Structure-Based Feature Extraction from Protein Databases

Gabriela Hristescu*
Department of Computer Science
Rowan University
Glassboro, NJ 08028, U.S.A.
hristesc@elvis.rowan.edu

Martin Farach-Colton*
Department of Computer Science
Rutgers University
Piscataway, NJ 08855, U.S.A.
farach@cs.rutgers.edu

Abstract

In this paper, we study the performance of the feature extraction method we developed for complex object databases, called Complex Object Feature Extraction *(*COFE*), with respect to protein datasets, using distance measures based on structural similarity between proteins. We first perform an assessment of the accuracy of six automatic protein comparison methods against the manually constructed classification of proteins, called SCOP. We then compare the quality of the feature spaces resulting from applying our developed feature extraction method against those obtained when a previously proposed method is used. The results on the considered dataset for five different structure-based distance spaces show that* COFE *provides significantly higher quality embeddings for four of them. We conclude that* COFE *proves to be a practical method for extracting high quality features from protein databases.*

1 Introduction

Despite the larger amount of information and faster accessibility to protein sequence databases, protein sequences do not capture sufficient enough information to accurately identify biological relationships between proteins. Research has shown [1] that, for more distantly related proteins, the sequences may have diverged to such an extent that structural comparison is needed in order to detect biological significant protein relationships. It is therefore important to extend similarity retrieval systems in protein databases to use comparison methods at the structural level. In recent years, a large number of automatic structural comparison methods have been proposed [15, 20, 8, 19, 3, 5, 6, 22, 14], each providing one or more measures of similarity or dissimilarity between proteins.

*Supported in part by NSF Award CCR-9820879.

In this paper, we are interested in studying the performance of the feature extraction method we developed for complex object databases, called *Complex Object Feature Extraction* (COFE) [9], with respect to protein datasets using distance measures based on structural similarity between proteins. We compare the quality of the resulting feature spaces against those obtained when a previously proposed feature extraction method, called FastMap [4], is applied. In the process of performing this study, we also evaluate five structural comparison methods. These structural comparison methods were selected from among the most used and referred to methods in the structural bioinformatics literature for which we could obtain the code. The evaluation method relies on the Cluster Preservation Ratio (CPR) [10] measure and the Structural Classification of Proteins (SCOP) [16], which was constructed manually by biologists on the basis of structural and evolutionary relationships between proteins.

2 Structure vs. Sequence

In a recent study [12], Levitt and Gerstein showed that statistics for sequence and structure comparison on pairs of proteins known to be related distantly indicate that structural comparison is able to detect almost twice as many distant relationships as sequence comparison. The same study also found that there are very few pairs with significant similarity in terms of sequence but not structure, whereas many pairs have significant similarity in terms of structure but not sequence. The higher sensitivity of protein similarity at the structural level, together with the increasing amount of information known about protein structures, have necessitated the development of robust automatic structural comparison methods.

3 Protein Structural Comparison Methods

Numerous methods have been proposed which compare proteins at the structural level and measure the degree of structural similarity between them [22]. They attempt to optimize a global alignment of the structures with respect to some similarity measure. However, there is no consensus in the structural bioinformatics community as to which approach or, for that matter, which measure of protein similarity is best. These methods cover a large spectrum of techniques for comparing proteins at the structural level. Some use an initial sequence superimposition to guide the structural alignment [15, 20]. Other approaches are based on comparing either scalar [8, 11] or vector [19] distance plots. Another method minimizes the soap-bubble surface area between protein backbones, like MINAREA [3]. Another proposed technique applies dynamic programming to pairwise distances between proteins [5]. Several other methods propose refining superimpositions of secondary structure elements (SSE's), like TOP [14], LOCK [22] and others [6]. The CE method [21] builds an alignment between two protein structures through a combinatorial extension of an alignment path defined by aligned fragment pairs. The THREECA method [24] relies on Geometric Hashing to compute the similarity between two proteins. More detailed expositions of the various types of structural comparison methods, as well as their differences and drawbacks, can be found in [20, 17, 3].

For our study we chose six protein comparison methods that can be installed and run locally. Five are based on structural comparison between proteins (TOP [14], LOCK [22], CE [21], MINAREA [3], THREECA [24]) and the sixth is the sequence comparison method based on the Smith-Waterman [23] algorithm.

4 Feature Extraction Methods

The *Complex Object Feature Extraction* (COFE) method [9] is a scalable method we have developed for extracting high quality features from complex object databases, like protein databases. The dimensionality of the resulting embedding can further be reduced using a greedy resampling method. A complete exposition of the method, including a detailed analysis and evaluation can be found in [9].

The FastMap method proposed by Faloutsos and Lin [4] is a feature extraction method that works both on data objects having a vector representation and on datasets characterized by inter-object (dis)similarities. It is therefore applicable to protein databases endowed with a distance function between proteins.

To compute the accuracy of a protein comparison method or a feature extraction method we use a measure that we call the *Cluster Preservation Ratio* (CPR) [10], which computes how well biologically significant clusters, as defined by a classification, are preserved by the embedding. We define the CPR of a query protein q_i to be:

$$CPR(q_i) = \frac{|E_{q_i} \cap C_{q_i}|}{|C_{q_i}|}$$

where:

- q_i is a query protein in cluster C_{q_i} in the original space
- $|C_{q_i}| = m_{q_i}$ and is defined by the classification
- E_{q_i} is the set of closest m_{q_i} proteins to q_i in the embedded space

We also investigate whether or not the closest k proteins to a query protein belong to the same class of proteins, as given by the manually constructed protein classification. This means a relaxation in the definition of the *Cluster Preservation Ratio* measure, by having $m_{q_i} = min\{|C_{q_i}|, k\}$. We therefore also consider the 5 closest (5CPR) and 3 closest (3CPR) proteins measures.

Out of the six considered protein comparison methods, four of them, namely TOP, LOCK, THREECA and Smith-Waterman, provide similarity scores between proteins while only two of them, CE and MINAREA, provide dissimilarity scores. Because both embedding methods rely on distances between proteins, for those automatic protein comparison methods providing similarity information, we transform the similarity score into a dissimilarity (distance) score. In addition to a protein distance function

$$d(A,B) = s(A,A) + s(B,B) - 2s(A,B)$$

proposed by Linial et al. [13], we also consider the inverse function and choose the best one for each method:

$$d(A,B) = \begin{cases} \frac{1}{s(A,B)} & \text{if A and B are different} \\ 0 & \text{if A and B are the same} \end{cases}$$

5 Methodology

We are interested in a method that gives an overall good (dis)similarity measure to use in the feature extraction method. This is why we need a way to evaluate the accuracy of a protein structure comparison method globally. We therefore use as measure the average *Cluster Preservation Ratio* [10] over a set of queries.

In devising the experiments we are faced with two important questions: how to choose the test dataset, and which classification of proteins to use. We selected the dataset to contain structurally related proteins with low sequence similarity. We chose to use the Structural Classification of Proteins (SCOP) [16], which is considered in the structural bioinformatics community to be the "gold standard" in defining the structural relationship between proteins from the Protein Data Bank (PDB) [2]. We use the PDB_SELECT [7] database, which provides lists of protein structures with sequence similarity within specified thresholds, to select a set of proteins belonging to certain SCOP defined families.

We selected a dataset of 125 protein structures from the 25% threshold PDB_SELECT list. The structures were chosen in such a way that 105 of them can be grouped in 11 subsets of structures of sizes varying from 6 to 22, where all structures from a subset belong to the same SCOP fold. The rest of 20 structures belong to SCOP folds not common to any other structure in the chosen dataset.

In Section 6 we perform an assessment of the accuracy of the six automatic protein comparison methods against the manually constructed classification of proteins, SCOP [16], and in Section 7 we analyze the performance of COFE on structure-based and sequence-based protein distance spaces and compare it with FastMap.

6 Structural Method Comparison

Figure 1 presents the results for the average *Cluster Preservation Ratio* (CPR) measure and its relaxations, the 5CPR and 3CPR, for the 105 queries. Three methods, TOP, LOCK and CE, show consistently higher accuracy in preserving the biological clusters as defined by the SCOP classification of proteins. As expected, concurring with the way the dataset is chosen, the Smith-Waterman method performs poorly. This confirms the fact mentioned earlier that sequence analysis does not provide sufficiently accurate information for similarity querying.

An interesting question is to what extent the methods complement each other or overlap in the results for the individual queries. In other words, what is the chance that a subset of these methods, combined, can perform better than any of the individual methods separately. To answer this question, we need to look beyond the global result reported in Figure 1.

Figures 2 and 3 further detail this comparison. Figure 2 shows the number of queries for which each method gives the best overall result (the highest CPR value over all the methods). Figure 3 reports for how many queries the entire cluster is preserved (CPR has a value of 100%).

Out of the 105 queries, 35 have the maximum value of the CPR as 100%. This represents 33.33% of the queries. We also notice that all these 100% values belong to one of the following 3 methods: CE, LOCK and TOP. 31 of these values are given by LOCK and TOP alone.

For 93 queries, the maximum value of the 5CPR is 100%. This represents 88.57% of the queries. We also notice that all these 100% values belong to one of the following 3 methods: CE, LOCK and TOP. 90 of these values are given by LOCK and TOP alone, 84 by CE and LOCK and 82 by CE and TOP.

For 99 queries, the maximum value of the 3CPR is 100%. This represents 94.28% of the queries. We also notice that all these 100% values belong to one of the following 3 methods: CE, LOCK and TOP. All 99 of these values are given by LOCK and TOP alone, 96 by CE and LOCK and 95 by CE and TOP.

7 Feature Extraction Comparison

We ran both the COFE and the FastMap methods on the selected dataset to produce 36-dimensional embeddings, apply a greedy resampling heuristic [9] on the extracted features and use the embedding with highest quality to represent the method in the comparison.

Figure 4 summarizes the results of comparing the quality of the embeddings for the two methods, COFE and FastMap, with the accuracy of the (dis)similarity measures between proteins produced by the protein comparison methods. For those

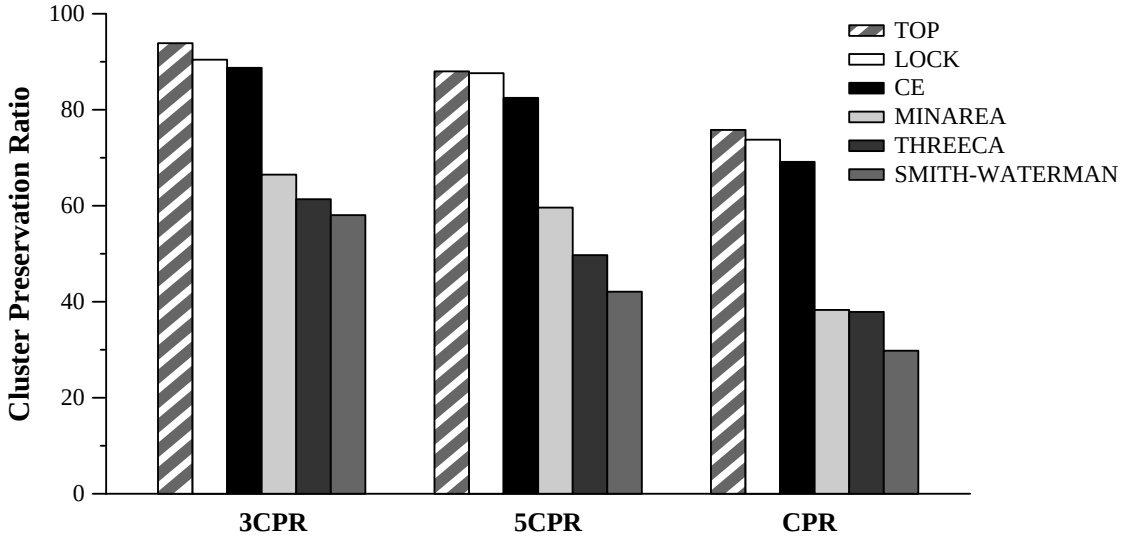

Figure 1: Average Cluster Preservation Ratio for six protein comparison methods.

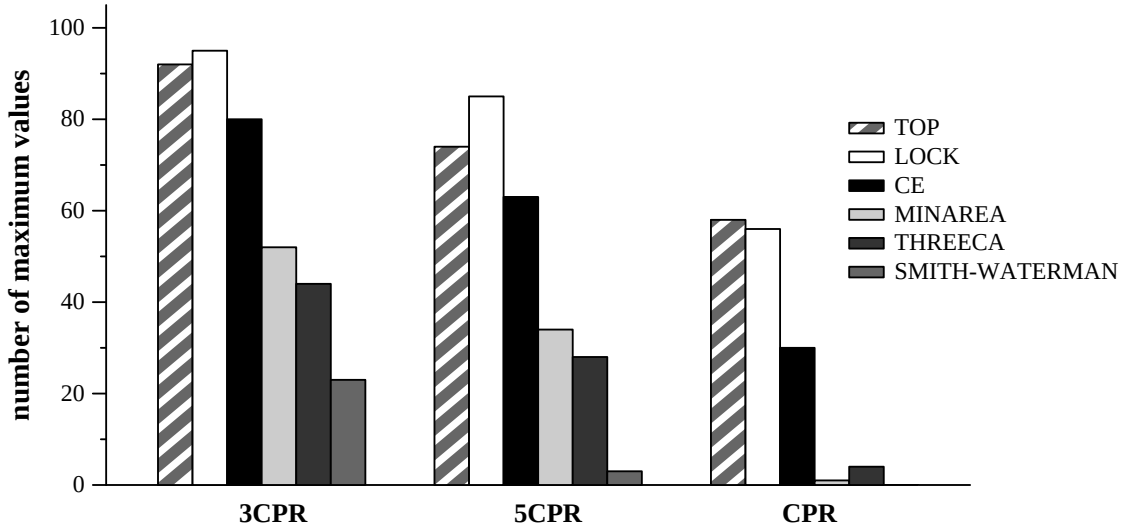

Figure 2: Number of maximum values for six protein comparison methods.

automatic protein comparison methods providing similarity information, we chose to compare the results for the method using the protein distance function that produces the highest quality embedding. We noticed that on the considered dataset, COFE works better with the inverse function for scores provided by most protein comparison methods, while FastMap performs better using the Linial et al. distance function.

Results show that for the three protein comparison methods that show higher accuracy in preserving the clusters, namely TOP, LOCK and CE, both COFE and FastMap show very good overall quality, although the embedding produces a loss in quality. COFE outperforms FastMap for all three methods. The loss in quality compared to the accuracy of the embedding methods is 32% for COFE and 40% for FastMap compared to the TOP method, 12% for COFE and 17% for FastMap compared to the LOCK method, and 25% for COFE and 41% for FastMap compared to the CE method. On the other hand, for the three protein comparison methods that already show low accuracy in preserving the clusters, namely MINAREA, THREECA and Smith-Waterman, the COFE method improves the quality of the embedding. The gain in the quality of the COFE embedding compared to the accuracy of the methods is 5% for MINAREA, 22% for THREECA and 14% for the Smith-Waterman method. FastMap also shows a substantial gain in

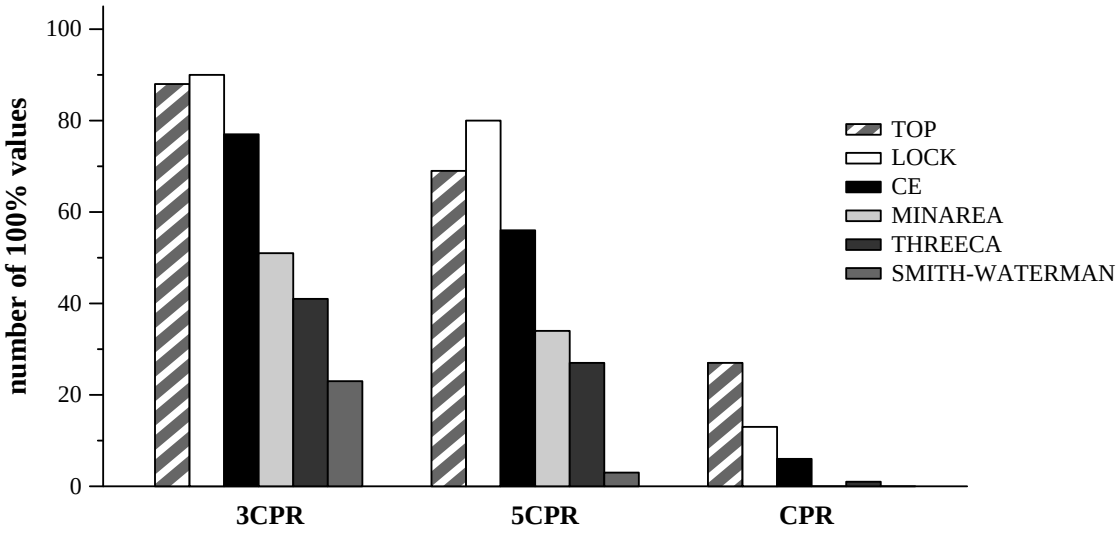

Figure 3: Number of 100% values for six protein comparison methods.

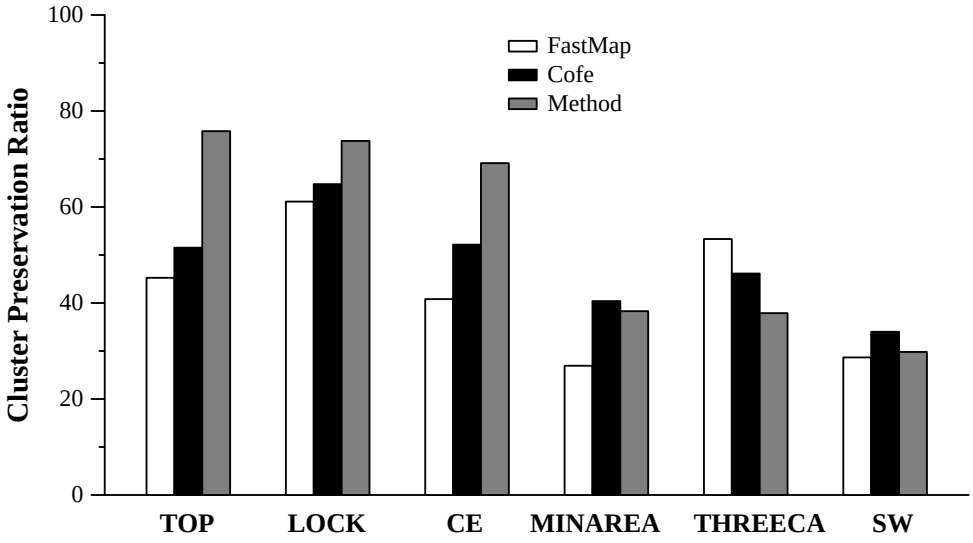

Figure 4: Comparing the CPR for the methods, the FastMap and COFE embeddings.

the quality of the embedding compared to the accuracy of the THREECA method by 41%.

8 Conclusions

An initial assessment of the accuracy of six automatic protein comparison methods against the manually constructed classification of proteins, SCOP [16], indicate that three structural protein comparison methods show higher accuracy in preserving biologically significant clusters, namely TOP, LOCK and CE. As expected, the Smith-Waterman method performs poorly, concurring with the way the dataset was chosen and with the theory that sequence analysis does not provide sufficiently accurate information for similarity querying.

We then performed a comparative experimental evaluation of our developed method, COFE, and a previously proposed method on sequence and structure-based protein distance spaces. The results on the considered dataset for five different structure-based distance spaces show that COFE provides significantly higher quality embeddings for four of them. We conclude that COFE proves to be a practical method for extracting high quality features from protein databases.

References

[1] C. Chothia and A.M. Lesk. The Relation between the Divergence of Sequence and Structure in Proteins. *EMBO J*, 5:823–826, 1986.

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

[3] A. Falicov and F. Cohen. A Surface of Minimum Area Metric for the Structural Comparison of Proteins. *Journal of Molecular Biology* 258:871–892, 1996.

[4] C. Faloutsos and King-Ip Lin. FastMap: A Fast Algorithm for Indexing, Data-mining and Visualization of Traditional and Multimedia Datasets. *ACM SIGMOD*, 24(2):163–174, June 1995.

[5] M. Gerstein and M. Levitt. Using Iterative Dynamic Programming to Obtain Accurate Pairwise and Multiple Alignments of Protein Structures. *Proc. 4th Intl. Conf. on Intelligent Systems for Molecular Biology*, 59–67, 1996.

[6] J.-F. Gibrat, T. Madej, and S.H. Bryant. Surprising Similarities in Structure Comparison. *Current Opinion in Structural Biology*, 6:377–385, 1996.

[7] U. Hobohm and C. Sander. Enlarged Representative Set of Protein Structures. *Protein Science* 3:522–524 ,1994.

[8] L. Holm and C. Sander. Protein Structure Comparison by Alignment of Distance Matrices. *Journal of Molecular Biology*, 233:123–138, 1993.

[9] G. Hristescu and M. Farach-Colton. COFE: A Scalable Method for Feature Extraction from Complex Objects. *Proc. 2nd International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2000)*, 358–371, 2000.

[10] G. Hristescu and M. Farach-Colton. Cluster-preserving embedding of proteins. *DIMACS Technical Report TR-99-50*, Rutgers University, 1999.

[11] U. Lessel and D. Schomburg. Similarities between Protein 3-D Structures. *Protein Eng.*, 7:1175–1187, 1994.

[12] M. Levitt and M. Gerstein. A Unified Statistical Framework for Sequence Comparison and Structure Comparison. *Proc. Natl. Acad. Sci.*, 95:5913–5920, 1998.

[13] M. Linial, N. Linial, N. Tishby, and G. Yona. Global Self Organization of All Known Protein Sequences Reveals Inherent Biological Signatures. *Journal of Molecular Biology*, 268:539–556, 1997.

[14] G. Lu. TOP: A New Method for Protein Structure Comparisons and Similarity Searches. *Journal of Appl. Cryst.*, 33:176–183 , 2000.

[15] B.W. Matthews and M.G. Rossmann. Comparison of Protein Structures. *Meth. Enzymol.*, 115:397–420, 1985.

[16] 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.

[17] C.A. Orengo. Classification of Protein Folds. *Curr. Op. in Structural Biology*, 4:429–440, 1994.

[18] C.A. Orengo and W.R. Taylor. A Local Alignment Method for Protein Structure Motifs. *Journal of Molecular Biology*, 233:488–497, 1993.

[19] C.A. Orengo and W.R. Taylor. SSAP: Sequential Structure Alignment Program for Protein Structure Comparison. *Meth. in Enzym.*, 266:617–635, 1996.

[20] R.B. Russell and G.J. Barton. Multiple Protein Sequence Alignment from Tertiary Structure Comparison: Assignment of Global and Residue Confidence Levels. *Proteins: Struct. Funct. Genet.* 14(2):309–23, 1992.

[21] I.N. Shindyalov and P.E. Bourne. Protein Structure Alignment by Incremental Combinatorial Extension (CE) of the Optimal Path. *Protein Engineering*, 11(9):739–747, 1998.

[22] A.P. Singh and D.L. Brutlag. Hierarchical Protein Structure Superposition using both Secondary Structure and Atomic Representations. *Fifth Intl. Conf. on Intelligent Systems for Molecular Biology*, Menlo Park, CA, 284–293, 1997.

[23] T. Smith and M. Waterman. The Identification of Common Molecular Subsequences. *Journal of Molecular Biology*, 147:195–197, 1981.

[24] C.J. Tsai, S.L. Lin, H. Wolfson, and R. Nussinov. A Dataset of Protein-Protein Interfaces Generated with a Sequence-order-independent Comparison Technique. *Journal of Molecular Biology*, 260:604–620, 1996.