Pattern Recognition Strategies for Molecular Surfaces. I. Pattern Generation Using Fuzzy Set Theory

THOMAS E. EXNER,1,2 MATTHIAS KEIL,2 JÜRGEN BRICKMANN2,3

1Department of Chemistry, Mathematical Chemistry Research Unit, University of Saskatchewan, 110 Science Place, Saskatoon, SK, Canada, S7N 5C9
2Department of Physical Chemistry, Darmstadt University of Technology, 64287 Darmstadt, Germany
3Darmstadt Center of Scientific Computing, 64287 Darmstadt, Germany

Received 12 November 2001; Accepted 21 January 2002

Abstract: A new method for the characterization of molecules based on the model approach of molecular surfaces is presented. We use the topographical properties of the surface as well as the electrostatic potential, the local lipophilicity/hydrophilicity, and the hydrogen bond density on the surface for characterization. The definition and the calculation method for these properties are reviewed shortly. The surface is segmented into overlapping patches with similar molecular properties. These patches can be used to represent the characteristic local features of the molecule in a way that is beyond the atomistic resolution but can nevertheless be applied for the analysis of partial similarities of different molecules as well as for the identification of molecular complementarity in a very general sense. The patch representation can be used for different applications, which will be demonstrated in subsequent articles.


Key words: molecular surface; molecular recognition; surface segmentation; fuzzy set theory; linguistic variables

Introduction

Molecular recognition problems are relevant for a variety of chemical structures and processes. An example is the biochemical specificity, such as the binding of enzymes to substrates or inhibitors and antibodies to antigens. The molecular recognition may often be described in terms of the key and lock image first introduced by Fischer.1 In the attempt to describe such recognition in a quantitative way, a large variety of different factors (energetic, entropic, and kinetic) come into play in a conceptual model approach. In principle, the tools for such an approach are available. ∆G-values or relative ∆G-values for the binding of a substrate to a given receptor can be calculated based on thermodynamic quantities and molecular simulations. Such calculations are still very expensive as far as computational effort is concerned. Therefore, they are only justified for such molecular aggregates for which a preselection process of molecular partners and their relative arrangements has been made on the basis of a simplified model scenario: one has to answer the question of what molecular part of the ligand, the “key,” may fit to what molecular part of the receptor, the “lock.” The answer is strongly related to the question of molecular similarity (if different keys should be tested as possible partners for an unknown “lock,” knowing that one or more molecules are binding) or molecular complementarity (if the “lock” is known and new, improved “keys” should be found). The search becomes more complicated when the structures of both molecules are known but the active site cannot be specified a priori. In this case, one is looking for complementarity between arbitrary regions of one molecule and regions of a second molecule without any knowledge of the search patterns.

How can the search for possible topographical similarity or complementarity of molecular surfaces effectively be done? There is no doubt that the “eyeball” technique used by human “searchers” is a very effective search procedure—for those instances in which it can be applied.2–4 One can easily compare different objects and analyze their similarity or dissimilarity without having explicitly defined criteria in hand.

An alternative to the eyeball technique is an algorithmic identification of similarity and complementarity. Several articles have been published on the analysis of global shape similarity of mol-
ecules in the last few years. In many cases, an algorithmic solution of the docking problem is not an easy task, while for the same system the eyeball technique can still be applied very effectively. This becomes even more obvious if one is interested in the analysis of partial similarity. The remarkable fact in this case is that one does not have to define the criteria for the search in advance. It seems to be obvious that the brain determines these criteria "on the fly." Presently, there are no computer algorithms available that have the capability for this autonomous determination of search criteria. It is a challenge to transfer the human efficiency into mathematical objects treatable by computer algorithms. In this article, new search patterns are presented, capable for the determination of binding sites. The formalisms are based on fuzzy set theory and the concept of molecular surfaces.

The article is organized as follows: in the remaining part of the first section we shortly review publications on search pattern definitions, which are related to the one proposed here. In the second section the basic principles of fuzzy set theory and fuzzy logic are given. The methods for the calculation of the different molecular properties used for the pattern generation are reported in the third section. Then, the segmentation algorithm generating the search patterns is discussed in detail. The results obtained are presented and discussed in the Results and Discussion section. The last section provides some concluding remarks and a short preview of applications for active site identification discussed in detail in subsequent articles.

For almost all docking algorithms, not relying on energy minimization, one has to define very specific search criteria. It is out of the scope of this article to review all the criteria described in the literature. Only those methods are considered in the short review that are based on concepts using molecular surfaces similar to the one presented here.

The DOCK program of Kuntz and coworkers was one of the earliest docking methods frequently applied; it is used in particular when the binding site of the protein is known and is based on a rigid body concept. A "negative image" of the receptor is used, consisting of a set of overlapping spheres. These spheres fill out the free space of the active site by increasing the sphere radii until the van der Waals surface of the receptor is hit. Orientations of the ligand are then generated by matching subsets of ligand atoms onto subsets of receptor sphere centers. A series of approaches utilizes interaction points as descriptors. The grid program developed by Goodford, for example, computes the interaction energies of a probe with user-specified properties at all points of a grid surrounding the molecule to indicate the desirability of placing various ligand atom types at the grid points. Other methods mainly differ by the probe properties used (hydrogen bond donor and acceptor, water probe, hydrophobic site, etc.) and the placement of the interaction points around the molecule.

The introduction of molecular surfaces leads to a variety of new definitions for search patterns. The solvent accessible surface is defined by rolling a water-sized sphere over the molecule, is often used in these approaches. The first descriptor, measuring the local shape features of the surface, was defined by Connolly. Within this approach a so-called solid angle is determined by picking a point on the surface, constructing a sphere of a particular radius around this point, and measuring the area of the sphere contained within the molecule. Local maxima and minima are defining holes and knobs of the surface, respectively. Complementary patterns of knob and holes are then used to generate possible complex structures. Hendrix and Kuntz improved this method by not only using local extrema. They applied the solid angle method to define shape regions as clusters of adjoining points with similar shape features. These shape regions are then used in an adapted version of the DOCK program as a shape-based filter.

Lin et al. described a surface by a set of critical points, derived from the Connolly algorithm for the generation of the surface accessible surfaces. Using this algorithm, a surface consists of connected domains of atomic size, called faces. Three different faces are defined: convex, saddle-shaped, and concave faces corresponding to domains, where the water probe touches tangentially one, two, or more atoms, respectively. The critical point is then obtained by projecting the gravity center of a face onto the surface, and an averaged normal vector can be assigned to this point. Using the geometric hashing algorithm, these critical points are used effectively in a docking algorithm.

Heiden and Brickmann as well as Duncan and Olsen defined shape descriptors, called surface topography index (STI) and shape index, respectively. These combine the two curvatures of the molecular surface to a value describing the concavity/convexity of the surface at a specific point. Based on linguistic variables introduced by fuzzy set theory, regions with similar values of the shape descriptors can be identified on molecular surfaces. Subdividing the molecular surface into discrete domains according to these similar regions can then be used to define search patterns suitable for active site identification. Cosgrove et al. also used patches enclosing an area of similar regional curvature to describe the surface of a molecule. Therefore, the curvatures of a certain surface point are approximated by fitting spheres to this point and neighbor points in a defined distance. The reciprocals of the radii of the biggest and smallest spheres built for one surface point are defined as the two so-called regional curvatures of the surface. The points are placed in one of five classes (convex, concave, saddle, cylinder, and flat) using certain cutoff values for these curvatures. Neighboring points belonging to the same class are then grouped together to build circular patches, which can be used for the comparison of different surfaces.

Another shape feature beside the convexity is the roughness of the surface. In order to describe this roughness, methods using fractal dimensions can be applied. Petit and Bowie showed that binding sites for small ligands are rougher than protein surfaces in general. Therefore, the roughness can also be used for binding site identification.

In addition to the shape-based criteria mentioned so far, other molecular properties can be used. Meyer et al. for example, used possible hydrogen bond donors and acceptors to define complementarity between different molecules, and the method of Heiden and Brickmann as described above can also use the electrostatic potential calculated on the surface or the local lipophilicity/hydrophilicity for the surface segmentation.

The new method presented in this article is based on fuzzy set theory and is suitable for many different molecular properties.
Fuzzy set theory has already been successfully applied in different areas of pattern recognition and at different stages of the recognition process. We shall demonstrate in particular that the concept of linguistic variables can be considered as very useful in molecular similarity search processes and in the study of complementarity.

Basic Principles of Fuzzy Logic

The concept of fuzzy set theory was introduced in 1965 by Zadeh. By now, a great number of applications have been reported in many different disciplines. Because the field is quite complex and under permanent development, fuzzy set theory cannot be fully introduced in this publication. We only present the concepts we actually use and refer to the literature for more detailed information.

Fuzzy Sets

Fuzzy set theory may be regarded as a generalization of classical set theory. A fuzzy set \( A \) is denoted by an ordered set of pairs. The first element denotes the basic variable \( x \) for the fuzzy set in the definition space \( X \) and the second element, the degree of membership. The latter is defined by a membership function \( \mu_A(x) \), with values lying within the range \( 0 \leq \mu_A(x) \leq 1 \) between zero and complete membership. This normalization is not necessary but is very helpful for the application described in this work.

\[
A = \{(x, \mu_A(x))|x \in X\} \tag{1}
\]

Fuzzy logic allows almost all types of functions for membership definition. In many applications, elements with a high membership lying above a certain value are important. Therefore, \( \alpha \)-level sets are defined as crisp sets of elements belonging to the fuzzy set \( A \) at least to the degree \( \alpha \):

\[
A_\alpha = \{x \in X|\mu_A(x) \geq \alpha\} \tag{2}
\]

Linguistic Variables

One of the basic tools for fuzzy logic is based on the concept of linguistic variables (LVs), whose values are not numbers but words of a natural or artificial language. LVs are groups of fuzzy sets with partially overlapping membership functions over a common (crisp) basic variable \( x \). In order to represent several classes (terms) within an LV, the membership functions should cover all the relevant space of the crisp basic variable \( x \). Generally a linguistic variable \( L \), classified by \( n \) fuzzy sets \( A_i \), can be defined as

\[
L = \{A_1, \ldots, A_n\} \tag{3}
\]

Decision Making in Fuzzy Environments

Usually, the information on which a decision should be based is a set of given crisp function values, like the topographical properties of a molecular surface. In addition, the decision itself should lead to a crisp value (whether the considered configuration is a docking configuration or not). Thus, decision making in fuzzy environments requires three steps:

- Fuzzification of crisp variables into linguistic variables.
- Fuzzy decision from different LVs using fuzzy operators.
- Defuzzification back to crisp values.

Many fuzzy operators have been suggested for fuzzy decisions. These suggestions vary with respect to the generality or adaptability of the operators as well as to the degree to which and how they are justified. In the next sections, the operators used are discussed in detail as far as is necessary. For further details see Zimmermann.

Molecular Descriptors

In this article the description of search patterns is based on molecular descriptors and properties located on the solvent accessible surfaces. The preparation of the surfaces follows a well-known process described elsewhere and results in triangular meshes in 3D space with location-dependent properties assigned to each surface point. Here, a brief introduction of the definition as well as the calculation or mapping of the descriptors on the surface is presented.

Shape Descriptors

The topographical properties of a surface can be quantified mathematically by the two canonical curvatures at each surface point. These canonical curvatures are defined by the eigenvalues of the Hessian matrix and describe the surrounding of the surface point in an atomistic scale. Because our interest is focused on the identification of active sites, the definition of global curvatures introduced by Zachmann et al. can be used to describe the shape feathers of larger surface regions involved in molecular recognition. A surface region of well-defined size around a surface point is approximated by a paraboloid, and the main curvatures of the paraboloid define the global curvatures of the surface point. These global curvatures may be interpreted as average curvatures of the corresponding surface region.

For the shape description, different possibilities combining the two curvatures to one characteristic value are presented in differential geometry. The most important and commonly used ones are the mean and the Gaussian curvature, defined in eqs. (4) and (5), respectively.

\[
H = \frac{1}{2} (k_1 + k_2) \tag{4}
\]

\[
K = k_1 \cdot k_2 \tag{5}
\]

with
In addition, two different methods describing the convexity of molecular surfaces based on curvatures described above are introduced almost at the same time.\textsuperscript{51–53} The STI was already mentioned in the Introduction. This index qualifies the convexity by increasing continuously through five basic shape descriptors (0...4) plus a flatness value (−1) and is calculated as follows:\textsuperscript{51}:

\begin{equation}
STI = \begin{cases} 
\frac{k_1 - k_2}{k_1} & \text{if } k_1 > 0 \text{ and } k_2 > 0 \\
0 & \text{if } k_1 > 0, k_2 \leq 0, \text{ and } |k_1| \geq |k_2| \\
\frac{k_1 + 3k_2}{k_2} & \text{if } k_1 > 0, k_2 \leq 0, \text{ and } |k_1| \leq |k_2| \\
1 & \text{if } k_1 \leq 0 \text{ and } k_2 < 0 \\
-1 & \text{if } k_1 = 0 \text{ and } k_2 = 0 
\end{cases}
\end{equation}

Figure 1. Comparison of the surface topography index (left) and the shape index (right). The values are shown color coded on the molecular surface. Red regions correspond to $STI = 4$ and $s = +1$, gray regions to $STI = 2$ and $s = 0$, and blue regions to $STI = 0$ and $s = -1$.

Figure 2. Hydrogen bond acceptor and donor sites and the hydrogen bond density mapped on the surface of methotrexate. On the left hand side acceptor surface patches are shown in red and donor surface patches in purple. Grey represents no hydrogen bonding ability. On the right hand side the hydrogen bond density is shown color coded on the molecular surface. The yellow-brown color qualifies low and the blue color high hydrogen bond density.
The five basic descriptors can then be assigned to uniform concave (bag, \(STI = 0\)), stretched concave (cleft, \(STI = 1\)), saddle typed (saddle, \(STI = 2\)), stretched convex (ridge, \(STI = 3\)), and uniform convex regions (knob, \(STI = 4\)).

The second descriptor was introduced by Ducan and Olson\(^{52,53}\) and is called shape index. It was originally defined using the local canonical curvatures, but the global curvatures\(^{67}\) can be used in an equivalent manner as shown in the work presented here:

\[
s = -\frac{2}{\pi} \arctan \left( \frac{k_1 + k_2}{k_1 - k_2} \right)
\]  

(7)

Although the definitions of the two descriptors are quite different, the qualitative agreement is astonishingly good. This is shown in Figure 1. Thus, the values 0, 2, and 4 of the STI meet the values \(-1, 0, \) and \(+1\) of the shape index, respectively. The values 1 and 3 of the STI nearly match with the values \(-0.5\) and \(+0.5\) of the shape index.

Additionally Ducan and Olson\(^{52,53}\) introduced a measure characterizing not the appearance (convexity, concavity) but the magnitude of the surface curvature. This so-called "curvedness" is defined as the root mean square of both curvatures:

\[
R = \sqrt{\frac{1}{2} \left( k_1^2 + k_2^2 \right)}
\]  

(8)

### Electrostatic Potential and Local Lipophilicity/Hydrophilicity

The definitions and derivations of these two molecular properties are described elsewhere\(^{68–71}\) and therefore only the equations, used for the calculation of the properties, are given here.

In order to calculate the electrostatic potential of a molecule in solution we use a shifted-force approach based on the Coulomb equation\(^{69}\):

\[
\varphi(\vec{r}) = \frac{1}{4\pi\epsilon} \sum_{i=1}^{N} q_i \left[ 1 - \left( \frac{d_i}{d_0} \right)^2 \right]^2 \text{ if } d_i \leq d_0
\]

\[
0 \quad \text{if } d_i > d_0
\]  

(9)

with

- \(d_i\): distance between atom \(i\) and surface point
- \(d_0\): cutoff parameters
- \(q_i\): partial charge of atom \(i\)
- \(N\): number of atoms

With many examples it could be shown that this equation gives similar results of the electrostatic potential for biomolecules in aqueous solution to those which have been calculated on the basis of the more accurate Poisson-Boltzmann equation.

The Poisson-Boltzmann equation is based on continuum theory, wherein the solute molecule is defined as a low dielectric cavity in a high dielectric medium representing the solvent. The charge distribution of the cavity is approximated by real and partial charges whose positions are given by the atom coordinates of the solute molecule, while the ion distribution in the solvent bulk phase is represented by a constant ionic strength. The disadvantage of this equation is that it must be solved numerically for almost all systems under consideration. Closed analytical solutions are possible only for very simple geometries such as planes or spheres. Very effective methods using finite difference and finite element methods have been developed.\(^{69,70,72–80}\) In this article a finite difference method based on the linearized Poisson-Boltzmann equation [eq. (10)] is used in addition to the modified Coulomb equation in order to investigate the effect of the calculation method on the search pattern generation:

\[
\nabla \cdot \left[ \epsilon(\vec{r}) \nabla \varphi(\vec{r}) \right] - \frac{\kappa^2}{\epsilon} \varphi(\vec{r}) + 4\pi \rho(\vec{r}) = 0
\]  

(10)

with

\[
\kappa = \sqrt{\frac{8\pi\epsilon N_A I}{1000k_BT}}
\]

and

- \(\epsilon(\vec{r})\): dielectric constant
- \(\varphi(\vec{r})\): electrostatic potential
- \(k_B\): Boltzmann’s constant
- \(T\): temperature
- \(\rho(\vec{r})\): charge density of the solute
- \(\kappa\): Debye-Hückel inverse length
- \(I\): ionic strength
- \(N_A\): Avogadro’s constant
- \(\epsilon\): elementary charge

In the literature, many definitions for the calculation of local lipophilicity exist.\(^{81,82}\) In this article the local lipophilicity/hydrophilicity is calculated on the surface according the free energy surface density concept.\(^{83–85}\) Therein, atomic lipophilicity contributions\(^{96–98}\) are mapped on the surface using the distance function introduced by Heiden et al.\(^{71}\):

\[
g_{\text{rad}}(d_i, c, \delta) = \frac{\epsilon^{-2\delta d_i} + 1}{\epsilon^{2\delta d_i} + 1}
\]  

(11)

with

- \(d_i\): distance between atom \(i\) and surface point
- \(N\): number of atoms
- \(c, \delta\): parameters

The local lipophilicity/hydrophilicity is then calculated as follows:
identification of binding sites of biomolecules. Therefore, the proposed fuzzification scheme is designed for this special application and was tested by the visual inspection of different biomolecular complex interfaces. The proposed classes and membership functions for the shape index, the curvedness, the electrostatic potential, the local lipophilicity/hydrophilicity, and the hydrogen bond density are shown in Figure 3. These definitions are the results of adaptation and optimization of the fuzzification scheme introduced by Heiden and Brickmann. Similar definitions can also be used for other molecular properties.

**Similarity/Dissimilarity of Molecular Properties Using Linguistic Variables**

Similarity/dissimilarity plays an important role in pattern recognition. The vagueness of the word itself already implies that there are numerous ways in which the dissimilarity of two objects may be defined. In this work, we use the definition of a dissimilarity function introduced by Heiden and Brickmann. Therein, the dissimilarity is defined by a weighted sum of the differences of the values of corresponding membership functions of both objects:

$$D_{LV}(A, x_A, x_B) = \frac{\sum_{i=1}^{n} w_i |\mu_{A_i}(x_A) - \mu_{B_i}(x_B)|}{\sum_{i=1}^{n} w_i (\mu_{A_i}(x_A) + \mu_{B_i}(x_B))}$$

with

- $A$: LV of corresponding type
- $x_A, x_B$: values of the basic variable of surface point $A$ and $B$, respectively
- $w_i$: weighting factors of class $i$, $0 \leq w_i \leq 1$
- $n$: number of classes of LV $A$

Using this definition, the dissimilarities of all possible objects fall in the interval $0 \leq D_{LV}(A, x_A, x_B) \leq 1$, and identical objects have a dissimilarity of $D_{LV}(A, x_A, x_B) = 0$. Correspondingly, the similarity can be defined as the complement of the dissimilarity:

$$S_{LV}(A, x_A, x_B) = 1 - D_{LV}(A, x_A, x_B)$$

**Segmentation of Molecular Surfaces**

The goal of the segmentation algorithm is to generate surface patches of well-defined size, called domains, which can be used for the comparison of characteristics of different molecules. The surface is divided by combining surface points, which are similar within a fuzzy limit according to a certain molecular descriptor. The algorithm is principally based on the growth of a surface domain, starting at a characteristic reference point. Following the
neighborhood information given by the triangular mesh, the domain ends up when the dissimilarity of the properties of a surface point to the ones of the reference point exceeds a given limit. The domain also ends when a specific distance to the reference point is reached. With this definition, a partial overlap of domains is possible and domains with similar shape and size are obtained.

In the first step of the segmentation procedure, the memberships of each surface point to the different classes of the LV are calculated. Therefore, the basic variable can be chosen from shape index, STI, electrostatic potential, local lipophilicity/hydrophilicity, and hydrogen bond density, and the membership functions of Figure 3 are used.

The following steps are carried out for one class of the linguistic after the other (a flow chart of the segmentation algorithm is shown in Fig. 4):

1. The surface points are sorted according to decreasing membership to the class under consideration.
2. The surface point with the highest membership is chosen as reference point for the new domain.
3. The dissimilarity and similarity of each surface point to the reference point are calculated using the dissimilarity function \([\text{eq. (13)}]\) and eq. (14), respectively.
4. Starting from the reference point and following the triangular network, surface points, identified as neighbors, are added.
As neighbors, we define surface points linked directly by a triangle side to one surface point assigned to the domain. If the distance between the new added point and the reference point is greater as a maximum distance, the point is rejected. The maximum distance is determined by a user-defined limit for the domain size. In order to calculate the distance, not the Cartesian but the shortest distance following the triangular mesh is used.

5. In addition, every selected point with a similarity to the reference point less than 0.5 is rejected.

6. For each remaining point the shortest way to the reference point following the triangular mesh is determined. The third criterion for rejecting a selected point is a surface point, lying on this way between the selected point and the reference point, with a similarity less than 0.5 to the reference point. All remaining points build the new domain.

7. Should the size of the resulting domain fall below a user-defined value, the domain is not considered any longer and the procedure continues at step 10.

Figure 5. Lewis structure of methotrexate.
Figure 6. Two typical surface domains generated with the linguistic variable topography. On the left hand side the whole molecular surface is shown color coded according to the shape index. The red color qualifies convex and grey saddle type regions. On the upper right hand side a very convex region is defined as domain. This domain is characterized as a knob. The domain on the lower side is a saddle type region.

Figure 7. Two typical surface domains generated with the linguistic variable hydrogen bond density. On the left hand side the whole molecular surface is shown color coded. The yellow-brown color qualifies low and the blue color high hydrogen bond density. On the right two domains corresponding to a very high hydrogen bond density are shown. Together these domains cover mostly the pteridine ring system of methotrexate.
8. For each domain point, a membership to the domain is calculated according to eq. 15. The criterion for the degree of membership is based on the distance to the reference point calculated in step 4.

\[ \mu_{\text{dom},i} = \frac{d'_{\text{max}} - d_i}{d_{\text{max}}} \]  

with

\[ d'_{\text{max}} = d_{\text{max}} + 0.1 \, \text{Å} \]

\[ \mu_{\text{dom},i} \]: membership of surface point \( i \)

\[ d_i \]: distance between point \( i \) and reference point following triangular mesh

\[ d_{\text{max}} \]: maximum distance of a surface point belonging to the domain

9. The domain is characterized by the average of the molecular properties, the center point, and the average surface normal. The center point is calculated as the center of mass of the surface points by defining an equivalent mass for each surface point.

10. The surface point with the highest membership not yet assigned to a domain of the investigated class of the LV is identified and the procedure restarts at step 3 using this new reference point.

In order to build domains with a high membership for a certain class, only surface points whose membership to the class is high enough should be used as a reference point. Therefore, the procedure for one class of the LV stops if all surface points not belonging to a domain of this class built so far have a membership smaller than a given value. The specific value depends on the basic variable and comes to 0.85 for the shape index and the STI, 0.7 for the local lipophilicity/hydrophilicity and hydrogen bond density, and 0.9 for the electrostatic potential. These values correspond approximately to the points of intersection of the membership functions of the different classes (see Fig. 3).

With the segmentation algorithm described here, the obtained surface patches are suitable to be used as search patterns for the identification of active sites. This is mainly due to the comparability of the surface patches of different molecules. There are two reasons for this comparability:

1. Due to the circular growth of a domain in the segmentation algorithm, each domain has a compact shape, which is similar to the one of other domains. Thus, there is no need to consider the actual shape of the domains in a comparison procedure, because the similarity is given a priori.

2. By using the membership of the surface points to a domain, the size of two different domains can be adjusted to each other. Therefore, only points with a high membership (\( \alpha \)-level set with large \( \alpha \)) are used for large domains, while an \( \alpha \)-level set with a smaller value \( \alpha \) is used for smaller domains. Additionally, multiple \( \alpha \)-level set combinations of the two domains with approximately equal size can be compared. This multiple \( \alpha \)-level set comparison can be accounted as an approximation of the comparison of the profile of the considered property over the surface patch.

Thus, the complexity of finding suitable search patterns for active site identification is reduced to the comparison of surface domains of well-defined size, shape, and properties.

### Results and Discussion

The applications of the search patterns will be presented in detail in subsequent articles. Here, only results using methotrexate (Fig. 5) taken from the pdb-entry 4dfr89,90 are shown as examples.

For an almost optimal representation of the molecular surface by the search patterns, the size of the domains is relevant. On one hand, the use of domains that are too small results in a huge amount of search patterns and a very ineffective comparison of the surface features. On the other hand, the specific surface features cannot be described by very large domains as the comparison of these local features is not possible. By these criteria, the optimal size of the domains was determined to be in the range between 50 and 150 Å\(^2\) depending on the application, wherein the search patterns should be used and the molecular property used for the segmentation.

The visual inspection of the resulting domains using the MOLCAD II module63–65 shows a good representation of local surface features. In Figures 6 and 7 representative domains of the molecular surface of methotrexate are presented. Therein, only the part of the domain with a membership higher than 0 is shown. For comparison purposes, the complete molecular surface is also displayed. In both pictures, it is demonstrated that the segmentation algorithm can be used to identify local surface features. The two domains in Figure 6 represent a very convex and a saddle type region, respectively. In Figure 7, two domains corresponding to a very high hydrogen bond density are shown. Together, these domains cover mostly the pteridine ring system of methotrexate (see Fig. 5), which builds up to five hydrogen bonds with the enzyme dihydrofolate reductase. It can also be seen that the domains have a nearly circular shape. This is very important for the applications of the search patterns, as mentioned above.

### Conclusion

In this article, a formalism is presented for the representation and a discrete classification of local features of a molecular surface within a scheme based on fuzzy set theory. Therein, surface patches, called domains, are generated with the aid of linguistic variables. In order to generate these domains, a variety of different molecular properties can be used. Among these, the topographical properties, the electrostatic potential, the local lipophilicity/hydrophilicity, and the capability to build hydrogen bonds are very important for molecular recognition. Although domains of only one test molecule are presented in this article, it has been shown that the local surface features are well represented by these do-
mains. In addition, the domains have a good comparability due to their nearly circular shapes and the fact that the sizes of the two domains can be adjusted to each other. Thus, it is demonstrated that this formalism is well suited for generating similarity and complementarity motifs, which can be used as search patterns in active site identifications.

The applications of the search patterns will be presented in two subsequent articles. In the first article, a fuzzy logic treatment for the prediction of structures of biomolecular complexes, at least the active site of proteins with a neuronal network. Domains generated by using the complete Protein Data Bank will be used as a training set for a backpropagation net.

The approach we are dealing with in this series of articles is clearly based on the concept of similarity and complementarity of well-shaped rigid bodies. It can be applied to a variety of fitting problems among molecules but also for larger complexes like viruses or catalytic surfaces. In any case, our treatments should lead to first guesses of possible structures of given partners. For a quantitative treatment of molecular interactions, subsequent technologies (like flexible docking routines based on force fields or free energy calculations) are necessary, which may start from these first guesses.

Acknowledgments

Thomas Exner gratefully acknowledges the Alexander von Humboldt Foundation for a Feodor Lynen research fellowship.

References

1. Fischer, E. Chem Ber 1894, 27, 2985.
44. Connolly, M. L. Biopolymers 1992, 32, 1215.