

In *Proc. GfKI 2004\**, Dortmund:

# Data Mining in Protein Binding Cavities

Katrin Kupas and Alfred Ultsch

Data Bionics Research Group,  
University of Marburg, D-35032 Marburg, Germany

**Abstract.** The molecular function of a protein is coupled to the binding of a substrate or an endogenous ligand to a well defined binding cavity. To detect functional relationships among proteins, their binding-site exposed physicochemical characteristics were described by assigning generic pseudocenters to the functional groups of the amino acids flanking a particular active site. These pseudocenters were assembled into small substructures and their spatial similarity with appropriate chemical properties was examined. If two substructures of two binding cavities are found to be similar, they form the basis for an expanded comparison of the complete cavities. Preliminary tests indicate the benefit of this method and motivate further studies.

## 1 Introduction

In a biological system multiple biochemical pathways are proceeded and regulated via the complementary recognition properties of proteins and their substrates. The ligand accommodates the binding cavity of the protein according to the lock-and-key principle. Two fold requirements are given: on the one hand, the ligand needs to fit sterically into the binding cavity of the protein. On the other hand, the spatial arrangement of ligand and receptor must correspond to a complementary physicochemical pattern.

The shape and function of a protein, e.g. of an enzyme together with its active site is not exclusively represented by a unique amino acid sequence. Accordingly, proteins with deviating amino acid sequence, even adopting a different folding pattern, can nevertheless exhibit related binding cavities to accommodate a ligand. 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. 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. In addition to the shape, it is required to

---

\* Copyright by Springer-Verlag, all rights are reserved.

code correctly the exposed physicochemical properties in a geometrical and also chemical sense.

In this paper, we describe a new algorithm to compare protein binding sites by the use of common local regions. These local regions form the basis for the further comparison of two binding cavities. Similar local regions among sets of spatially arranged descriptors of two binding cavities provide a coordinate system which will be used in the next step to perform other substructure searches for related cavities. Once a convincing match is detected, it can be assumed that the two active centers are capable to bind similar ligands and thus exhibit related function.

The paper is organized as follows: In Chapter 2, other approaches to classify binding cavities are reviewed. Chapter 3 describes the underlying theory and concept of our algorithm used for cavity matching. The local region in descriptor space is defined. In Chapter 4, some preliminary results of a binding cavity matching are presented. Conclusions are given in Chapter 5.

## 2 Other approaches

Previously reported approaches to classify binding cavities can be assigned to three categories, according to the information they use for the classification:

1. Sequence alignments.
2. Comparison of folding patterns and secondary structure elements.
3. Comparison of 3D substructural epitopes.

### Sequence alignments

If two proteins show high sequence identity one can assume structural and most likely also functional similarity among them. The mostly applied procedures were presented by Needleman and Wunsch (1970) and by Waterman (1984). Nevertheless, they are computationally and memory-wise quite demanding, so that often heuristic methods are used such as FASTA (Pearson and Lipman (1988), Pearson (1990)) and BLAST (Altschul et al. (1990)). These procedures do not find an optimal solution in all cases, but generally reveal good approximative results.

### Comparison of folding patterns and secondary structure elements

In general, sequence alignment methods are only capable to detect relationships among proteins if sequence identity exceeds beyond 35%. To classify more distant proteins, information about their three-dimensional structure has to be incorporated to the comparison. Many methods, that establish classification and assignment of proteins to structural families, exploit global fold similarities. Hierarchical procedures have been developed which classify proteins according to their folding, their evolutionary ancestors, or according to their functional role. These systems operate either automatically or are

dependent on manual intervention. Many of the classification schemes treat proteins as being composed by domains and classify them in terms of the properties of their individual domains (Ponting and Russell (2002)). Important approaches for such classifications are: SCOP (Murzin et al. (1995), Lo Conte et al. (2002)), CATH (Orengo et al. (1997, 2000)), FSSP/DALI (Holm and Sander (1996), Holm (1998)), MMDB (Gibrat et al. (1996)). The ENZYME (Bairoch (2000)) and BRENDA database (Schomburg et al. (2002)) annotate proteins with respect to the catalyzed reaction.

### Comparison of 3D substructural epitopes

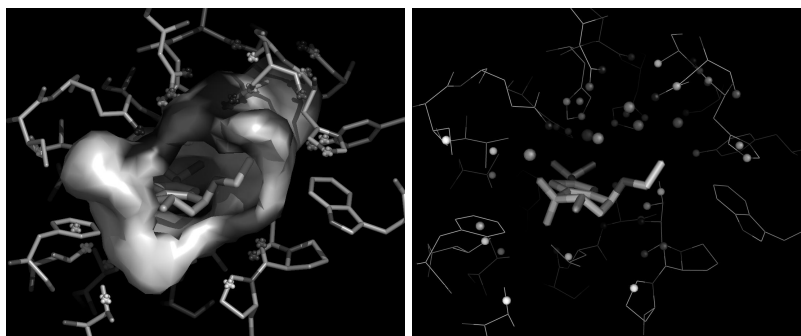
Beyond these relationships, proteins can possess a similar function even if they do not have any sequence and/or folding homology in common. Accordingly, methods that compare proteins only with respect to their folding pattern cannot detect such similarities. Procedures which seek for similar substructures in proteins are better adapted to discover similarities in such cases.

The first group of algorithms comprises methods that scan protein structural databases in terms of pre-calculated or automatically generated templates. A typical example of such a template is the catalytic triad in serine proteases. A substantial advantage to restrict to relatively small templates is due to the fact that even large data collections can be scanned efficiently. Some of the best known procedures based on templates are ASSAM introduced by Artymiuk et al. (1993,2003), TESS/PROCAT by Wallace et al. (1996, 1997), PINTS by Stark and Russell (2003), DRESPAT by Wangikar et al. (2003) as well as the methods of Hamelryck (2003) and Kleywegt (1999).

The second group includes approaches to compare substructural epitopes of proteins which operate independent of any template definition. For the similarity search the whole proteins or substructures are used. The group of Ruth Nussinov and Haim Wolfson developed many approaches to compare entire receptor structures or substructures. The individual methods essentially differ whether the protein structure is represented by their  $C_{\alpha}$ -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 (Bachar et al. (1993)) for common substructure detection. They perform completely independent of sequence or fold homology. The approach of Rosen (1998) permits an automatic comparison of binding cavities. Kinoshita et al. (2003) use a graph-based algorithm to compare the surfaces of two proteins. Other methods, such as GENE FIT of Lehtonen et al. (1999) and the approach of Poirrette et al. (1997) use genetic algorithms to optimally superimpose proteins in identified substructure ranges.

### 3 Theory and Algorithm

The algorithm builds on the approach of Schmitt et al. (2002). 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), hydrogen-bond 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.



**Fig. 1.** Surface (left) and pseudocenters (right) of a binding cavity with bound ligand

The idea for this algorithm resides on the concept that common substructures of two binding cavities have an arrangement of these pseudocenters in common. Therefore local regions are regarded. They are composed by a center under consideration and the three nearest neighboring centers forming a pyramid. The pyramid was chosen as similarity measure because it corresponds to the smallest spatial unit spanned by these four centers.

Systematically every pseudocenter in a cavity is selected as the current center and forms a local region with its three nearest neighbors. Following the procedure, the binding cavities are partitioned by all local regions to be possibly inscribed. Accordingly, the mutual comparison of binding cavities is reduced to a multiple comparison of the different local regions.

Therefore, the spatial and physicochemical characteristics of the local regions are considered separately. The spatial features of the local regions were chosen under the aspects of using a minimum number of descriptors to identify local regions and producing minimal measuring errors. A set of six spatial descriptors was tested. Three of them, the height of the pyramid, the area of the triangle spanned by the three neighbor centers and the distance between the root point of the height and the barycenter of this triangle, have been chosen.

Every pair of local regions of two cavities with the same physicochemical and spatial properties forms the basis for the comparison of the two cavities. These two local regions are matched and the score of the appropriate overlay of the cavities is calculated.

1. Two pyramids with appropriate chemical and spatial characteristics of different binding cavities give rise to a coordinate transformation, which optimally superimposes both pyramids. That means if pyramid A consists of the points  $(A_1, A_2, A_3, A_4)$  and pyramid B of the points  $(B_1, B_2, B_3, B_4)$ , then a rotation/translation has to be found, so that the sum of the squares of the distances from  $A_i$  to  $B_i$  adopts a minimal value (Prokrustes analysis).
2. 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 score for this pair of pyramids.
3. The superpositioning is done for all pairs of pyramids with the same physicochemical and sterical properties of these two cavities. The maximum of the resulting scores is determined.
4. Relating this maximum score to the "maximally achievable score", i.e. the number of pseudocenters in the smaller of the two cavities, gives an estimate of the maximally achievable score.

This algorithm has a set of advantages contrary to a consideration only of the individual pseudocenters. With a Prokrustes analysis concerning the individual pseudocenters the coordinate transformation must be accomplished for all pairs of chemically identical pseudocenters and the best match has to be calculated. By consideration of local regions four suitable pseudocenters are given, which have to be matched. Thus the number of computations for the superpositioning of the two cavities is reduced. Only those overlaps with a match of all four pseudocenters have to be computed. A local region composed of four centers gives a good initialization for the Prokrustes analysis. Fewer degrees of freedom exist for the coordinate transformation. A further advantage is that not all binding cavities of a data base have to be considered. Only those cavities containing a suitable local region come into consideration for the surface overlay of the cavities. The remaining cavities without a suitable local region are not consulted.

## 4 First Results

The approach based on local regions for the comparison of protein active sites has been tested with four pairs of binding cavities with well known common substructures (Siemon (2001)). Other similarities between the proteins than these pairs were not expected. The proteins from where the binding cavities

had been extracted are the following:

an Adenylate Kinase (1ake.2), an allosteric Chorismate Mutase (1csm.3), the Chorismate Mutase of E. Coli (1ecm.5), a Bovine-Actin-Profilin Complex (1hlu.1), a heat shock cognate Protein (1kay.1), Trypsin (1tpo.1), the Uridylate Kinase (1ukz.1) and Proteinase K (2prk.2) (Protein Data Base code (PDB)).

The pairs of proteins with well-known common substructures in their binding cavities are 1ake.2/1ukz.1 (Kinases), 1csm.3/1ecm.5 (Isomerases), 1hlu.1/1kay.1 (Hydrolases) and 1tpo.1/2prk.1 (Serine Proteinases). The resulting scores after mutual match are shown in Figure 1.

	1ake.2	1csm.3	1ecm.5	1hlu.1	1kay.1	1tpo.1	1ukz.1	2prk.1
1ake.2	—	21.1	20.7	11.5	13.5	19.4	<b>69.1</b>	26.2
1csm.3	21.1	—	<b>41.4</b>	14.0	21.1	12.3	17.5	9.5
1ecm.5	20.7	<b>41.4</b>	—	13.8	17.2	0.0	20.7	17.2
1hlu.1	11.5	14.0	13.8	—	<b>29.7</b>	14.9	12.7	11.9
1kay.1	13.5	21.0	17.2	<b>29.7</b>	—	17.9	13.6	21.4
1tpo.1	19.4	12.3	0.0	14.9	17.9	—	14.9	<b>28.6</b>
1ukz.1	<b>69.1</b>	17.5	20.7	12.7	13.6	14.9	—	19.1
2prk.1	26.2	9.5	17.2	11.9	21.4	<b>28.6</b>	19.1	—

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

The numbers are given as percentage with respect to the maximally achievable score (see section 3.3).

The table shows that those cavities which are known to possess common substructures also achieve the best scores, whereas the best fit found for the other cavities reveals in most of the cases significantly smaller values. 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.

## 5 Conclusions

We presented a new algorithm to find common substructures and compare protein binding cavities. The cavities are partitioned into small local regions with spatial and physicochemical properties. They are formed by pseudocenters assigned to five different physicochemical qualities. The local region exists of a center under consideration and its three nearest neighbors. The physicochemical characteristics of the local regions are the combination of the physicochemical attributes assigned to each pseudocenter. The spatial characteristics were described by the height of the pyramid, the area of the

triangle spanned by the three neighbor centers and the distance between the root point of the height and the barycenter of this triangle.

The advantage of this algorithm is, that only those cavities are observed, that share a common local region, expressed in terms of the pyramid. Such substructures, which are represented by pseudocenters widely distributed over the cavity and so aren't biologically relevant, are a priori excluded from the consideration. An advantage of the use of an ESOM for classifying protein binding cavities is the fast comparison of one individual cavity with the entire database. All candidates of proteins of the whole database sharing in common similar local regions are identified in one step.

The comparison of four pairs of binding cavities with well-known common substructures led to promising results. It can be assumed that the approach of dividing protein binding cavities into local regions and comparing them is capable to detect similar substructures in the cavities.

## References

- ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W. and LIPMAN, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.*, 215, 403-10.
- ARTYMIUK, P. J., GRINDLEY, H. M., RICE, D. W. and WILLETT, P. (1993). Identification of tertiary structure resemblance in proteins using a maximal common subgraph isomorphism algorithm. *J. Mol. Biol.*, 229, 707-21.
- ARTYMIUK, P. J., SPRIGGS, R. V. and WILLETT, P. (2003). Searching for patterns of amino acids in 3D protein structures. *J. Chem. Inf. Comput. Sci.*, 43, 412-21.
- BACHAR, O., FISCHER, D., NUSSINOV, R. and WOLFSON, H. (1993). A computer vision based technique for 3-D sequence-independent structural comparison of proteins. *Protein Eng.*, 6, 279-88.
- BAIROCH, A. (2000). The ENZYME database in 2000. *Nucleic Acids Res.*, 28, 304-5.
- GIBRAT, J. F., MADEJ, T. and BRYANT, S. H. (1996). Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.*, 6, 377-85.
- HAMELRYCK, T. (2003). Efficient identification of side-chain patterns using a multidimensional index tree. *Proteins*, 51, 96-108.
- HOLM, L. and SANDER, C. (1996). The FSSP database: fold classification based on structure-structure alignment of proteins. *Nucleic Acids Res.*, 24, 206-9.
- HOLM, S. (1998). Touring protein fold space with Dali/FSSP.
- KINOSHITA, K. and NAKAMURA, H. (2003). Identification of protein biochemical functions by similarity search using the molecular surface database eF-site. *Protein Sci.*, 12, 1589-95.
- KLEYWEYGT, G. J. (1999). Recognition of spatial motifs in protein structures. *J. Mol. Biol.*, 285, 1887-97.
- LEHTONEN, J. V., DENEISSIOUK, K., MAY, A. C. and JOHNSON, M. S. (1999). Finding local structural similarities among families of unrelated protein structures: a generic nonlinear alignment algorithm. *Proteins*, 34, 341-55.
- LO CONTE, L., BRENNER, S. E., HUBBARD, T. J., CHOTHIA, C. and MURZIN, A. G. (2002). SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res.*, 30, 264-7.

- MURZIN, A. G., BRENNER, S. E., HUBHARD, T. and CHOTHIA, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, 247, 536-40.
- NEEDLEMAN, S.B. and WUNSCH, C.D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, 48, 443-453.
- ORENGO, C. A., MICHIE, A. D., JONES, S., JONES, D. T., SWINDELLS, M. B. and THORNTON, J. M., (1997). CATH—a hierarchic classification of protein domain structures. *Structure*, 5, 1093-108.
- ORENGO, C. A., PEARL, F. M., LEE, D., BRAY, J. E., SILLITO, I., TODD, A. E., HARRISON, A. P. and THORNTON, J. M. (2000). Assigning genomic sequences to CATH. *Nucleic Acids Res* 28, 277-82.
- PEARSON, W. R. and LIPMAN, D. J. (1988). Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A*, 85, 2444-8.
- PEARSON, W. R. (1990). Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol*, 183, 63-98 .
- POIRETTE, A. R., ARTYMIUK, P. J., RICE, D. W. and WILLETT, P. (1997). Comparison of protein surfaces using a genetic algorithm. *J. Comput. Aided Mol. Des.*, 11, 557-69.
- PONTING, C. P. and RUSSELL, R. R. (2002). The natural history of protein domains. *Annu. Rev. Biophys. Biomol. Struct.*, 31, 45-71.
- ROSEN, M., LIN, S. L., WOLFSON, H. and NUSSINOV, R. (1998). Molecular shape comparisons in searches for active sites and functional similarity. *Protein Eng.*, 11, 263-77.
- RUSSELL, R. B. (1998). Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. *J. Mol. Biol.*, 279, 1211-27.
- SCHMITT, S., KUHN, D. and KLEBE, G. (2002). A new method to detect related function among proteins independent of sequence and fold homology. *J. Mol. Biol.*, 323, 387- 406.
- SCHOMBURG, I., CHANG, A. and SCHOMBURG, D. (2002). BRENDA, enzyme data and metabolic information. *Nucleic Acids. Res.*, 30, 47-9.
- SIEMON, R. (2001). Einige Werkzeuge zum Einsatz von selbstorganisierenden Neuronalen Netzen zur Strukturanalyse von Wirkstoff-Rezeptoren. *Diplomarbeit, 13.2.2001, FB Mathematik u. Informatik*
- STARK, A., SUNYAEV, S. and RUSSELL, R. B. (2003). A model for statistical significance of local similarities in structure. *J. Mol. Biol.*, 326, 1307-16.
- STARK, A. and RUSSELL, R. B. (2003). Annotation in three dimensions. PINTS: Patterns in Non-homologous Tertiary Structures. *Nucleic Acids Res.*, 31, 3341-4.
- WALLACE, A. C., LASKOWSKI, R. A. and THORNTON, J. M. (1996). Derivation of 3D coordinate templates for searching structural databases: application to Ser-His-Asp catalytic triads in the serine proteinases and lipases. *Protein Sci.*, 5, 1001-13.
- WALLACE, A. C., BORKAKOTI, N. and THORNTON, J. M. (1997). TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. *Protein Sci.*, 6, 2308-23.
- WANGIKAR, P. P., TENDULKAR, A. V., RAMYA, S., MALI, D. N. and SARAWAGI, S. (2003). Functional sites in protein families uncovered via an objective and automated graph theoretic approach. *J. Mol. Biol.*, 326, 955-78.

WATERMAN, M.S. (1984). General methods for sequence comparison. *Bull. Math. Biol.*, 46, 473-500.