Classifying Protein Folds using Multi-Level Information of Protein Structures

Zeyar Aung* and Kian-Lee Tan†
Department of Computer Science, National University of Singapore, 3 Science Drive 2, Singapore 117543.

Abstract

We present a new scheme for classifying three-dimensional (3D) protein folds. It is a 3-step scheme using the different levels of protein structure information in the respective steps. The experimental results show that we can achieve an average accuracy of 98.8% using it. We compare our proposed method against the other two protein structure classification schemes, namely *SGM* and *CPMine*.

Keywords: 3D protein structures classification, abstract representation, filter-and-refine, nearest neighbor classification.

1 Background

A protein fold is a description of the overall 3D shape and the topological arrangement of a protein. Classification of the protein folds is an important task in bioinformatics. Knowledge about the fold class of a protein can give insights into its functions, which is useful in many applications such as drug discovery.

First, we have a database of deposited 3D protein structures whose fold classes have been already determined (e.g., by *SCOP* [7]). After we have newly solved the 3D structure of a protein (by *X-ray Crystallography* etc.), we may want to know which of the known fold classes it belongs to. Then, we use the knowledge of the relationship between the 3D structures and the fold classes of the existing proteins in the database in order to predict the fold class of the new protein.

People usually use the detailed structural alignment tools such as *DALI* [6] and *CE* [9] to look for the structurally similar proteins in the database, and then assign the fold class of the most similar one to the new protein. They are quite accurate, but slow especially when the database involved is large. Alternatively, people use the database search tools such as *PSI* [3] and *ProteinDBS* [4], or the dedicated protein structure classifiers such as *SGM* [8] and *CPMine* [1]. These methods use some form of abstract information (such as multi-dimensional vectors) of the 3D structures, rather than the direct 3D coordinates of them. They are fast enough to handle the large databases, but relatively less accurate.

Our objective is to develop a protein fold classifier that can offer the near accuracy to the detailed alignment methods, while much faster than them (even though not as fast as the dedicated schemes). We propose a multi-step scheme using the relevant type of information and algorithm for each step.

2 Method

2.1 Filtering

In the first step, we represent a 3D protein structure in a very abstract form called *Protein Abstract* (*PA*). It is a 6-tuple consisting of the attributes shown in Table 1.

PA can be used to roughly distinguish a protein from one fold class to that from another. For example, the proteins belonging to All-alpha Class have very high *Helix ratio* and *Helix count ratio* values as opposed to All-beta Class proteins which have very low values of them.

The PA of the query protein is compared against those of all the proteins in the database. The threshold value for each PA attribute is pre-calculated for each distinct fold class in the database. Only the proteins which are similar enough (according to the respective thresholds) to the query PA are passed to the next step. On average, for 71% of the proteins in the database are pruned away in this filtering step.

Table 1.Attributes of Protein Abstract (PA).

Sr.	Attribute
1	No. of amino acid (AA) residues
2	No. of SSEs
3	SSE content (total length of all SSEs as a ratio of no. of residues)
4	Helix ratio (total length of all helices as a ratio of total SSE length)
5	Helix count ratio (no. of helices as a ratio of no. of SSEs)
6	SSE sequence (string of H's and E's)

2.2 Refinement

In the second step, we represent a protein structure as a relatively more detailed yet still abstract structure called a *CPset* (*Contact Pattern Set*). A *Contact Pattern* (*CP*) is a description about the interaction of a pair of SSEs (secondary structure elements) in a protein. A CP formed by the interaction of two SSEs *a* and *b* consists of the attributes as shown in Table 2.

Generally, the fold class of a protein can be determined by the types (helix – H or sheet – H), forms and arrangements of its constituent SSEs. These features of the interacting SSEs are effectively captured in CP representation. The proteins belonging to the same fold class have a greater number of similar CP pairs than the proteins from the different fold classes.

Table 2. Attributes of Contact Pattern (CP).

Sr.	Attribute	Upper Bound	#Bins
1	Type of CP (HH, HE, EH, EE)	3	4
2	Difference between starting positions of *a* and *b* in AA sequence	800	8
3	Difference between positions of *a* and *b* in SSE sequence	48	12
4	Angle between *a* and *b*	180.0	16
5	Distance between midpoints of *a* and *b*	50.0	2
6	Nearest vertex-pair distance of *a* and *b*	50.0	8
7	Other vertex-pair distance of *a* and *b*	80.0	2
8	Mean of Cα – Cα distances in CP	64.0	8
9	Standard deviation of Cα – Cα distances in CP	16.0	2
10	Contact density of CP	1.0	2

A protein with n SSEs contains $n(n-1)/2$ distinct CPs. We represent each CP as a 23-bit integer by discretizing and concatenating its attribute values. Thus, a protein structure can be encoded as an ordered set of integer-valued CPs which is the CPset. Discretization of CPs loses some information; but the discretized CPs still maintain sufficient information to serve the

*e-mail: zeyaraun@comp.nus.edu.sg
†e-mail: tankl@comp.nus.edu.sg

purpose of structural classification. Discretization allows compact representation of the CPs which enables efficient processing, as well as the approximate matching of the original CPs by simply the exact matching of their discrete versions.
The CPsets of proteins filtered from the first step are compared against that of the query using a linear-time merging algorithm. The similar score between two CPsets is calculated based on the number of common CPs they contain.
The coarse score for a protein is calculated taking both its PA similarity score and CPset similarity score into account.

2.3 Alignment

In the final step, the best scoring protein (according to the coarse score) from each fold class is aligned with query using *DALI* alignment tool [6]. (We use the standalone version of DALI called *DaliLite* [5].) The detailed 3D coordinates of the proteins are used in this alignment step. The fold class of the best scoring protein (according to DALI) is returned as the answer.
The number of proteins to be aligned with DALI is at most the number of all distinct fold classes. On average, only 1.6% of the proteins in the database are needed to be aligned.
The overview of the proposed scheme is illustrated in Figure 1.

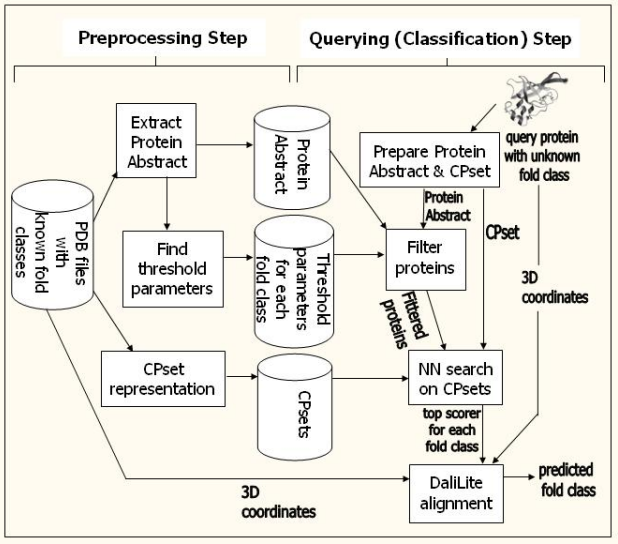

Figure 1. Overview of the new 3-step scheme.

3 Results and Discussion

We conduct a 10-fold cross-validation classification experiment on a database with 600 proteins. We select 15 Folds each with 40 members (with less than 40% sequence homology) from *ASTRAL* database [2]. The same experiment is also done on pure DALI search (i.e. without step 1 and 2), *SGM* [8] and *CPMine* [1]. The experimental results are benchmarked against the widely-used *SCOP* [7] manual classification system. The results show that the 3-step scheme offers an average accuracy of 98.8% if we take the top 3 scorers into account and 94.7% if we take only the topmost scorer into account. The average time and accuracy comparison of the methods are shown in Figure 2.
The new scheme is 6.7 times faster than the pure DALI search whilst providing the very close accuracy. (The time reduction is not proportional to the percentage of proteins aligned in the third step, since DALI search also applies its own filtering mechanism.)
The accuracy of the new scheme is better than those of SGM and CPMine. The 3-step scheme is slower than them. However, by executing only the first two steps, we can make it faster than SGM, and as fast as CPMine. Although the accuracy of this 2-step process is somewhat lower than that of the original 3-step scheme, it is still better than those two schemes.

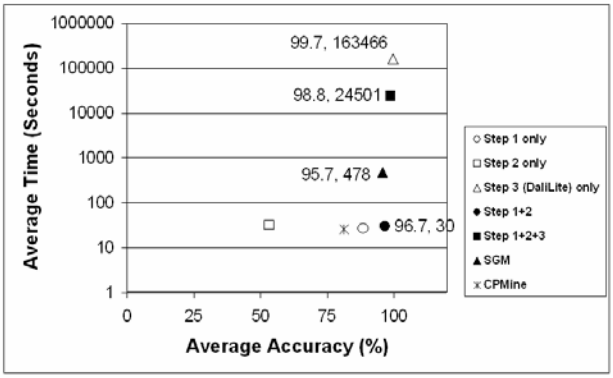

Figure 2. Average time vs. accuracy comparison of methods.

4 Conclusion

We have presented a new multi-step scheme for protein fold classification. The experimental result shows that it is both accurate and efficient. After some future works, it can become a useful tool in the age of very large protein structure databases.

References

1. AUNG, Z. AND TAN, K. L. 2004. Automatic protein structure classification through structural fingerprinting. In *Proc. 4th IEEE Symposium on Bioinformatics and Bioengineering*, pages 508–515.

2. BRENNER, S. E., KOEHL, P., AND LEVITT, M. 2000. The ASTRAL compendium for sequence and structure analysis. *Nucl. Acids Res.* 28, 254–256.

3. CAMOGLU, O., KAHVECI, T., AND SINGH, A. K. 2003. PSI: indexing protein structures for fast similarity search. *Bioinformatics* 19, 81i–83i.

4. CHI, P. H., SCOTT G., AND SHYU C. R. 2004. A fast protein structure retrieval system using image-based distance matrices and multidimensional index. In *Proc. 4th IEEE Symposium on Bioinformatics and Bioengineering*, pages 522–529.

5. HOLM, L. AND PARK. J. 2000. DaliLite workbench for protein structure comparison. *Bioinformatics* 16, 566–567.

6. HOLM, L. AND SANDER, C. 1993. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* 233, 123–138.

7. HUBBARD, T. J. P., AILEY, B., BRENNER, S. E., MURZIN, A. G., AND CHOTHIA, C. 1997. SCOP: a structural classification of proteins database. *Nucl. Acids Res.* 25, 236–239.

8. RØGEN, P. AND FAIN, B. 2003. Automatic classification of protein structure by using Gauss integrals. *Proc. Natl. Acad. Sci. USA* 100, 119–124.

9. SHINDYALOV, I. N. AND BOURNE, P. E. 1998. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.* 11, 739–747.