IN *PROC. ICBA 2004*[*]; *FORT LAUDERDALE, FLORIDA, USA*:

AN ALGORITHM FOR FINDING SIMILARITIES IN PROTEIN ACTIVE SITES

K. KUPAS[†] AND A. ULTSCH

Data Bionics Research Group,
University of Marburg, D-35032 Marburg, Germany,
E-mail: kupas@informatik.uni-marburg.de,
ultsch@informatik.uni-marburg.de

G. KLEBE

Institute of Pharmaceutical Chemistry,
University of Marburg, D-35032 Marburg, Germany,
E-mail: klebe@mailer.uni-marburg.de

A new method has been developed to find similar substructures in protein binding cavities. It is based on the idea that protein function is intimately related to the recognition and subsequent response to the binding of an endogenous ligand in a well-characterized binding pocket. It can therefore be assumed that proteins having similar binding cavities also bind similar ligands and exhibit related function. For the comparison of the binding cavities, the binding-site exposed physicochemical characteristics are described by assigning generic pseudocenters to the functional groups of the amino acids flanking a particular active site. These pseudocenters are assembled into small substructures. To find substructures with spatial similarity and appropriate chemical properties, an emergent self-organizing map is used for clustering. Two substructures which are found to be similar form the basis for an expanded comparison of the complete cavities. Preliminary results with four pairs of binding cavities show that similarities are detected correctly and motivate further studies.

Keywords: Functional Comparison of Proteins, *De Novo* Design, Data Mining, SOM, U-Matrix

[†]Work supported by the Deutsche Forschungsgemeinschaft (DFG).

1. Introduction

The molecular function of a protein is coupled to the binding of a substrate or an endogenous ligand to a well defined binding cavity. This requires highly conserved molecular recognition patterns from the receptor. 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.

The three-dimensional structure has to be regarded as a prerequisite for a reliable comparison of proteins. Such structures are available for many examples from X-ray crystallography. In literature, different methods based on the description of the spatial protein structures in terms of a reduced set of appropriate descriptors have been reported. 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 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 [1] for common substructure detection. They operate completely independent of any sequence or fold homology. The approach of Rosen *et al.* [2] permits an automatic comparison of binding cavities. Kinoshita *et al.* [3] use a graph-based algorithm to compare the surfaces of two proteins. Other methods, such as GENE FIT of Lehtonen *et al.* [4] and the approach of Poirrette *et al.* [5] use genetic algorithms to optimally superimpose proteins in identified substructure ranges. All these approaches only use descriptors for the shape of the protein. But in addition to the shape, it is required to code correctly the exposed physicochemical properties in a geometrical and also chemical sense.

In this paper, we describe a new algorithm for the comparison of substructures in protein binding cavities. We use well-placed pseudocenters representative for a small set of physicochemical properties and the mutual distance between two such centers [6]. Every cavity then is described with also geometrical and physico-chemical properties. The pseudocenters are assembled into small substructures. We used a cluster algorithm based on emergent self-organizing maps (ESOM) for finding similar substructures. By using ESOM for clustering the local pseudocenter assemblies, an all-against-all comparison also for very large data samples is possible. Two

substructures found to be similar provide a coordinate system which will be used in the next step to superimpose the related cavities and score the detected match.

2. Theory and Algorithm

2.1. *Local regions in binding cavities*

The protein binding cavities are characterized by the descriptors developed by Schmitt *et al.* [6]. 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). 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 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. Every cavity comprises on the mean about 100 pseudocenters and accordingly 100 local regions.

The number of four pseudocenters for the local region has been chosen because this is the smallest possible pseudocenter assembly having spatial properties. The mutual distances between the four pseudocenters of a local region form a pyramid with a triangular basis. The local region is described sufficiently by three spatial properties, the area of the basis triangle, the height and the skewness of the pyramid. The physicochemical properties of the local regions correspond to the physicochemical properties assigned to the four pseudocenters.

2.2. *Emergent self-organizing maps*

To visualize high-dimensional data, a projection from the high dimensional space onto two dimensions is needed. 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 [7]. In ESOM a large number of neurons is used. The weight vectors of the neurons are thought 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 [8]. 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 of the 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 [8]. All figures in the following are tiled displays.

The U-Matrix has become the canonical tool for the display of the distance structures of the input data on ESOM [7]. 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, clusters in the data set are found in the valleys, mountain ranges point to cluster boundaries.

2.3. *Clustering of the local regions*

To find similar local regions, a self-organizing map is trained with the spatial features of the local regions. 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 are also from different regions in data space. The U-Matrix for this map is calculated. All data points found in coherent regions on the U-Matrix with small U-heights are assigned to one cluster. Larger U-heights are the cluster boundaries. Those data points lying in regions with large U-heights are distant to each other and not assigned to any cluster. All local regions lying in the same cluster have the same spatial properties.

2.4. *Scoring of the match*

Every pair of local regions of two cavities within the same cluster and with suitable physicochemical properties forms the basis for the further comparison of the two cavities. The two local regions give rise to a coordinate transformation, which optimally superimposes both pyramids. Subsequently this coordinate transformation is applied to the whole cavities. Then, every pair of pseudocenters of the two cavities, which mutually match chemically and fall close to each other beyond a threshold of 1 Å is counted. This number of successful matches is the absolute score for this pair of pyramids. The superpositioning is done for all pairs of pyramids with the same physico-

chemical and sterical properties of these two cavities. The maximum of the resulting absolute scores is determined. Relating this maximum to the "maximally achievable score", i.e. the number of pseudocenters in the smaller of the two cavities, gives then the final score of the two cavities.

3. First Results

The algorithm has been tested with four pairs of binding cavities with well known common substructures (1ake/1ukz, 1csm/1ecm, 1hlu/1kay and 1tpo/2prk (Protein Data Base code (PDB))). Other similarities between the binding cavities belonging to different pairs of proteins were not expected.

Those eight cavities have been divided into local region as described in Chapter 2.1. Because Euclidean distances are used in the learning phase of the ESOM, the distributions of the spatial features must be straight. Therefore a preprocessing step has to be done.

With this preprocessed data an ESOM with 50×82 neurons has been trained. The corresponding U-Matrix of this map can be seen in Figure 1. For reasons of clarity the data points are not shown in this figure. The

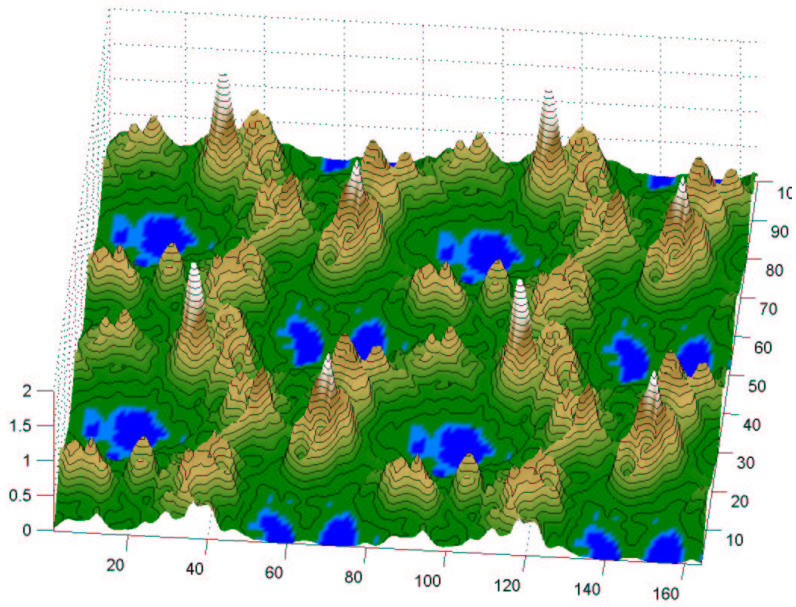

Figure 1. U-Matrix of the spatial features of the local regions corresponding to four pairs of binding cavities in 3D view.

U-Matrix shows clusters of local regions. All data points lying in a valley on the U-Matrix have similar spatial features and belong to one cluster.

Two local regions from different cavities within the same ESOM-cluster and suitable physicochemical properties are a match for the corresponding cavities. All matches have been scored as described in Chapter 2.4. The resulting final scores are shown in Table 1.

Table 1. Resulting scores of a mutual comparison of four pairs of binding cavities with well-known common substructures.

	1ake	1csm	1ecm	1hlu	1kay	1tpo	1ukz	2prk
1ake	—	21.1	20.7	11.5	13.5	19.4	**69.1**	26.2
1csm	21.1	—	**41.4**	14.0	21.1	12.3	17.5	9.5
1ecm	20.7	**41.4**	—	13.8	17.2	0.0	20.7	17.2
1hlu	11.5	14.0	13.8	—	**29.7**	14.9	12.7	11.9
1kay	13.5	21.0	17.2	**29.7**	—	17.9	13.6	21.4
1tpo	19.4	12.3	0.0	14.9	17.9	—	14.9	**28.6**
1ukz	**69.1**	17.5	20.7	12.7	13.6	14.9	—	19.1
2prk	26.2	9.5	17.2	11.9	21.4	**28.6**	19.1	—

Table 1 shows that those cavities which are known to possess common substructures achieve the best scores, whereas the best fit found for the other cavities reveals in most of the cases a significantly smaller value. The results have been examined by an expert. The coordinate transformations and the matching pseudocenters of the known pairs of binding cavities were identical with the estimated analogy.

4. Conclusions

We presented a new algorithm to find common substructures in protein binding cavities. The cavities are partitioned into small local regions of four pseudocenters with spatial and physicochemical properties. Regarding the mutual distances between the four pseudocenters, the local regions exhibit the shape of a pyramid with a triangular basis. The physicochemical characteristics of the local regions are the combination of the physicochemical attributes assigned to each pseudocenter. The spatial descriptors are the area of the basis triangle, the height and the skewness of the pyramid. 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 scoring of the match.

The comparison of four pairs of binding cavities with well-known common substructures led to promising results. The matches for cavities be-

longing to the same pair achieved higher scores than those for cavities from different pairs of proteins. It was possible to detect the similarities between the eight binding cavities correctly.

The algorithm has a set of advantages compared to other applications 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 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 four pseudocenters, that means one local region, in common. This implies a significant speed up, since not every cavity in the database is subjected to the pairwise comparison.

References

1. Bachar, O., Fischer, D., Nussinov, R. & Wolfson, H. (1993). A computer vision based technique for 3-D sequence-independent structural comparison of proteins. *Protein Eng* **6**, 279-88.
2. Rosen, M., Lin, S. L., Wolfson, H. & Nussinov, R. (1998). Molecular shape comparisons in searches for active sites and functional similarity. *Protein Engineering* **11**, 263-77.
3. Kinoshita, K. & Nakamura, H. (2003). Identification of protein biochemical functions by similarity search using the molecular surface database eF-site. *Protein Sci* **12**, 1589-95.
4. Lehtonen, J. V., Denessiouk, K., May, A. C. & Johnson, M. S. (1999). Finding local structural similarities among families of unrelated protein structures: a generic nonlinear alignment algorithm. *Proteins* **34**, 341-55.
5. Poirrette, A. R., Artymiuk, P. J., Rice, D. W. & Willett, P. (1997). Comparison of protein surfaces using a genetic algorithm. *J Comput Aided Mol Des* **11**, 557-69.
6. Schmitt, S., Kuhn, D. & Klebe, G. (2002). A new method to detect related function among proteins independent of sequence and fold homology. *J. Mol. Biol.* **323**, 387- 406.
7. Kohonen, T. (1982). Self-Organized formation of topologically correct feature maps. *Biological Cybernetics* **43**, 59-69.
8. Ultsch, A. (2003). Maps for the Visualization of high-dimensional Data Spaces. *Proc. Workshop on Self organizing Maps, Kyushu, Japan* 225 - 230 .
9. Ultsch, A. (1992). Self-organizing neural networks for visualization and classification. *Proc. Conf. Soc. for Information and Classification*