Efficient Similarity Retrieval for Protein Binding Sites based on Histogram Comparison

Thomas Fober[1*], Marco Mernberger[1*], Gerhard Klebe[2] and Eyke Hüllermeier[1]
[1]*Department of Mathematics and Computer Science*
[2]*Department of Pharmacy*
Philipps-Universität, 35032 Marburg, Germany

Abstract: We propose a method for comparing protein structures or, more specifically, protein binding sites using a histogram-based representation that captures important geometrical and physico-chemical properties. In comparison to hitherto existing approaches in structural bioinformatics, especially methods from graph theory and computational geometry, our approach is computationally much more efficient. Moreover, despite its simplicity, it appears to capture and recover functional similarities surprisingly well.

1 Introduction

With the steady improvement of structure prediction methods, the inference of protein function based on structure information becomes more and more important. The comparison of protein structures, for which quite a number of methods have already been proposed, is a central task in this regard. One class of methods focuses on geometrical aspects and, correspondingly, makes use of tools from computational geometry. As examples of this type of approach, we mention geometric hashing [RW97] and labeled point cloud superposition [FH09]. Another idea is to use graphs as formal models of molecular structures. Here, the focus is more on the physical and chemical properties, which are often modeled as nodes of a graph, while geometrical or topological properties are captured in a more indirect way via weighted edges. Typical examples of this approach include measures based on sub-graph isomorphism [NB07], graph edit distance [FMKH09], and graph kernels [G08].

Geometrical and graph-based approaches are appealing, especially since they produce more than just a numerical degree of similarity. Usually, they also provide useful extra information, e.g., correspondences between basic structural units. The price to pay is a high computational complexity. In fact, many of the aforementioned methods lead to NP-hard optimization problems and scale poorly with the size of the structures. This complexity prevents them from being used in large-scale studies like cluster analysis requiring all-against-all comparisons.

*The first two authors contributed equally to this work.

A possible alternative to methods of the above kind is offered by *feature-based* approaches in which a protein structure is first represented in terms of a fixed number of features, that is, a vector of fixed dimensionality. The comparison of objects is thus reduced to the comparison of feature vectors. Since the original object cannot be recovered from a finite number of features, this transformation normally comes with a significant loss of information. Consequently, it is unclear to what extent the similarity of the original structures is mirrored by the similarity of their respective feature vectors. On the other hand, this approach has an obvious advantage with regard to complexity, as feature vectors can be compared quite efficiently.

In this paper, we propose a feature-based approach to the comparison of protein binding sites. More specifically, our idea is to summarize important information about the geometrical and physico-chemical properties of protein binding sites in terms of histograms. This idea is largely motivated by the successful use of similar approaches in the field of image processing, where the distribution of the brightness or the colors of a picture are represented in terms of histograms [RTG00, VB00]. A similar approach has also been applied in the field of structural bioinformatics [SSS$^+$07] for the analysis of homologous proteins.

2 Modeling Protein Binding Sites

Our approach builds upon CavBase [SKK02], a database for the automated detection, extraction, and storing of protein cavities (hypothetical binding sites) from experimentally determined protein structures. In CavBase, a set of points is used as a first approximation to describe a binding pocket.

The geometrical arrangement of the pocket and its physicochemical properties are first represented by predefined *pseudocenters* – spatial points that represent the center of a particular property. The type and the spatial position of the centers depend on the amino acids that border the binding pocket and expose their functional groups. Currently, CavBase considers seven types of pseudocenters (hydrogen-bond donor, acceptor, mixed donor/acceptor, hydrophobic aliphatic, metal ion, pi, aromatic).

Pseudocenters can be regarded as a compressed representation of areas on the cavity surface where certain protein-ligand interactions are experienced. Consequently, a set of pseudocenters is an approximate representation of a spatial distribution of physicochemical properties.

3 Transforming Protein Binding Sites into Histograms

A histogram h is a partition of a set of observations $\mathcal{O} \subset \mathcal{X}$ into a finite number of discrete units. Formally, h can be represented as a $\mathcal{B} \longrightarrow \mathbb{R}$ mapping, where $\mathcal{B}$ is a finite set of *bins*, and $h(b)$ denotes the number (fraction) of observations falling into bin

b. We call a histogram h normalized if $\sum_{b\in\mathcal{B}} h(b) = 1$. Each bin b is associated with a subset $X[b]$ of the domain $\mathcal{X}$, so that $h(b) = |\mathcal{O} \cap X[b]|$ before normalization and $h(b) = |\mathcal{O}|^{-1}|\mathcal{O} \cap X[b]|$ in the normalized case. The set of bins is assumed to form a partition of $\mathcal{X}$, i.e., $X[a] \cap X[b] = \emptyset$ for $a \neq b$ and $\bigcup_{b\in\mathcal{B}} X[b] = \mathcal{X}$.

To obtain histograms from a protein binding site, we will use two important properties, namely its distribution of pseudocenters and the distribution of distances between pseudocenters, thereby capturing both, the physico-chemical properties as well as the geometry of the binding site.

To combine both pseudocenter and distance information, our representation is based on sets of pairwise distances: $D_{i,j}$ is the set of all distances between pseudocenters of type i and j, with $1 \leq i \leq j \leq n_p$ (n_p denoting the number of pseudocenter types). To obtain a corresponding histogram $h_{i,j}$, we use $\mathcal{B} = \{1, \ldots, d_{\max}\}$ and let $X[b] = [b-1, b[$. All histograms are normalized to ensure equal weights (except empty histograms). Thus, a structure is represented by a set of $n = n_p(n_p+1)/2$ histograms.

4 Distance Measures

Consider two structures represented, respectively, by histograms $g_1, \ldots, g_n$ and $h_1, \ldots, h_n$. Moreover, let δ be a distance measure suitable for comparing histograms. The overall distance between the two structures can then be obtained by aggregating the distances $\delta(g_i, h_i)$, for example in terms of the Euclidean norm of the vector

$$(\delta(g_1, h_1), \ldots, \delta(g_n, h_n)) \quad .$$

In the literature, two types of distance measures on histograms are distinguished, namely *bin-by-bin* and *cross-bin* measures. The former are rather simple and only compare values in the same bin. The distance between two histograms is then defined by the sum of distances for all bins. Cross-bin measures, on the other hand, also compare values in different bins. In order to aggregate these distances, they also require the existence of a *ground distance* on $\mathcal{B}$; in our case, we can simply define $|a - b|$ as distance between bins a and b.

Since cross-bin measures proved superior to bin-by-bin measures in a previous study [FH10], we focus on the former type. More precisely, we consider the Quadratic Form Distance,

$$d_{QF}(g, h) = \sqrt{(\vec{g} - \vec{h})^T A (\vec{g} - \vec{h})} \, ,$$

where A is a matrix whose entries $a_{i,j}$ specify the similarity between bins b_i and b_j with

$$a_{i,j} = 1 - \frac{d_{i,j}}{\max_{i,j}\{d_{i,j}\}} \quad ,$$

the Earth Mover's Distance,

$$d_{EMD}(g,h) = \begin{cases} \min\left\{\sum_{\mathcal{B}_n} f_{i,k} \,|\{f_{i,k} : (i,k) \in \mathcal{B}_n\}\right\} & \\ \text{subject to:} & \\ \sum_{k:(i,k)\in\mathcal{B}_n}(f_{i,k} - f_{k,i}) = g(b) - h(b) & \forall\, b \in \mathcal{B} \\ f_{i,k} \geq 0 & \forall\, (i,k) \in \mathcal{B}_n \end{cases}$$

and Cumulative Distributions. The latter approach replaces the original histogram h by the corresponding cumulative distribution, defined by $H(b) = \sum_{a\leq b} h(a)$, and then measures the distance on these distributions. Here, we use the Kolmogorov-Smirnov distance

$$d_{KS}(g,h) = \max_{b\in\mathcal{B}}\{|G(b) - H(b)|\}$$

and the match distance

$$d_M(g,h) = \sum_{b\in\mathcal{B}} |G(b) - H(b)|.$$

5 Experimental Results

In our experiments, we first used a dataset from a previous study designed to assess the performance of global structural alignment methods. This dataset contains 355 protein binding sites comprising two classes of proteins, ATP binding and NADH binding proteins. Binding sites known to bind the corresponding ligands in similar conformation were derived from CavBase; in case of multiple binding sites belonging to the same structure, only one representative was selected at random. See [FMKH09] for a more thorough description of the dataset.

As a second, more complex dataset (Table 1), we selected a number of different, highly populated functional enzyme classes according to the ENZYME database [BWF$^+$00]. Protein structures belonging to the selected classes were derived from the Protein Data Bank and corresponding cavities where extracted from CavBase.

Since CavBase may contain multiple cavities for the same protein, not all of them being functionally important, we selected only those binding sites that contained at least two residues belonging to the catalytic center of the protein according to the catalytic activity atlas annotation (CSA) version 2.2.12 [PBT04]. In case of multiple instances for the same structure, we took the binding site with the largest number of catalytic residues.

5.1 Classification Performance on a Two-Class Problem

As a first proof-of-concept, we assessed the performance of our distance measure on a two-class classification problem, namely of ATP- versus NADH-binding proteins. More precisely, we used a k-nearest neighbor (k-NN) classifier combined with different cross-bin measures to discriminate the two classes. As performance criteria, we measured the

EC number	Function	Number of proteins
2.1.1.45	thymidylate synthase	153
3.4.21.4	trypsin	373
3.4.23.16	HIV-1 retropepsin	291
3.4.24.27	thermolysin	70
1.9.3.1	cytochrome-c oxidase	233
4.2.1.1	carbonate dehydratase	316
3.4.25.1	proteasome endopeptidase	167
2.6.1.1	aspartate transaminase	106

Table 1: Dataset of 8 different EC classes.

accuracy of the methods in terms of their classification rates (determined through leave-one-out cross validation) as well as their efficiency in terms of runtime.

For comparison, we also applied kernel methods (the shortest path (SP) kernel [BK05], the random walk (RW) kernel [G08] and the fingerprint (FP) kernel [FMM$^+$09]), graph-based methods (the iterative graph alignment (IGA) [WHKK07] and the evolutionary graph alignment (GAVEO) [FMKH09]) and geometric approaches (the labeled point cloud superposition (LPCS) [FH09]).

Table 2 summarizes the results of these approaches. As can be seen, there are clear differences in terms of performance: The highest classification accuracy is achieved by LPCS, followed by the fingerprint kernels. The graph-alignment methods (IGA and GAVEO) perform less strongly, and the worst classification rates are produced by the graph kernels.

The runtime reported in the table includes the time needed for an all-against-all comparison of the 355 structures and the time needed to perform a leave-one-out cross validation. As can be seen, all methods require at least one day.

k	RW	SP	LPCS	FP	IGA	GAVEO
1	0.597	0.606	0.935	0.842	0.766	0.789
3	0.597	0.628	0.916	0.882	0.718	0.766
5	0.597	0.634	0.890	0.873	0.724	0.780
runtime (h)	1149.88	171.14	361.58	35.98	2136.88	> 5000

Table 2: Classification rates and runtime in hours of a k-NN classifier using different values of k and different distance measures.

Table 3 summarizes the results for our histogram approach using different cross-bin distance measures and bins of size 1 (as they will be used in the whole work). Interestingly, the accuracy values are quite high, even outperforming some of the competitor methods, although LPCS still performs best. However, considering the runtime efficiency of the histogram approach, the results show that we can retrieve comparably good results within only a fraction of the time.

k	d_{QF}	d_M	d_{KS}	d_{EMD}
1	0.862	0.865	0.859	0.772
3	0.856	**0.882**	0.854	0.749
5	0.845	0.865	0.837	0.732
runtime (h)	0.785	0.470	0.472	11.53

Table 3: Classification rates of cross-bin measures on the NADH/ATP data set.

Rank	pdb code	Protein	score
1	2ACK	Acetylcholinesterase (AChE)	0
2	1AX9	Acetylcholinesterase (AChE)	0.180
3	1GQS	Acetylcholinesterase (AChE)	0.203
⋮	⋮	⋮	⋮
98	2V98	Acetylcholinesterase (AChE)	0.402
99	1ZGC	Acetylcholinesterase (AChE)	0.404
100	1G6R	Aspartate aminotransferase (mAspAT)	0.405

Table 4: Top ranks retrieved by querying the CavBase with the main pocket of 2ACK. Omitted entries contained exclusively acetylcholinesterases.

5.2 Database Querying

In a second experiment, we applied our approach on the task of querying the complete CavBase for similar structures. Given the simplicity of the approach, one may doubt its suitability for a task of this kind.

We chose the main pocket of acetylcholinesterase from *T. californica* (pdb code: 2ACK) as a query structure. This protein has previously been used to query the CASTp database with a similarity measure that combines structural similarity with evolutionary conservation [BAL03]. Binkowski et al. retrieved further acetylcholinesterase structures on all top ranks, a result they attributed to the uniqueness of the protein structure.

Table 4 shows some results of our query using the match distance. Surprisingly, and despite the simplicity of our approach, the top 99 ranks are exclusively occupied by other acetylcholinesterase structures before the first false positive shows up on position 100. This is consistent with the results of Binkowski et al. and suggests that important information is indeed captured by our histogram representation.

5.3 Discriminating Enzyme Classes

The third experiment investigates whether our approach can be used to discern binding pockets of different enzyme classes. To this end, we selected several highly populated enzyme classes from the Protein Data Bank and calculated the corresponding distance matrix using our histogram approach with the match distance.

k	d_{QF}	d_M	d_{KS}
1	0.941	0.944	0.945
3	0.920	0.919	0.926
5	0.905	0.912	0.916

Table 5: Classification accuracy on the multi-class enzyme dataset.

Since the class information is known, we visualize the distance matrix by means of a heat map, which is shown in Figure 1. Again, it can be seen that important information is captured by the histogram approach, as several classes show a high similarity within the class.

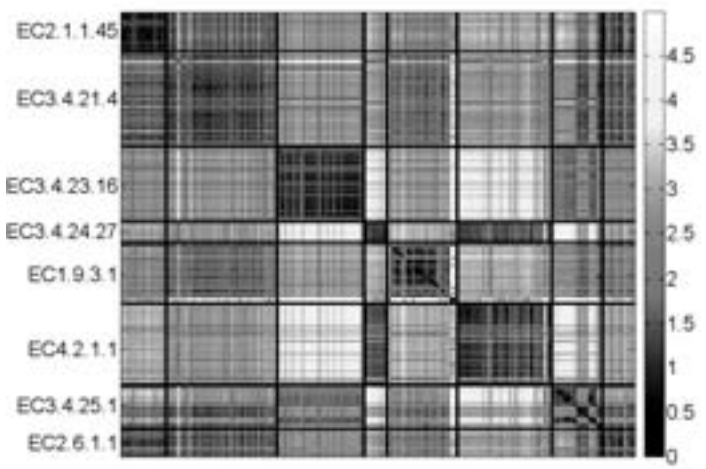

Figure 1: Heat map depicting the distance matrix based on match distance for the EC dataset. Different EC classes are seperated by black lines.

Based on the above distance matrix, we additionally performed a hierarchical clustering using repeated bisection and subsequent k-way refinement. Comparing the resulting clustering with the original EC class yields a Rand index of $R = 0.8633$, indicating that the clustering is in good agreement with the real class structure.

Finally, the distance matrix was again used for a nearest neighbor classification, this time on a multi-class problem. Table 5 shows the classification accuracies for a leave-one-out cross validation, using different distance metrics.

6 Conclusions

In this paper, we have introduced a very simple though extremely efficient method for comparing protein structures in terms of a histogram-based representation. The main interest of the paper is probably less the method itself, but more its strong performance in

our experimental studies on classification and retrieval. In light of the simplicity of the representation and the distinctive loss of information it implies, this performance was unexpected. On the other hand, it is true that similar representations have been used quite successfully in other fields, too, where the loss of information is arguably not smaller.

Due to its runtime efficiency and scalability, our approach is amenable to applications that cannot be tackled by other methods. It can be used as a kind of filter, for example, to preselect structures from very large datasets, thereby reducing the amount of data to be processed afterward by more complex structure comparison algorithms. Using the method for clustering, as we have already done in our experiments, is another example. Indeed, the need for an all-against-all comparison does usually prevent the use of computationally complex methods here.

Acknowledgements: The authors like to thank the reviewers for useful suggestions that helped to improve the paper and, moreover, for bringing the approach of Sander *et al.* to their attention.

References

[BAL03] T.A. Binkowski, L. Adamian, and J. Liang. Inferring Functional Relationships of Proteins from Local Sequence and Spatial Surface Patterns. *Journal of Molecular Biology*, 332(2):505–526, 2003.

[BK05] K.M. Borgwardt and H.P. Kriegel. Shortest-Path Kernels on Graphs. In *International Conference on Data Mining*, pages 74–81, Houston, Texas, 2005.

[BWF+00] 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.

[FH09] T. Fober and E. Hüllermeier. Fuzzy Modeling of Labeled Point Cloud Superposition for the Comparison of Protein Binding Sites. In *Proc. IFSA/EUSFLAT–2009*, pages 1299–1304, Lisbon, Portugal, 2009.

[FH10] T. Fober and E. Hüllermeier. Similarity Measures for Protein Structures based on Fuzzy Histogram Comparison. In *WCCI–2010, World Congress on Computational Intelligence*, Barcelona, 2010.

[FMKH09] T. Fober, M. Mernberger, G. Klebe, and E. Hüllermeier. Evolutionary Construction of Multiple Graph Alignments for the Structural Analysis of Biomolecules. *Bioinformatics*, 25(16):2110–2117, 2009.

[FMM+09] T. Fober, M. Mernberger, V. Melnikov, R. Moritz, and E. Hüllermeier. Extension and Empirical Comparison of Graph-Kernels for the Analysis of Protein Active Sites. In *Lernen, Wissen, Adaptivität*, pages 30–36, Darmstadt, Germany, 2009.

[G08] T. Gärtner. *Kernels for Structured Data*. World Scientific, Singapore, 2008.

[NB07] M. Neuhaus and H. Bunke. *Bridging the Gap between Graph Edit Distance and Kernel Machines*. World Scientific, New Jersey, 2007.

[PBT04] C.T. Porter, G.J. Bartlett, and J.M. Thornton. The Catalytic Site Atlas: A Resource of Catalytic Sites and Residues Identified in Enzymes using Structural Data. *Nucleic Acids Research*, 32:129–133, 2004.

[RTG00] Y. Rubner, C. Tomasi, and L. J. Guibas. The Earth Mover's Distance as a Metric for Image Retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

[RW97] I. Rigoutsos and H. Wolfson. Geometric Hashing. *IEEE Computational Science Engineering*, 4:1070–9924, 1997.

[SKK02] S. Schmitt, D. Kuhn, and G. Klebe. A New Method to Detect Related Function Among Proteins Independent of Sequence and Fold Homology. *Journal of Molecular Biology*, 323(2):387–406, 2002.

[SSS+07] O. Sander, T. Sing, I. Sommer, A.J. Low, P.K. Cheung, P.R. Harrigan, T. Lengauer, and F.S. Domingues. Structural Descriptors of gp120 V3 Loop for the Prediction of HIV-1 Coreceptor Usage. *PLoS Computational Biology*, 3(3):555–564, 2007.

[VB00] C. Vertan and N. Boujemaa. Using Fuzzy Histograms and Distances for Color Image Retrieval. In *Challenge of Image Retrieval*, pages 1–6, Brighton, United Kingdom, 2000.

[WHKK07] N. Weskamp, E. Hüllermeier, D. Kuhn, and G. Klebe. Multiple Graph Alignment for the Structural Analysis of Protein Active Sites. *IEEE Transactions on Computational Biology and Bioinformatics*, 4(2):310–320, 2007.