Journal of Computer-Aided Molecular Design **18:** 697–708, 2004.
DOI 10.1007/s10822-004-6553-x

Comparison of substructural epitopes in enzyme active sites using self-organizing maps

Katrin Kupas[1], Alfred Ultsch[1,*] & Gerhard Klebe[2]
Data Bionics Research Group, Department of Computer Science, University of Marburg, Germany; [2]*Department of Pharmaceutical Chemistry, University of Marburg, Germany*

Received 14 September 2004; accepted in revised form 22 November 2004

Key words: data mining, de novo design, functional comparison of proteins, self-organizing neural networks, U-matrix

Summary

This paper presents a new algorithm to compare substructural epitopes in protein binding cavities. Through the comparison of binding cavities accommodating well characterized ligands with cavities whose actual guests are yet unknown, it is possible to draw some conclusions on the required shape of a putative ligand likely to bind to the latter cavities. To detect functional relationships among proteins, their binding-site exposed physicochemical characteristics are described by assigning generic pseudocenters to the functional groups of the amino acids flanking the particular active site. The cavities are divided into small local regions of four pseudocenters having the shape of a pyramid with triangular basis. To find similar local regions, an emergent self-organizing map is used for clustering. Two local regions within the same cluster are similar and form the basis for the superpositioning of the corresponding cavities to score this match. First results show that the similarities between enzymes with the same EC number can be found correctly. Enzymes with different EC numbers are detected to have no common substructures. These results indicate the benefit of this method and motivate further studies.

Introduction

In a biological system multiple biochemical pathways are proceeded and regulated via the complementary recognition properties of proteins and their substrates. This requires highly conserved molecular recognition patterns on the side of the receptor. It can therefore be assumed that proteins having similar binding cavities also bind similar ligands and exhibit related function. Through the comparison of newly determined proteins with structures which are well known, the function of the new protein and a suitable ligand can be predicted.

The shape and function of a protein is not exclusively represented by one unique amino acid sequence. Proteins with deviating amino acid sequence, even adopting a different folding pattern, can nevertheless exhibit related binding cavities and accordingly related function. Low sequence homology does not imply any conclusions on binding site differences or similarities. For this reason one has to regard the three-dimensional structure as a prerequisite for a reliable comparison of proteins.

In literature, many approaches for the comparison of protein substructures have been presented. The first group of algorithms comprises methods that scan protein structural databases in terms of pre-calculated or automatically generated templates. A typical example for such a template is the catalytic triad in serine proteases. A substantial advantage to restrict to relatively small

*To whom correspondence should be addressed. E-mail: ultsch@mathematik.uni-marburg.de

templates is due to the fact that also large data collections can be scanned efficiently. Some of the best known procedures based on templates are ASSAM introduced by Artymiuk et al. [1, 2], TESS/PROCAT by Wallace et al. [3, 4], PINTS by Stark, Russell et al. [5–7], DRESPAT by Wangikar et al. [8] as well as the methods of Hamelryck [9] and Kleywegt [10]. The second group includes approaches to compare substructural epitopes of proteins which operate independently of any template definition. For the similarity search entire proteins or substructures are used. The group of Ruth Nussinov and Haim Wolfson developed a whole bunch of approaches to compare entire receptor structures or substructures. The individual methods essentially differ in whether the protein structure is represented by their C_α-atoms or grid points on their solvent-accessible surface, or by so-called 'sparse critical points', a compressed description of the solvent-accessible surface. In each case, the different procedures use geometric hashing [11] for common substructure detection. They operate completely independent of any sequence or fold homology. The approach of Rosen et al. [12] permits an automatic comparison of binding cavities. Kinoshita et al. [13] use a graph-based algorithm to compare the surfaces of two proteins. Other methods, such as GENFIT of Lehtonen et al. [14] and the approach of Poirrette et al. [15] use genetic algorithms to optimally superimpose proteins in terms of identified substructure ranges.

All these approaches need either a predefined template or use only a rough approximation for the description of the shape of the protein. But in addition to the shape, it is required to code correctly the exposed physicochemical properties both in a geometrical and also in a chemical sense.

Self-organizing maps are widely used for visualization and clustering of high-dimensional data. Gasteiger et al. use them for the prediction of aqueous solubility of organic compounds [16]. Schneider et al. mapped whole surface cavities of metalloproteinases containing zinc cations in their active sites by a self-organizing neural network and classified them [17].

In this paper, we describe a new algorithm to detect common substructural epitopes across protein binding sites. We use well-placed pseudocenters representative for a small set of physicochemical properties and the mutual distance between two such centers [18]. Every cavity is described with respect to its geometrical and physicochemical properties.

The description of the binding pockets by a set of local pseudocenter assemblies leads to large data sets of high dimensionality. For finding similarities, a cluster algorithm is needed, which can handle large and high-dimensional data samples. Classical clustering algorithms need either a predefined number of clusters or are too time consuming for large data sets. We used a cluster algorithm based on emergent self-organizing maps (ESOM) evaluating both distance and density structure of the data. ESOM is a topology preserving projection from the high-dimensional space onto two dimensions. The number of clusters can be derived from a visualization of this projection. By using ESOM for clustering the local pseudocenter assemblies, an all-against-all comparison also for very large data samples is possible. The generation of easy to understand decision rules for each cluster allows a validation of the clustering.

Similar substructures among sets of spatially arranged pseudocenters of two binding cavities then provide a coordinate system which will be used in the next step for the superimposition of the related cavities. Once a convincing match is detected, it can be assumed that both active sites are capable to bind similar ligands and thus exhibit related function.

The paper is organized as follows: First the underlying theory and concept of our algorithm used for cavity matching are described. The local region in descriptor space based on subsets of pseudocenters is defined and the principles of ESOM clustering are described. Then the data are given and the results obtained for this data set are presented, followed by a summary.

An algorithm for the comparison of binding cavities

For the comparison of substructural epitopes in protein binding cavities the cavities are divided into local regions. Accordingly, the mutual comparison of binding cavities is reduced to a multiple comparison of the segmented local regions. The descriptors of the local regions are their physicochemical properties along with a set of spatial properties. Each local region is described

by a vector of dimension 21 (see below). A projection into two dimensions is used to visualize this high-dimensional data set. Therefore an emergent self-organizing map (ESOM) is used, which is a self-organizing projection from the high-dimensional data space onto a grid of neuron locations [19]. This grid of neuron locations is called a map.

To visualize the information of the map, either the U- or the P-Matrix are used [19, 20]. The U-Matrix shows the distance information among the data. The P-Matrix reflects the density information of the data. The data set of local regions is clustered using the U- and the P-Matrix. To validate the clustering, decision rules are generated for the clusters with the sig*-algorithm [21]. A valid cluster possesses a meaningful decision rule. Local regions within the same cluster are used to align the corresponding binding cavities to score the detected match.

Local regions in binding cavities

The protein binding cavities are characterized in terms of the descriptors developed by Schmitt et al. [18]. The physicochemical properties of the cavity-flanking residues are condensed into a restricted set of generic pseudocenters corresponding to five properties essential for molecular recognition: hydrogen-bond donor (DO), acceptor (AC), mixed donor/acceptor (DA), hydrophobic aliphatic (AL) and aromatic (PI) features. The pseudocenters express the features of the 20 different amino acids in terms of five well-placed physicochemical properties.

Local regions are composed of four such pseudocenters, a center under consideration and its three nearest neighboring centers. Systematically every pseudocenter in a cavity is selected as center under consideration and forms a local region with its three nearest neighbors. Following this procedure, the binding cavities are partitioned into all local regions to be possibly inscribed.

The mutual distances between the four pseudocenters of a local region form pyramids with a triangular basis. Such local regions may be described by, for example, the distances between the four pseudocenters, the perimeters and the areas of the four triangles, the volume, the height and the skewness of the pyramid. The skewness of the pyramid is defined by the distance between the centroid of the basis triangle and the root point of the height. The physicochemical properties of the local regions correspond to the physicochemical properties assigned to the four pseudocenters. A local region is therefore described by a 21-dimensional vector. Every cavity comprises on average about 100 pseudocenters and accordingly 100 local regions.

Emergent self-organizing maps

To visualize high-dimensional data, a projection from the high-dimensional space onto two dimensions is needed. There are many algorithms which project a high-dimensional data space into two or three dimensions. Examples are PCA and ICA for linear projections and MDS and Sammon's Mapping for nonlinear projections. We decided to use ESOM because they have been demonstrated to unfold data sets of complex and intertwined cluster structures [22].

The emergent self-organizing map (ESOM) is a projection onto a grid of neurons, called map. The map of an ESOM preserves the neighborhood relationships of the high-dimensional data [19]. In ESOM a large number of neurons is used. The weight vectors of the neurons are thought of as sampling points of the data. ESOM is able to handle large and high-dimensional data sets.

In order to avoid bordering effects, toroid map grids are used [20]. The often used finite grid as map has the disadvantage that neurons at the rim of the map have very different mapping qualities compared to neurons in the center of the map. This is due to the fact that there is a different number of neighboring neurons in the center vs. the border. It is important during the learning phase and structures the projection.

To visualize toroid maps, four instances of the grid are tiled and displayed adjacently. This is called a tiled display [20]. All figures in the following are tiled displays.

Visualization of the self-organizing map

The U-Matrix has become the canonical tool for the display of the distance structures of the input data on ESOM [19]. The U-Matrix is a display of U-heights on top of the grid positions of the neurons on the map. Small U-heights mean small distances between the data points, large U-heights mean large distances. Accordingly, distance-based

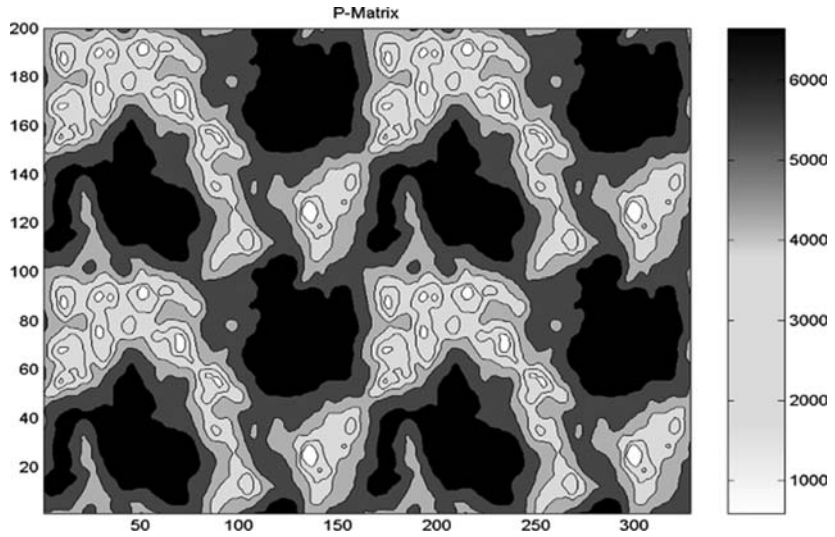

Figure 1. The figure shows a P-Matrix for the data set of local regions from 72 binding cavities described in the text. Regions with a large data density are colored in dark, those with little density are light colored. The P-Matrix is shown with a small number of colors here. A more discriminative coloring scheme allows the identification of density differences insight the dark regions.

clusters in the data set are found in the valleys, mountain ranges point to cluster boundaries.

To reflect the density structure of the data, the P-Matrix has been developed [20]. The P-heights are a measurement for the data density. Large P-heights reflect regions in the data with large density. Small P-heights mean regions with little density. To construct the P-Matrix, for every neuron the points enclosed in a hypersphere around this neuron are counted. The radius of this hypersphere is calculated using information optimal sets [20].

Figure 1 shows a P-Matrix for the data set of local regions from 72 binding cavities described below. Dark colored are regions with high density. Light colored are regions with low density.

Clustering of the local regions

To find similar local regions a distance- and density-based clustering using ESOM is performed. The topology preservation of the underlying SOM algorithm assures that data points lying next to each other on the map are also neighbors in data space. Those falling into different regions on the map originate also from different regions in data space.

Data points found in coherent regions with large P-heights are assigned to one cluster. Smaller heights on the P-Matrix describe the cluster boundaries. Local regions falling into areas with small P-heights are not assigned to any cluster. Accordingly, the local regions are clustered with respect to the density of the data set.

The ESOM clustering has some advantages in comparison to other cluster algorithms. Other cluster algorithms, e.g. K-means, split the data into a preset number of clusters. The ESOM clustering does only produce clusters, if there are clusters in the data set. The number of clusters arises from the analysis. Another advantage is that ESOM can handle very large and high-dimensional data sets. Hierarchical cluster algorithms, for example, are not able to handle such large data sets. They are too time consuming.

Validity of the clusters

Many cluster algorithms such as, for example, K-means split the data set into a preset number of clusters. It is not known, however, whether the generated clusters are meaningful.

We evaluate whether the ESOM clustering is reasonable. Decision rules are generated for every cluster. The rule generation is done with the sig*-algorithm [21]. If the decision rules for the clusters are meaningful, we anticipate that the clustering of the local regions is valid. Then all those local regions falling into the same cluster have similar properties. Those falling into different clusters

have different properties. Local regions originating from different cavities assigned to the same cluster establish a match among these cavities. They are used for the surface superposition of the corresponding cavities to score the detected match.

Scoring of the match

Two local regions found to be similar by the algorithm give rise to a coordinate transformation which optimally superimposes both local regions. This transformation is then applied to the whole cavities. Every pair of pseudocenters from the two cavities falling close to each other with a threshold of 1 Å and having the same physicochemical properties is counted. This number of suitable pairs of pseudocenters is the absolute score for the match.

To filter out matches which are meaningless, a threshold for the score is used. Subpockets in binding cavities have a minimum size of about nine pseudocenters. All matches having an absolute score smaller than nine are not expected to detect a common substructure of the two cavities. Therefore only those matches with an absolute score higher than eight are regarded.

The obtained absolute score is related to the maximum possible score, this is the number of pseudocenters in the smaller one of the two cavities. The procedure is done for all pairs of similar local regions of the two cavities. The maximum of these relative scores is used as total score for the comparison of both cavities.

The data

The data set used in our study consists of 72 binding cavities from enzymes in the PDB (Protein Data Base). These cavities have been generated with the program LIGSITE [23], whose algorithm is implemented in the protein–ligand database RELIBASE [24], and are stored in the database CAVBASE [18].

For all considered proteins an assignment in terms of the six EC numbers is given. For convenience the subclasses corresponding to the different EC numbers have been labelled as 1a to 6b. The corresponding PDB codes of the proteins in this data set are listed in Table A1 in Appendix 1.

The 72 cavities have been divided into local regions. This results in a data set of 8462 local regions. The spatial features of these local regions show some correlations. Correlated features contain the same information. Therefore one of the two correlated features can be omitted.

The distances between the four pseudocenters and the perimeters of the four triangles are correlated. We decided to use the perimeters as the only descriptor for the four triangles. Due to limited accuracy of the crystal structures used to assign the various pseudocenters, deviations in the mutual distances can be cancelled out to some extent in the summation. The volume of the pyramid is given by its height and the area of the corresponding base triangle. The perimeters and the areas of the triangles are to some extent also correlated.

In summary, a local region is described sufficiently by three spatial properties, the area of the basis triangle, the height and the skewness of the pyramid, and the physicochemical properties of the four pseudocenters located at the edges of the pyramid.

Clusters are mainly defined using an appropriate distance metric. In order to use Euclidean distances for the spatial properties the distributions of the features need to be taken into account. Applying a square root transformation to height and skewness of the pyramid and a cubic root transformation to the area of the basis triangle results in an approximate normal distribution. These transformations are used for the preprocessing of the three spatial properties before calculating Euclidean distances. All values have then been standardized.

Figure 2 shows a plot of the three spatial properties after transformation. On first sight no clearly separated clusters concerning only the distances between the data points can be detected. Some regions, however, possess higher density than others. So a cluster algorithm combining distance- and density-information will lead to much better results than a single distance-based one would.

Results

ESOM clustering of the data

The spatial and physicochemical properties were considered separately. A toroid ESOM with

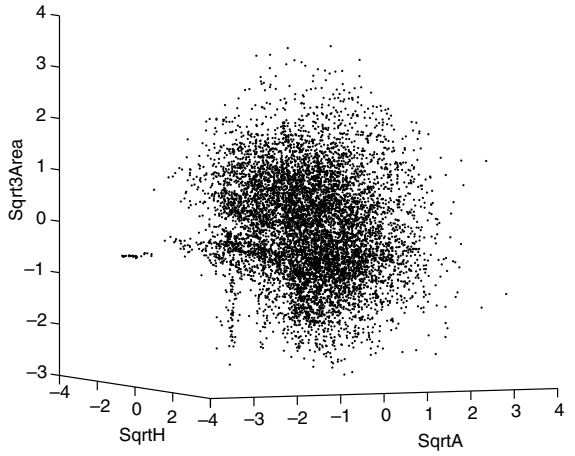

Figure 2. The plot shows the three spatial properties of the local regions after transformation. SqrtH is the square root of the height, SqrtA is the square root of the skewness and Sqrt3Area is the cubic root of the area of the basis triangle. No clearly separated clusters can be detected concerning only the distances between the data points. But there are density differences in the data.

100×164 neurons has been trained with the three spatial features of the 8462 local regions. Figure 3 shows the U-Matrix of the toroid ESOM. Regions with large U-heights (large distances between the data points) have light color, regions with small U-heights (small distances between the data points) have dark color. Clusters in the data should be found in the regions with small U-heights, colored in dark.

The corresponding P-Matrix of the ESOM trained with the three spatial features of the local regions is shown in Figure 1. Regions with high density are shown in dark color, regions with low density are shown in light color. For reasons of clarity, the P-Matrix is shown with a smaller number of colors here. A more discriminative coloring scheme allows the identification of density differences inside the dark regions.

The local regions are clustered as described above. The P-Matrix suggests 16 clusters of local regions. The clusters are found in the dark regions of the U-Matrix, that means regions with small distances between the data points. The U-Matrix with the colored data points of these clusters is shown in Figure 4. Data points having the same color belong to the same cluster. Data points with different color belong to different clusters.

The clusters found were validated using decision rules generated with the sig*-algorithm. As an

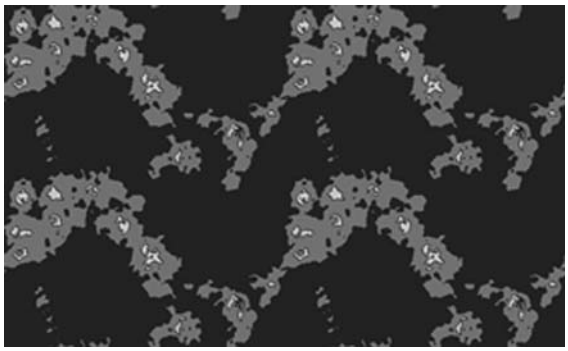

Figure 3. The figure shows the U-Matrix of an ESOM trained with the spatial features of the local regions. Dark grey-shaded are those regions where small distances between the data points occur. Those with large distances between the data points are light grey-shaded. Clusters in the data can be found in the dark regions of the U-Matrix.

example, these rules describe cluster No. 5 as follows: cluster 5:

Area in [−1.549346, −0.9513708]
and
Height in [−0.3348052, 0.2966688]
and
Skewness in [−0.3319888, 0.228202]

Values around zero in the rules mean an average value, positive numbers mean a large value and negative numbers mean a small value. Apparently, cluster No. 5 contains local regions with a very small-sized basis triangle, an average height and skewness.

The intervals for the data distribution of the spatial properties in the cluster decision rules have been translated into the 7 features very small, small, medium small, medium, medium large, large and very large. The conditions for the 16 clusters obtained by this coarse classification are listed in Table 1.

Table 1 shows that the 16 detected clusters are different. Accordingly, the classification of the spatial features is a valid clustering. Cluster No. 3, for example, contains local regions with a small basic triangle, but a very large height and skewness. This means that the local regions in this cluster are very stretched and skewed. One of the centers is distant from the other three centers. The other three centers are close together. Cluster No. 16, however, contains local regions with small area, height and skewness. The four centers are close together. The pyramids are small and straight. Examples for the shapes of local regions

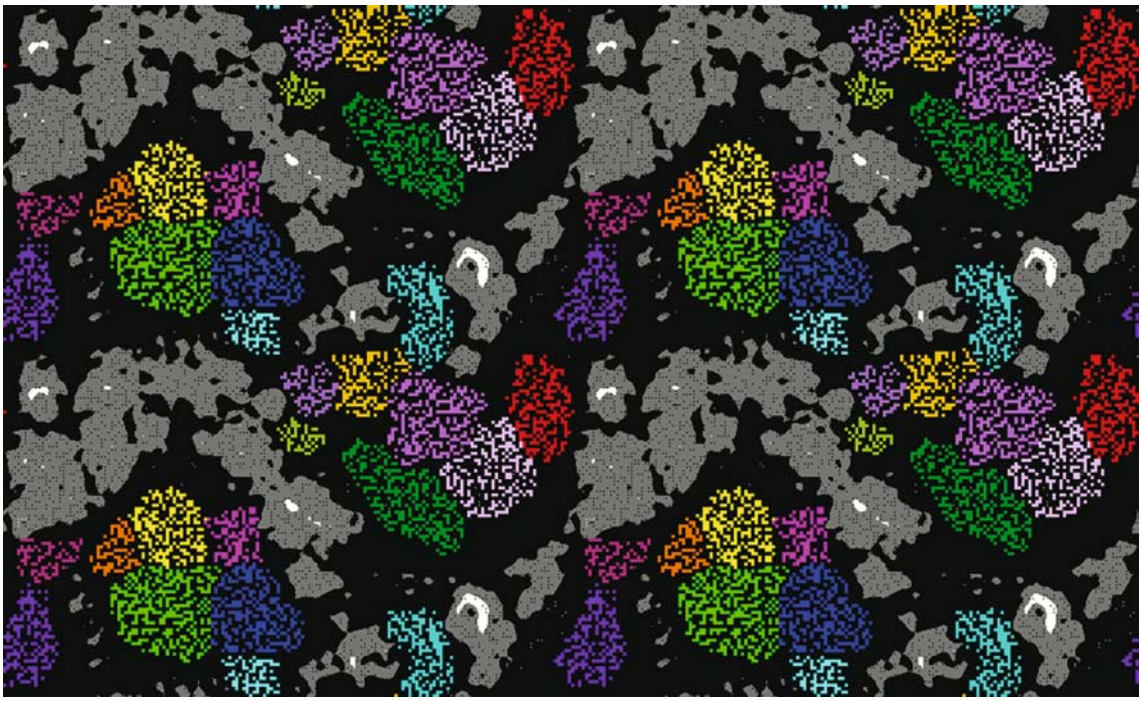

Figure 4. The figure shows the U-Matrix with the colored clusters of local regions generated with the P-Matrix. All data points with the same color belong to one cluster. Data points with different colors belong to different clusters. All clusters are lying in regions with small distances between the data points, and are surrounded by regions with large distances.

found in cluster No. 3 and No. 16 are sketched in Figure 5.

The clusters defined by the spatial properties of the local regions are further subdivided. All local regions possessing the same physicochemical properties in one spatial cluster are assigned to the same chemical subcluster. This results in subclusters of local regions with similar spatial and an identical composition of physicochemical properties.

Table 1. Results of the sig* rules for the different clusters.

Cluster No.	Basis triangle	Height	Skewness
1	medium small	large	large
2	small	medium small	medium large
3	small	large	large
4	medium	medium large	medium
5	very small	medium	medium
6	medium large	medium	large
7	large	medium	medium small
8	small	medium	very large
9	small	medium small	large
10	large	medium	small
11	medium large	medium	medium small
12	small	medium	medium
13	medium	small	medium
14	small	small	medium
15	large	small	medium large
16	small	small	small

Table 2 shows the number of subclusters and the number of each physicochemical property for the center under consideration of the local regions in each of the 16 ESOM clusters.

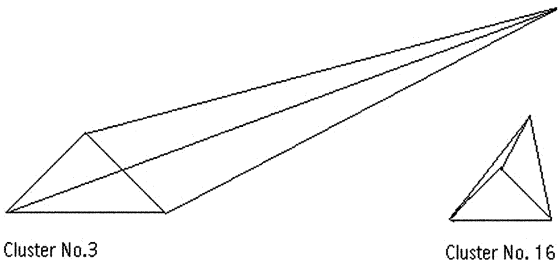

Figure 5. The figure shows the different shapes of local regions from different clusters, as an example from cluster No. 3 and from cluster No. 16.

Table 2. The number of subclusters and the number of physicochemical properties assigned to the pseudocenters under consideration of the local regions in each ESOM cluster.

Cluster No.	♯ Subclusters	♯ DO	♯ AK	♯ DA	♯ PI	♯ AL
1	124	69	59	25	41	65
2	112	96	98	18	62	24
3	80	54	31	8	18	23
4	161	97	115	46	91	47
5	48	15	31	11	30	6
6	64	23	27	9	20	19
7	124	60	74	15	48	18
8	45	15	16	4	8	25
9	60	18	25	5	41	18
10	120	102	72	16	52	11
11	129	87	74	18	78	24
12	128	78	83	29	78	39
13	137	96	81	26	108	30
14	70	35	38	15	77	5
15	59	31	34	7	21	7
16	97	57	101	16	80	7

Comparison of ESOM and EC classification

The obtained clustering by ESOM has to be compared to the EC classification of the enzymes considered in our data set. The EC classification has been taken from the ENZYME database [25]. In this database enzymes have been functionally annotated according to the chemical reaction being catalyzed, the used substrates and formed products and the possibly involved cofactors. In several subclusters local regions from proteins of mainly one single EC class are found. A subcluster is called characteristic for EC class c, if more than 80% of the cavities from EC class c are observed in this subcluster ($c \in \{1a, 1b, 1c, 2a, 2b, 3a, 3b, 4a, 4b, 5a, 5b, 6a, 6b\}$).

The clusters containing characteristic subclusters are listed in Table 3. EC class 5a is not listed here, because the data set contains only one cavity corresponding to EC class No. 5a. Table 3 shows that the characteristic subclusters for most of the EC classes are only found in some P-Matrix clusters.

Characteristic subclusters are significant, if the corresponding EC class is not over-represented in the cluster. A binomial model has been used to identify significant differences from the expected vs. the given number of local regions from EC class c within one cluster, $c \in \{1a, 1b, 1c, 2a, 2b, 3a, 3b, 4a, 4b, 5a, 5b, 6a, 6b\}$. The error value α was set to 0.05.

Those P-Matrix clusters which contain a significant number of local regions from the different EC classes are shown in Table 4. Column 'under' shows those clusters with a significantly smaller number of local regions than expected by the binomial model. The cluster numbers with significantly more local regions than expected by the binomial model are listed in column 'over'.

Significant subclusters can be identified using Tables 3 and 4. They are listed in Table 3, but not

Table 3. The cluster numbers containing characteristic subclusters for EC class c.

EC class	Cluster No.
1a = 1. 1. 1.21	–
1b = 1. 1. 1.42	1, 7, 13
1c = 1.14.13. 2	7, 9, 10, 12, 13
2a = 2. 7. 1.37	–
2b = 2. 7. 4. 9	–
3a = 3. 4.21.62	4
3b = 3. 4.23.20	4, 12
4a = 4. 2. 1. 1	4, 12
4b = 4. 4. 1.11	3, 4, 5, 11, 12, 13
5b = 5. 3. 1. 5	1, 2, 4, 6, 7, 10, 11, 13, 16
6a = 6. 3. 2. 3	1, 2, 4, 12
6b = 6. 3. 2. 9	2, 7, 9, 12

Table 4. The cluster numbers containing significant numbers of local regions.

EC Subclass No.	under	over
1a = 1. 1. 1.21	2, 9, 16	–
1b = 1. 1. 1.42	–	–
1c = 1.14.13. 2	–	11
2a = 2. 7. 1.37	16	–
2b = 2. 7. 4. 9	9, 16	15
3a = 3. 4.21.62	–	16
3b = 3. 4.23.20	14, 16	–
4a = 4. 2. 1. 1	–	4, 12
4b = 4. 4. 1.11	–	13
5a = 5. 4. 2. 1	–	12
5b = 5. 3. 1. 5	15	16
6a = 6. 3. 2. 3	–	–
6b = 6. 3. 2. 9	–	9

in the 'over' column of Table 4. For example, all characteristic subclusters for EC class 1c are significant, because they are not belonging to cluster No. 11, which falls into the 'over' column in Table 4. In contrast, all characteristic subclusters for EC class 4a are not significant, because the clusters they are belonging to fall into the 'over' column of Table 4.

Most of the characteristic subclusters are significant, e.g. all those for EC class 1b. However, the following subclusters are not significant: all entries of EC class 4a, those of EC class 4b in cluster No. 13, of EC class 5b in cluster No. 16 and of EC class 6b in cluster No. 9.

Many of the significant subclusters appear only in some prominent clusters. In these cases, the local regions defining the corresponding significant subclusters are good candidates for an appropriate classification of the proteins, since they reveal a very similar denomination as the EC classification scheme which is based on a totally different concept. Most important, the latter scheme requires additional information beyond structural data and it can hardly operate in a fully automized way as our method does.

Scoring of the matches found within the characteristic subclusters

The matches found within the characteristic subclusters are scored as described above. The scoring is done for every pair of local regions within one characteristic subcluster.

As expected, the matches where the corresponding cavities belong to the same EC class have very high score. The superimposition of two cavities based on their shared local regions is very convincing. As an example for such a good match, Figure 6 shows the superimposition for the cavities 1apv.1 and 1apw.1 belonging to EC class 3b (Pepsin-like Aspartate Protease). The local regions found in the corresponding subcluster, which were the basis for the superimposition of the two cavities, are colored in white. The absolute score for this match is 60, that means 60 pairs of suitable pseudocenters, the relative score is 0.97 (97%).

There are also local regions in characteristic subclusters from proteins belonging to different EC classes. They are matches for proteins exhibit-

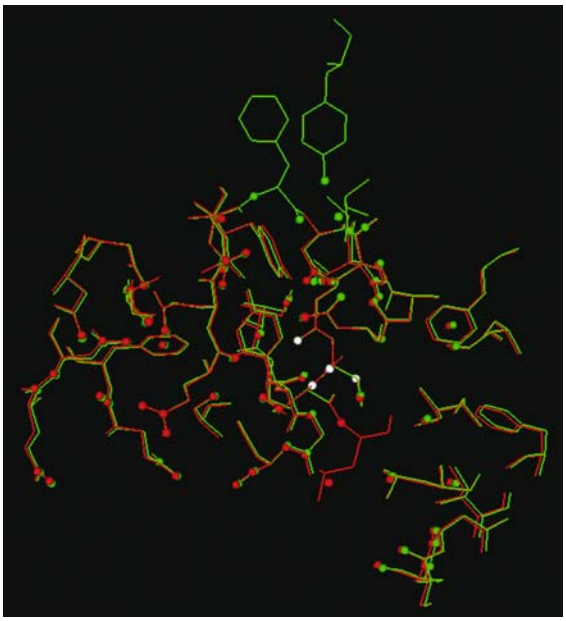

Figure 6. The figure shows the result of the superimposition of the cavities 1apv.1 and 1apw.1 with the same EC number. The pseudocenters of the local regions found within the corresponding subcluster which were the basis for this superimposition are colored in white. The amino acids and pseudocenters of the cavity of 1apv.1 are colored in red, those of the cavity of 1apw.1 are colored in green. The two cavities have 60 pseudocenters in common, so this superimposition obtains an absolute score of 60 and will not be neglected in the scoring step.

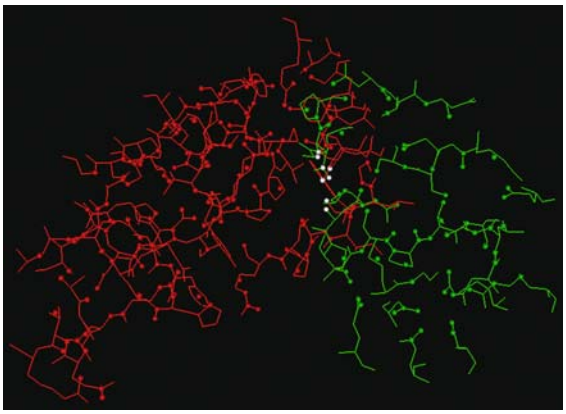

Figure 7. The figure shows the result of the superimposition of the cavities 1bf3.2 and 1gsa.2 having different EC numbers. The pseudocenters of the local regions found within the corresponding subcluster which were the basis for this superimposition are colored in white. The amino acids and pseudocenters of the cavity of 1bf3.2 are colored in red, those of the cavity of 1gsa.2 are colored in green. The two cavities have only the four pseudocenters of the local regions in common. So this superimposition obtains an absolute score of four and will be neglected in the scoring step.

ing different function. In general such subclusters could be indicative that two proteins possess related binding pockets that might accommodate similar ligands and give rise to unexpected cross-reactivity and side effects in drug action.

The matches for cavities belonging to enzymes from different EC classes have a much lower score. Only a few pseudocenters of the cavities show suitable correspondence. Frequently, they represent only the four pseudocenters belonging to the matching local region. As an example for such a restricted match, Figure 7 shows the superimposition for the cavities 1bf3.2 (EC class 1c: p-Hydroxybenzoate Hydroxylase) and 1gsa.2 (EC class 6a: Glutathione Synthase). The two cavities have only the four pseudocenters of the local regions in common, the absolute score is 4. The relative score for these two cavities is 0.03 (3%).

With the constraint to have an absolute score higher than nine, only those matches are remaining where the corresponding cavities belong to the same EC class. All those of cavities from different EC classes have a score smaller than nine and are eliminated. If there are two cavities from the same EC class within a significant subcluster for another EC class, the match for these two gets of course a high score and is not eliminated. So with the significant subclusters and the scoring, the local regions are clustered with respect to the EC class of the corresponding cavities.

Summary

This paper presents a new algorithm to compare substructural epitopes in protein binding cavities. The cavities are partitioned systematically into small local regions of pseudocenters. Each local region is composed by one center under consideration and its three nearest neighbors. Regarding the mutual distances between the four pseudocenters, the local regions exhibit the shape of a pyramid with a triangular basis. They are described by three spatial features and the four physicochemical properties assigned to the four contributing pseudocenters. The local regions are clustered with ESOM to detect groups of similar local regions. Two local regions originating from different cavities but found within the same cluster suggest a match for the corresponding cavities. They are used to superimpose the two cavities for the subsequent scoring of the match.

The data set used for proof-of-concept consists of local regions extracted from 72 different cavities. It contains enzymes originating from all 6 EC classes. Similarities and dissimilarities in this data set are well known. The spatial and physicochemical properties were analyzed separately. An emergent self-organizing map (ESOM) with 100×164 neurons has been trained with the spatial properties of the local regions. The U- and the P-Matrix of the ESOM show that the data groups into 16 clusters of local regions based on their spatial properties. The decision rules generated with the sig*-algorithm are used to check whether the clustering is valid. Other methods such as SVM may be better for classification. But the classification cannot be verified. The sig*-algorithm generates interpretable rules. With these rules the validity of the clusters can be proved easily.

The P-Matrix clusters were subdivided into subclusters with respect to the physicochemical properties of the local regions. Some of these subclusters contain local regions from over 80% of the cavities corresponding to enzymes of one EC class. They are called characteristic for this EC class. This result is, apart from one case, not dependent on the distribution of the different EC classes in the clusters. Those clusters, which contain statistically significantly more local regions from one EC class than expected, are different from the clusters containing characteristic subclusters for the corresponding EC class. These characteristic subclusters are called significant. They are likely candidates for a classification of the proteins with respect to their function.

Local regions are rather small subunits. So the probability to find similar local regions in two different cavities is relatively high. Considering a data set of enzymes well-spread over examples from different EC classes makes it rather likely to find members of the various EC classes in every subcluster of the clusters. For the selected data set this is actually found for the classes 1a, 2a and 2b. Interestingly enough, many significant subclusters are found.

All matches found in the significant subclusters have been analyzed with respect to the actual scoring of the superimposition of the correspond-

ing cavities. Those matches within one significant subcluster that agree to a very low score can be traced back to the small size of the local regions. Only matches with a score of more than nine matching pseudocenters have been regarded. Those matches are remaining where the corresponding cavities belong to the same EC class. So with the ESOM clustering and the scoring of the matches within the significant subclusters afterwards the local regions are to some degree clustered with respect to the EC class of the corresponding cavities.

The algorithm suggests a number of advantages compared to other techniques used for the comparison of protein binding cavities.

(1) With the ESOM clustering all local regions of cavities of an entire database can be compared in one step.
(2) The local region is a good starting point for the surface superposition required for the scoring of the resulting match. Based on four pseudocenters, which have to be matched, the coordinate transformation is determined more precisely than regarding only pairs of identical pseudocenters.
(3) The time consuming surface superpositioning has to be done only for those cavities which share a minimum of at least four pseudocenters, means one local region, in common. This implies a significant speed up, since not every cavity in the database must be subjected to the computationally demanding pairwise comparison.

Acknowledgements

Our special thanks are directed to the Deutsche Forschungsgemeinschaft (DFG) for supporting the present project in the framework of the NeuroCav project (UL 159/3-1).

References

1. Grindley, H.M., Artymiuk, P.J., Rice, D.W. and Willett, P., J. Mol. Biol., 229 (1993) 707.
2. Spriggs, R.V., Artymiuk, P.J. and Willett, P., J. Chem. Inf. Comput. Sci., 43 (2003) 412.
3. Wallace, A.C., Laskowski, R.A. and Thornton, J.M., Protein Sci., 5 (1996) 1001.
4. Wallace, A.C., Borkakoti, N. and Thornton, J.M., Protein Sci., 6 (1997) 2308.
5. Russell, R.B., J. Mol. Biol., 279 (1998) 1211.
6. Stark, A., Sunyaev, S. and Russell, R.B., J. Mol. Biol., 326 (2003) 1307.
7. Stark, A. and Russell, R.B., Nucleic Acids Res., 31 (2003) 3341.
8. Wangikar, P.P., Tendulkar, A.V., Ramya, S., Mali, D.N. and Sarawagi, S., J. Mol. Biol., 326 (2003) 955.
9. Hamelryck, T., Proteins, 51 (2003) 96.
10. Kleywegt, G.J., J. Mol. Biol., 285 (1999) 1887.
11. Bachar, O., Fischer, D., Nussinov, R. and Wolfson, H., Protein Eng., 6 (1993) 279.
12. Rosen, M., Lin, S.L., Wolfson, H. and Nussinov, R., Protein Eng., 11 (1998) 263.
13. Kinoshita, K. and Nakamura, H., Protein Sci., 12 (2003) 1589.
14. Lehtonen, J.V., Denessiouk, K., May, A.C. and Johnson, M.S., Proteins, 34 (1999) 341.
15. Poirrette, A.R., Artymiuk, P.J., Rice, D.W. and Willett, P., J. Comput.-Aided Mol. Des., 11 (1997) 557.
16. Yan, A. and Gasteiger, J., J. Chem. Inf. Comput. Sci., 43 (2003) 429.
17. Stahl, M., Taroni, C. and Schneider, G., Protein Eng., 13(2) (2000) 83.
18. Schmitt, S., Kuhn, D. and Klebe, G., J. Mol. Biol., 323 (2002) 387.
19. Kohonen, T., Biol. Cybernetics, 43 (1982) 59.
20. Ultsch, A., Proc. Workshop on Self-organizing Maps, Kyushu, Japan, 2003, pp. 225–230.
21. Ultsch, A., Proc. Conf. Soc. for Information and Classification, 1992.
22. Ultsch, A., U*-Matrix: a tool to visualize clusters in high dimensional data. Technical Report No. 36, Department of Mathematics and Computer Science, University of Marburg, 2004.
23. Hendlich, M., Rippmann, F. and Barnickel, G., J. Mol. Graph. Model, 15 (1997) 359.
24. Hendlich, M., Bergner, A., Gunther, J. and Klebe, G., J. Mol. Biol., 326 (2003) 607.
25. Bairoch, A., Nucleic Acids Res., 28(1) (2000) 304.
26. Siemon, R., Einige Werkzeuge zum Einsatz von selbstorganisierenden Neuronalen Netzen zur Strukturanalyse von Wirkstoff-Rezeptoren. Diploma Thesis, Department of Mathematics and Computer Science, University of Marburg, 2001.

Appendix 1: PDB codes of the proteins in the data set

Table A1. The EC numbers, the names of the proteins within this EC class and the corresponding PDB codes of the proteins in our data set.

EC Number	Name of the proteins	PDB codes of the proteins
main number 1		
1a = (1, 1, 1, 21)	Aldose Reductase	1ah0.1, 1ah3.1, 1ah4.1, 1az1.1, 1az2.1, 1ef3.1, 1eko.1, 2acr.1, 2acs.1, 2acu.1
1b = (1, 1, 1, 42)	Isocitrate Dehydrogenase	1ai2.3, 1ai3.1, 1bl5.3, 1cw1.4, 1gro.4
1c = (1, 14, 13, 2)	p-Hydroxybenzoate Hydroxylase	1bf3.2, 1cj2.2, 1cj3.2, 1d7l.1, 1ius.2, 1pxa.3, 1pxb.2
main number 2		
2a = (2, 7, 1, 37)	Kinases Ser,Thr,Tyr	1bkx.2, 1atp.2, 1cdk.7, 1ydr.3, 1hck.1, 1b38.1, 1gol.1, 1phk.1, 1csn.1, 1ir3.1, 2src.1, 1fin.5, 1fin.7
2b = (2, 7, 4, 9)	Thymidylate Kinase	1e9a.1, 1e9d.1, 1e9e.1, 1tmk.4, 2tmk.4, 3tmk.18
main number 3		
3a = (3, 4, 21, 62)	Subtilase	1bfu.1, 1bh6.1, 1sua.2, 1sue.2
3b = (3, 4, 23, 20)	Pepsin-like Asp Protease	1apt.1, 1apu.1, 1apv.1, 1apw.1, 1bxo.1, 1bxq.1
main number 4		
4a = (4, 2, 1, 1)	Carbonic Anhydrase	1cil.1, 1g52.1, 1g54.1, 1i90.1, 1azm.1, 1bzm.1
4b = (4, 4, 1, 11)	Methionine γ Lyase	1e5e.4, 1e5f.3
main number 5		
5a = (5, 4, 2, 1)	Phosphoglycerate Mutase	1e59.3
5b = (5, 3, 1, 5)	D-Xylose Isomerase	1xii.1, 1xli.2, 1xyc.2, 1xym.2, 5xim.7, 5xin.7
main number 6		
6a = (6, 3, 2, 3)	Glutathione Synthase	1gsa.2, 2hgs.2
6b = (6, 3, 2, 9)	Udp-N-Acetylmuramoyl-L-Alanine: D-Glutamate Ligase	1eeh.1, 1uag.1, 2uag.1, 3uag.1