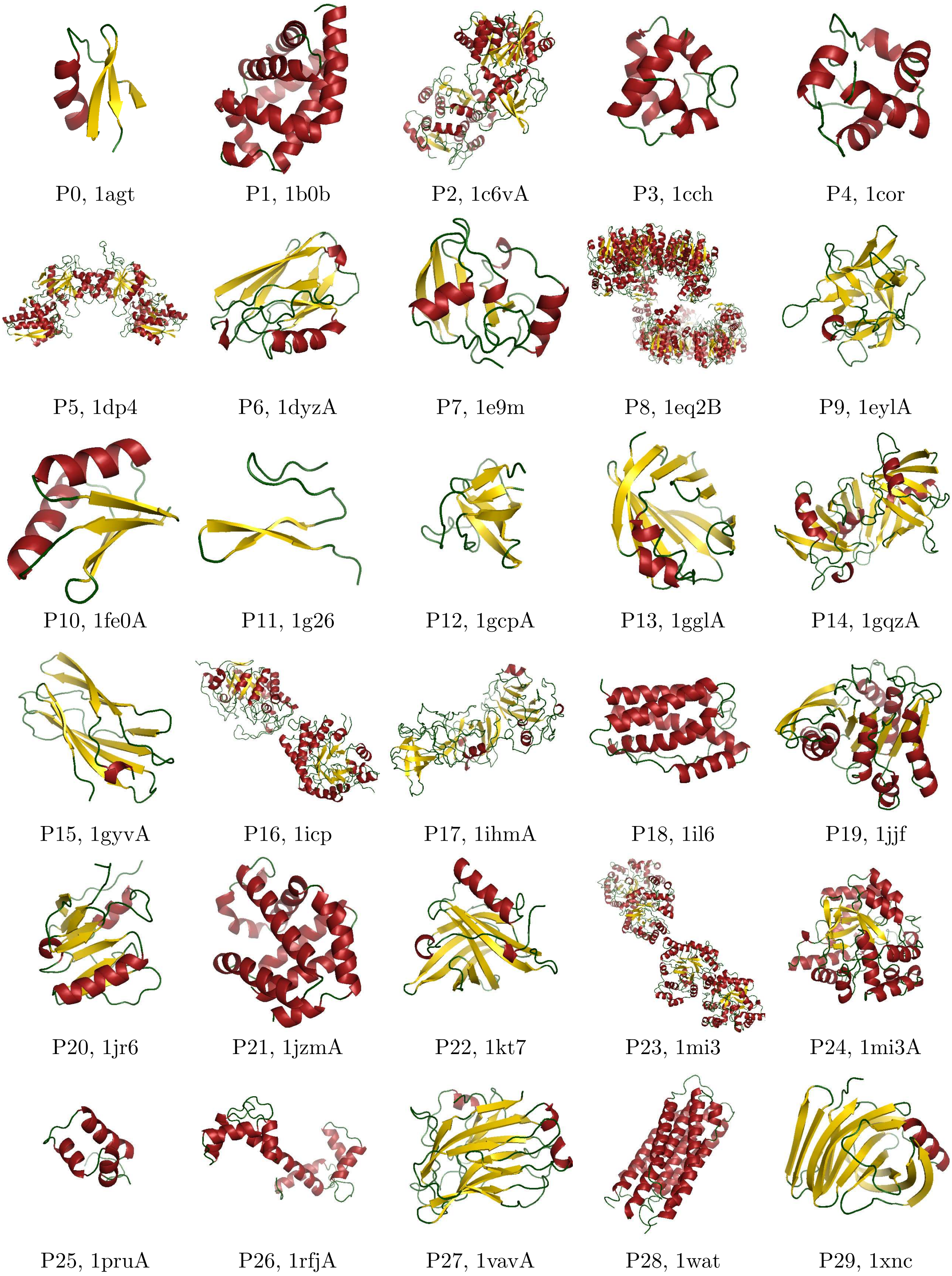

Figure 2. The evaluation dataset.