Segmentation of protein surfaces using fuzzy logic

Wolfgang Heiden and Jürgen Brickmann

Institut für Physikalische Chemie I, Technische Hochschule Darmstadt, Darmstadt, Germany

An algorithm has been developed that can be used to divide triangulated molecular surfaces into distinct domains on the basis of physical and topographical molecular properties. Domains are defined by a certain degree of homogeneity concerning one of these properties. The method is based on fuzzy logic strategies, thus taking into consideration the smooth changes of the properties considered along complex macromolecular surfaces. Scalar qualities assigned to every node point on a triangulated surface are translated into linguistic variables, which can then be processed using a special fuzzy dissimilarity operator.

Possible applications are demonstrated using surface segmentation for properties like electrostatic potential, lipophilicity and shape for the analysis of serine proteinase substrate/inhibitor specificity.

Keywords: automatic segmentation, molecular surfaces, fuzzy logic, surface topology, surface shape, electrostatic potential, local lipophilicity

INTRODUCTION

The selectivity of proteins interacting with macromolecular compounds is related to both physical and topological properties, which can, in principle, be attributed to electrostatic and hydrophobic interactions, and steric relations, respectively. There are a vast number of papers dealing with the calculation and display of several physical properties on molecular surfaces.[1-7] The importance of topographical complementarity in molecular recognition has also been stressed widely,[8-11] leading to automated docking procedures dominantly based on shape.[12-14] Although modern computer graphics provide many features that help the scientist in analyzing macromolecular structure-activity relationships intuitively, detailed examination of three-dimensional (3D) protein structures in regard to physical and topographical surface qualities is still a time-consuming process that requires expertise in molecular modeling strategies. Even with tools like transparent, colored surfaces and a variable 3D cursor,[15] the segmentation of a molecular surface into more or less homogeneous domains (according to certain qualities) is a task not easy to perform interactively.

In addition, color-coded representations of physical qualities on molecular surfaces may be misleading. Keen border lines separating regions of different quality may be seen where there are none, just as a consequence of the color coding and of common abilities to distinguish between different colors.

The automatic segmentation and classification of molecular surfaces into distinct domains can be helpful for the analysis of selective protein-protein interactions. However, straightforward strategies for surface segmentation do not work in most cases without user interaction,[16] because of the lack of unequivocally defined borders between different surface regions as well as a consequence of the overall surface complexity.

In this work, we present the results of an attempt to overcome these difficulties. An algorithm has been developed based on the principles of fuzzy logic. This relatively new mathematical tool was designed for the treatment of exactly the kind of problems mentioned earlier.

BASIC PRINCIPLES OF FUZZY LOGIC

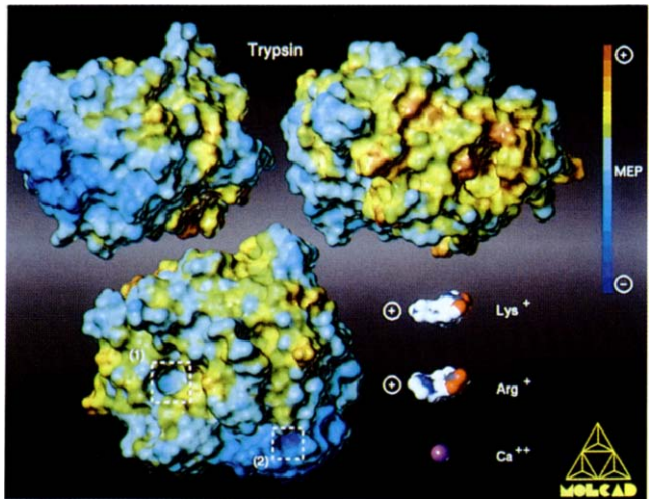
The concept of *fuzzy logic* was introduced almost 30 years ago by Zadeh.[17] Lying dormant for many years, it has been rediscovered in the mid-80s for regulation in microelectronics, automatic process regulation or in operations research. By now, fuzzy set theory has many applications in a large variety of different domains. Because the field is quite complex and in development, the basics of fuzzy logic can not be discussed fully in this paper. We refer to the literature[18,19] for detailed representations. Here, we only present those concepts that are directly used for segmentation of triangulated molecular surfaces.

Fuzzy sets

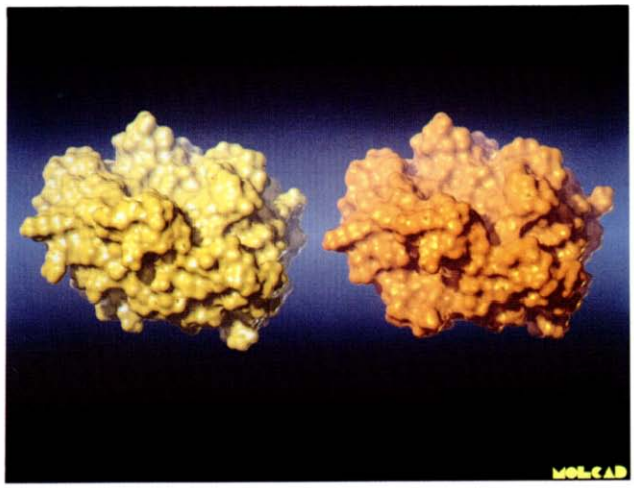
Fuzzy set theory may be seen as a generalization of classical set theory, each element of a fuzzy set, $\tilde{A}$, being defined by a function value, x, in definition space, X, together with its degree of membership to $\tilde{A}$. The latter is defined by a membership function, $\mu_{\tilde{A}}(x)$, with values that lie normally within a range $0 \leq \mu_{\tilde{A}}(x) \leq 1$ between zero and complete membership.

$$\tilde{A} = \{(x, \mu_{\tilde{A}}(x)) \mid x \in X\} \quad (1)$$

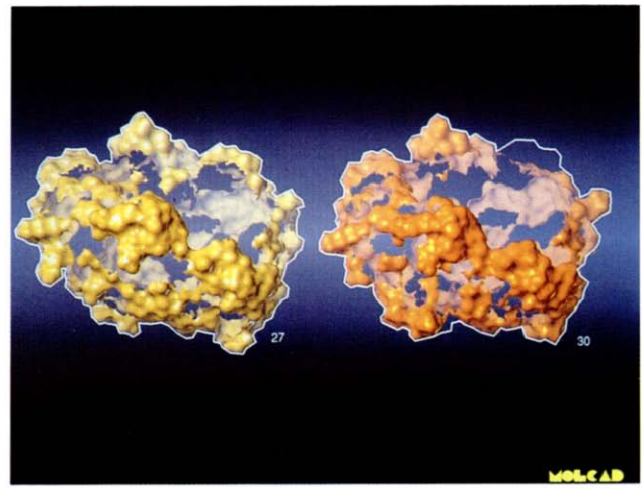
Address reprint requests to Professor Brickmann at Institut für Physikalische Chemie I, Technische Hochschule Darmstadt, Petersenstr. 20, D-64287 Darmstadt, Germany.
Received 15 September 1993; accepted 27 October 1993



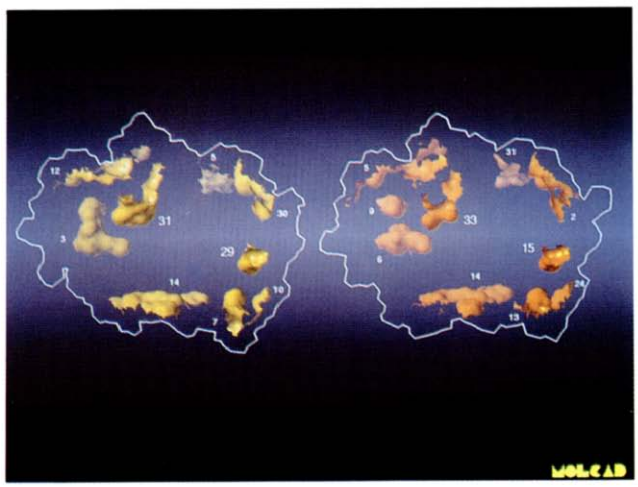
Color Plate 1. Characteristic regions on the surface of trypsin. Colors represent the MEP, varying from blue (negative) to red (positive). The protein is shown in different orientations. The specificity pocket (1) for substrate recognition (Lys, Arg) and the Ca^{2+} binding pocket (2) are labeled.



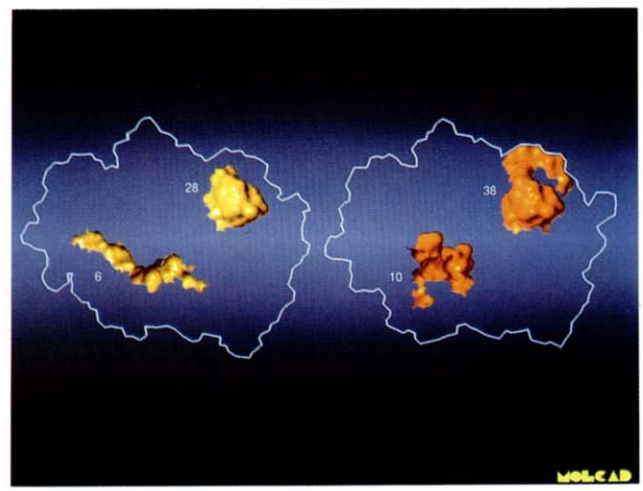
a



b



c



d

Color Plate 2. Automatic segmentation of trypsin (left) and trypsinogen (right) by local shape (given by an STI). Numbers at single domains refer to Table 1. Related domain pairs are named by the domain numbers trypsin/trypsinogen, respectively. The outline of the complete surface (a) is displayed as a white line around the domains in (b) through (d). (a) Triangulated molecular surface of trypsin and trypsinogen (solid representation). (b) Core domains. (c) Several corresponding segments of significant size. Specificity pocket (31/33) and Ca^{2+} binding pocket (29/15) are highlighted. (d) One pair of corresponding segments (6/10) shows significant structural changes during trypsinogen activation.

In classical (crisp) sets, $\mu_{\tilde{A}}(x)$ can only be 0 or 1, while fuzzy logic allows almost any type of function for membership definitions.

Linguistic variables

One of the most important tools in applications of fuzzy set theory is the concept of *linguistic variables* (LV).[20] These are groups of fuzzy sets with partially overlapping membership functions over a common (crisp) basic variable x. In order to represent several classes within an LV, the membership functions should cover all the relevant definition space of the basic variable x with membership function values $0 < \mu_{\tilde{A}}(x) < 1$ (Figure 1). (Values of 0 or 1 are assigned to the rest of the definition space in all membership functions.) The overlap of these functions defines the fuzziness. Generally, a linguistic variable, $\mathcal{L}$, classified by n fuzzy sets $\tilde{A}_i$, can be defined as

$$\mathcal{L} = \{\tilde{A}_1, \ldots, \tilde{A}_n\} \tag{2}$$

Or, together with Equation (1)

$$\mathcal{L} = \{(x, \mu_{\tilde{A}1}(x)), \ldots, (x, \mu_{\tilde{A}n}(x)) \mid x \in X\} \tag{3}$$

Decision making in fuzzy environments

Usually, the information on which a decision should be based is given by crisp function values. (For molecular surface segmentation, this means certain scalar qualities are assigned to every node point on a triangulated surface.) Also, the decision itself shall again lead to a crisp value (in this case, the binary decision between continuation or limitation of a surface domain). However, in order to apply fuzzy logic tools to a problem, it has to be defined by linguistic variables. Thus decision making requires three steps:

(1) fuzzification of crisp variables into linguistic variables
(2) fuzzy decision from different LV using fuzzy operators
(3) defuzzification back to a crisp value

The details of these steps are discussed with the specific application patterns as far as necessary. (For further details see Reference 19.)

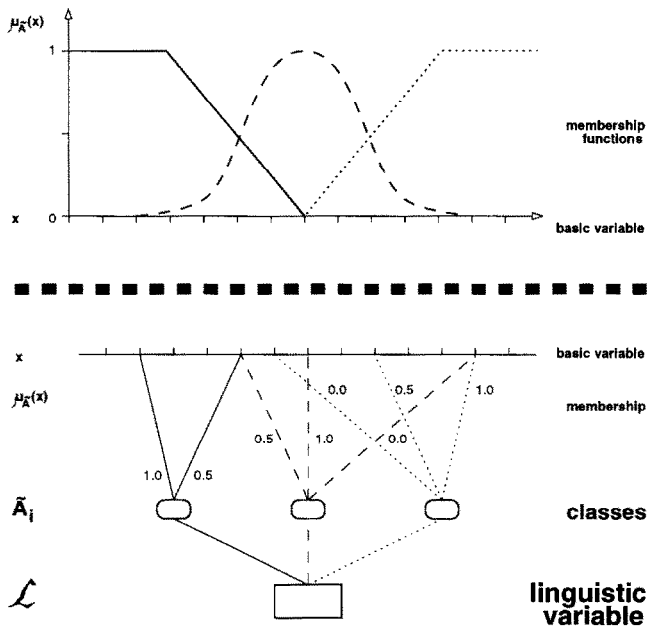


Figure 1. Schematic representation of a linguistic variable with three elements (classes) whose membership functions partially overlap. Membership functions in relation to a common basic variable x (upper). Definitions from the projection of membership function values (lower).

SEGMENTATION OF MOLECULAR SURFACES

A rather simple way to subdivide molecular surfaces into discrete domains has already been introduced in context with the search for topologically significant surface points, by using a new concept of *global canonical curvatures*.[21] The surface distance following the triangle mesh (which was used for the definition of the global curvatures[21]) gave the limit for a surface domain around a certain reference point as a "ring" of selected neighbors. This approach can be improved using LVs instead of the global curvatures. Now the surface distance is replaced by dissimilarity of the LV as the criterion for determination of surface domain limits.

Dissimilarity of 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, D, of two objects, a and b, may be defined, according to the actual problem. However, in any case the following expression can be formulated for quantifying dissimilarity for all possible objects a and b:

$$D(a, b) \geq 0 \tag{4}$$

$D(a, b) = 0$ means that both objects are identical.

$$D(a, b) = 0 \iff a = b \tag{5}$$

In order to map an intuitively defined approach into a formal concept, we designed a dissimilarity function, D_{LV}, for linguistic variables, which fulfils several conditions:

(1) The dissimilarity, D_{LV}, of two arbitrary linguistic variables of the same type, $\mathcal{A}$, $\mathcal{B}$, should be defined as generally as possible. This may be done, for example, by summarizing the absolute values of the differences of all membership function values, $\Delta\mu_i$, over every n class of both LVs.

$$D_{LV}(\mathcal{A}, \mathcal{B}) = \sum_{i=1}^{n} \Delta\mu_i \tag{6}$$

with

$$\Delta\mu_i = |\mu_{\tilde{A}i}(x) - \mu_{\tilde{B}i}(x)| \tag{7}$$

(2) Independent weighting of single classes should be possible by the employment of weighting factors, w_i.

$$D_{LV}(\mathcal{A}, \mathcal{B}) = \sum_{i=1}^{n} w_i \cdot \Delta\mu_i \tag{8}$$

(3) As all membership function values of different classes of an LV do not necessarily overlap (Figure 1), the

difference in pairs of single function values might pretend identity, if the membership for this LV class is zero for both objects. In order to preclude this, the partial differences of different classes, i, can be weighted by the absolute value of the membership, $\Sigma\mu_{\tilde{\imath}}$.

$$\Sigma\mu_{\tilde{\imath}} = \mu_{\tilde{A}i}(x) + \mu_{\tilde{B}i}(x) \tag{9}$$

A reasonable choice of the normalization of the membership functions is $\mu_{\tilde{\imath}} = 0$ for nonmembership and $\mu_{\tilde{\imath}} = 1$ for complete membership. Although this is not necessarily so, only such μ_τ types will be considered here. Analogously, D_{LV} should also vary between absolute identity ($D_{LV} = 0$) and a maximum dissimilarity (defined here for reasons of consistency as $D_{LV} = 1$). This is ensured by Equation 10, because the sum of both membership function values (denominator) can never become smaller than their difference (counter).

$$D_{LV}(\mathcal{A}, \mathcal{B}) = \frac{\Sigma_{i=1}^{n} w_i \Delta\mu_{\tilde{\imath}}}{\Sigma_{i=1}^{n} w_i \Sigma_{\mu\tilde{\imath}}} \tag{10}$$

(4) A special feature of the dissimilarity function defined in Equation 10 reveals itself at a transition from one LV class into another. Such a transition within a segment should be possible, but is rendered difficult by the fuzzy segmentation criterion. Thus, it seems reasonable to give differences higher values if one membership is zero, compared to an identical absolute difference within a class.

For calculating the dissimilarity, D_{LV}, of two linguistic variables, $\mathcal{A}, \mathcal{B}$, the following function (which is an extended writing of Equation 10) has been defined according to the previously named criteria, fusing the fuzzy decision (i.e., the combination of fuzzy sets by fuzzy operators leading again to fuzzy sets[22]) and defuzzification (i.e., the translation of the fuzzy decision sets back to a crisp—i.e., scalar—value) within a single step.

$$D_{LV}(\mathcal{A}, \mathcal{B}) = \frac{\Sigma_{i=1}^{n} w_i |\mu_{\tilde{A}i}(x) - \mu_{\tilde{B}i}(x)|}{\Sigma_{i=1}^{n} w_i (\mu_{\tilde{A}i}(x) + \mu_{\tilde{B}i}(x))} \tag{11}$$

with

$\mathcal{A}, \mathcal{B}$ = linguistic variables of the same type
$\mathcal{A} = \{(x, \mu_{\tilde{A}1}(x)), \ldots, (x, \mu_{\tilde{A}n}(x))\}$
$\mathcal{B} = \{(x, \mu_{\tilde{B}1}(x)), \ldots, (x, \mu_{\tilde{B}n}(x))\}$
$0 \leq \mu_{\tilde{A}i}, \mu_{\tilde{B}i} \leq 1$
$0 \leq w_i \leq 1$
w_i = weighting factor for class i
n = number of classes in $\mathcal{A}, \mathcal{B}$

From this definition it follows that

$$0 \leq D_{LV}(\mathcal{A}, \mathcal{B}) \leq 1 \tag{12}$$

Fuzzification of molecular surface properties

Segmentation of molecular surfaces by fuzzy dissimilarity criteria according to certain surface qualities first requires the translation of a scalar quality value into a linguistic variable. An adequate fuzzification of each quality is essential for the success of the whole procedure and has to be defined very carefully. The definition of membership functions should be guided by intuitive insight into the actual problem. There is, of course, no unique way for such a definition. Whether a fuzzification scheme is adequate or not can be finally decided only by the success of the procedure.

Segmentation by the molecular electrostatic potential (MEP) projected onto protein surfaces (as the basic variable), for example, is possible using the fuzzification scheme displayed in Figure 1, but there may be arbitrarily many other procedures leading to reasonable results. In the present case, the linguistic variable $\mathcal{L}_{mep}$ = local molecular electrostatic potential is built from the classes (fuzzy sets)—negative, neutral and positive potential—represented by solid, dashed and dotted lines, respectively.

Better results can be achieved, however, by adding the classes highly negative and highly positive potential to both sides of the MEP spectrum. The neutral membership function may be simplified using linear quality dependence, leading to the linguistic variable

$$\begin{aligned}\mathcal{L}_{mep} = \{ \quad & (\text{highly negative}, \mu_{\tilde{-}\,\tilde{-}}(\text{MEP})); \\ & (\text{negative}, \quad \mu_{\tilde{-}}(\text{MEP})); \\ & (\text{neutral}, \quad \mu_{\tilde{0}}(\text{MEP})); \\ & (\text{positive}, \quad \mu_{\tilde{+}}(\text{MEP})); \\ & (\text{highly positive}, \mu_{\tilde{+}\,\tilde{+}}(\text{MEP})) \quad \}\end{aligned}$$

as shown in Figure 2.

A molecular lipophilicity potential (MLP) can be encoded in a similar way in a three-class linguistic variable

$$\begin{aligned}\mathcal{L}_{lipo} = \{ \quad & (\text{hydrophilic}, \mu_{\tilde{h}}(\text{MLP})); \\ & (\text{neutral}, \quad \mu_{\tilde{n}}(\text{MLP})); \\ & (\text{lipophilic}, \quad \mu_{\tilde{l}}(\text{MLP})) \quad \}\end{aligned}$$

The shape of molecular surfaces can be described in many ways.[8-14,21] The definitions differ profoundly, so that each one may require its own fuzzification scheme. The shape analysis presented in this article was performed on the basis of a surface topography index (STI), defined by the authors as a single quality for convexity increasing continuously through five basic shape descriptors *plus* a flatness value as a measure of shape intensity.[23,24] This suggests a six-class linguistic variable

$$\begin{aligned}\mathcal{L}_{sti} = \{ \quad & (\text{bag}, \quad \mu_{\tilde{B}}(\text{STI})); \\ & (\text{cleft}, \quad \mu_{\tilde{C}}(\text{STI})); \\ & (\text{saddle}, \quad \mu_{\tilde{S}}(\text{STI})); \\ & (\text{ridge}, \quad \mu_{\tilde{R}}(\text{STI})); \\ & (\text{nob}, \quad \mu_{\tilde{N}}(\text{STI})); \\ & (\text{plateau}, \quad \mu_{\tilde{P}}(\text{absolute curvature})) \quad \}\end{aligned}$$

as shown in Figure 3.

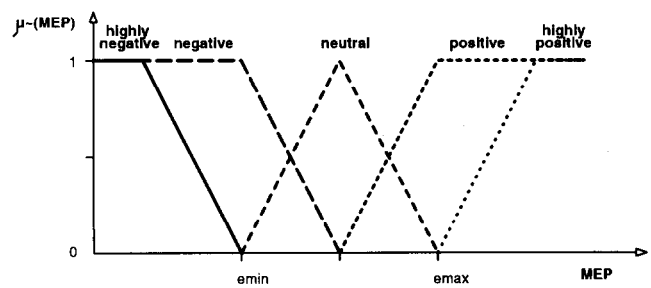


Figure 2. Calculation of the linguistic variable $\mathcal{L}_{\text{mep}}$ from MEP. emin and emax define the relevant quality range.

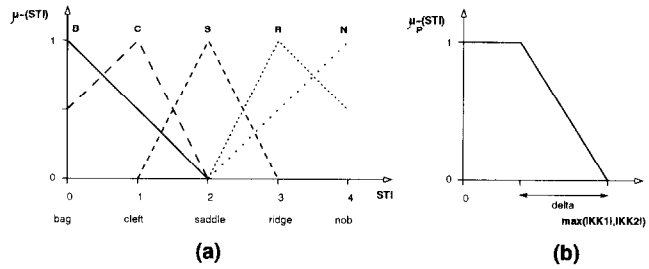
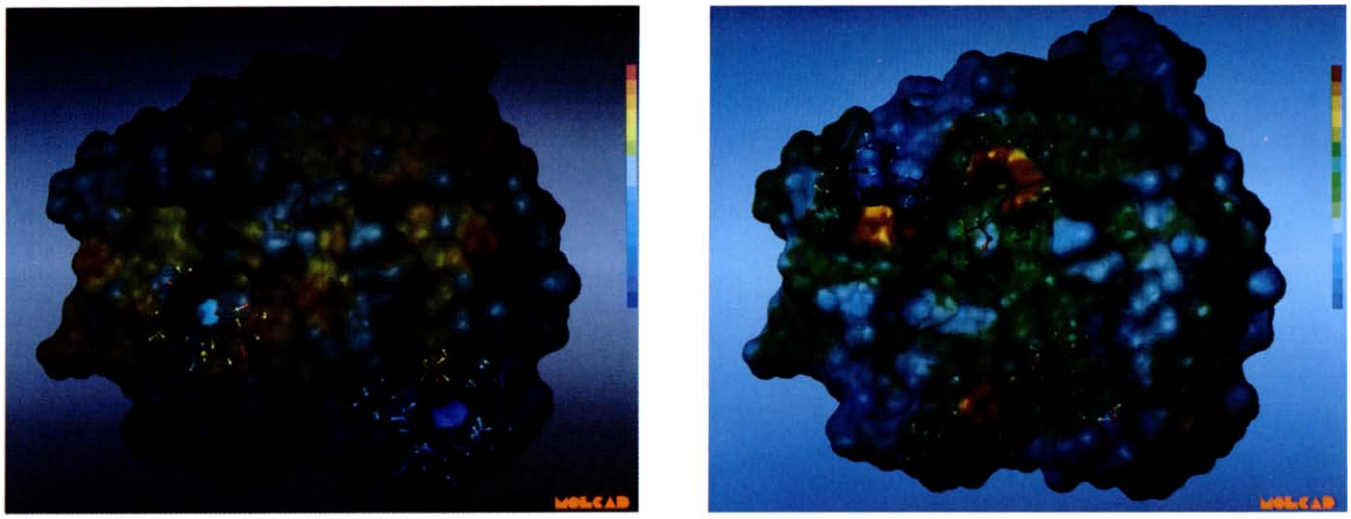
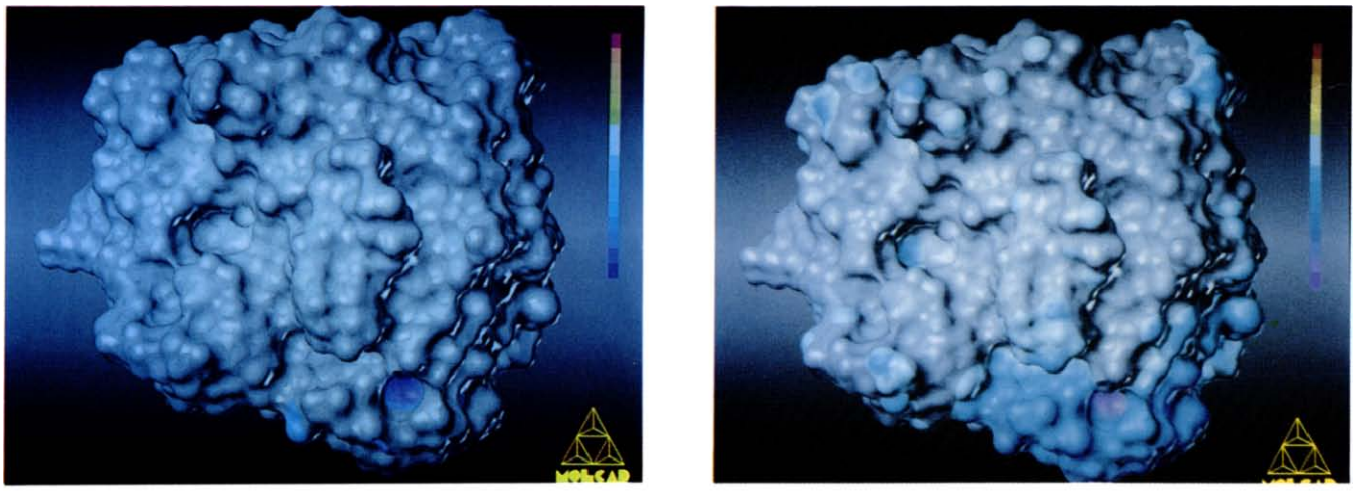


Figure 3. Linguistic variable $\mathcal{L}_{sti}$ for local shape. (a) Membership functions for five shape classes. (b) Membership function for the additional class, plateau, *from the absolute regional curvature. Position and slope of the relevant function range are provided by delta.*



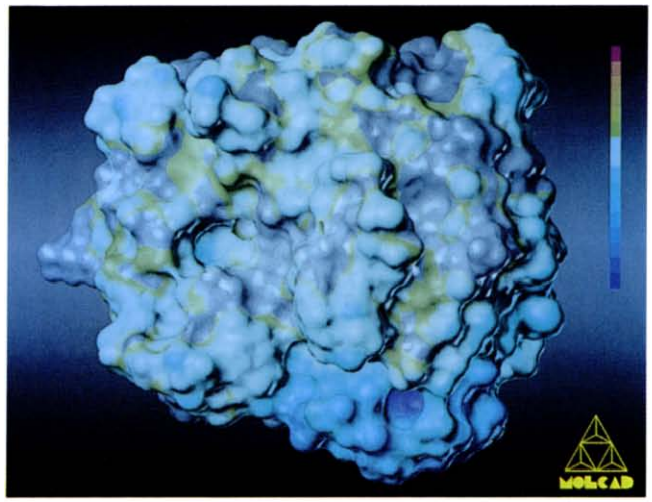
a b

Color Plate 3. MEP and MLP domains on the trypsin surface. A transparent surface of trypsin is shown color coded according to the MEP (a) and MLP (b), given by Equations 13 and 14, respectively. Certain surface domains (found automatically) are displayed solid and highlighted, together with the amino acid residues contributing to these surface segments in *ball-and-stick* representation. MEP: bottom of specificity pocket (central left) and Ca^{2+} binding pocket (lower right). MLP: three lipophilic spots near the active site.

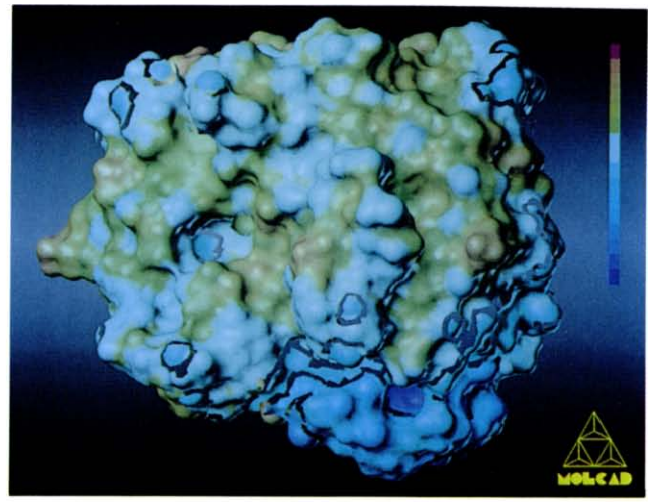


a b

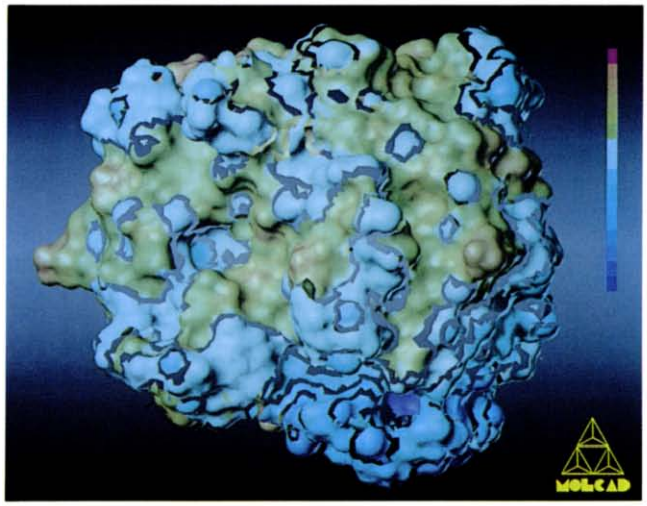
Color Plate 4. Crisp (a)–(c) and fuzzy (d)–(f) MEP surface analysis of trypsin. The orientation of the molecule is chosen similar to that of Color Plate 1 with both the specificity and the Ca^{2+} binding pocket visible. Crisp contour cuts are projected onto a grey trypsin surface for regions with MEP values (a) <0 kcal/mol, (b) <150 kcal/mol and (c) <200 kcal/mol. The results of fuzzy segmentation are shown for (d) $D_{LV} > 0.5$; emin,emax $= \pm 150$ kcal/mol, (e) $D_{LV} > 0.3$; emin,emax $= \pm 150$ kcal/mol and (f) $D_{LV} > 0.3$; emin,emax $= \pm 200$ kcal/mol.



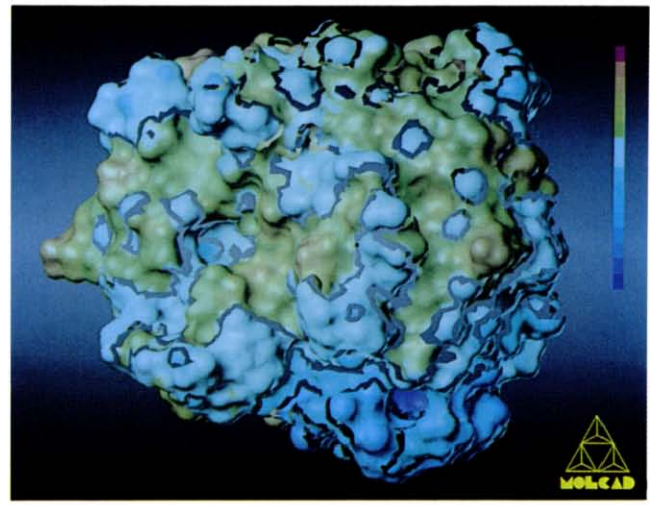
c



d



e



f

Color Plate 4. *Continued*

Automatic segmentation of triangulated surfaces

For an automatic segmentation of molecular surfaces into distinct domains, the computer program Molecular Analysis by Fuzzy Intelligence Algorithms (MAFIA) has been developed on the basis of the dissimilarity of two linguistic variables defined in Equation 11. Surfaces given as a triangular mesh in 3D space with location-dependent qualities assigned to each surface point (which is a node between adjacent triangles) are divided into separate, homologous domains. Neighboring domains differ with regard to a certain surface quality, whose value is characteristic for each domain (within a fuzzy limit).

The algorithm is based principally on the growth of a surface domain, starting at a characteristic reference point (e.g., the point with the highest MEP absolute not yet assigned to another domain). Linguistic variables are assigned in advance to each surface point and are updated continuously for an average of the actual domain. Following the neighborhood information given by the triangular mesh, the domain ends when the dissimilarity of a surface point to the domain average, or its direct neighbor within the domain, exceeds a given limit. The borders of other domains already defined also put an end to segment growth.

Sequentially working its way through all triangle node points, the program achieves complete segmentation of a triangulated surface. Segmentation can be performed for the MEP, MLP and STI based on global curvatures.[21]

APPLICATIONS

Some applications of the new method are demonstrated by an analysis of the digestive enzyme trypsin and its inactive precursor trypsinogen. Trypsin is known to cleave peptide bonds, specifically at the C side of Lys or Arg side chains, and is also known to bind a Ca^{2+} ion with high affinity (Color Plate 1).

The 3D structures were taken from the Brookhaven Protein Data Bank[25] using the entries 1TPO and 2TGT for trypsin and trypsinogen, respectively. Triangulated surfaces were derived either from triangulation of dotted surfaces[26]

from Connolly's MS program[27] or by a new method based on 3D electron density maps,[28,29] after addition of H-atoms to each protein structure by standard distances and angles. The MEP was calculated from atomic partial charges taken from a protein standard library, based on *ab initio* calculations by Rullmann.[30] The MEP at each surface point was defined by the following distance function

$$\mathrm{MEP} = \sum_{i=1}^{N} \frac{q_i}{d_i} \tag{13}$$

with N being the number of atoms and q_i the Coulomb partial charge assigned to atom i. d_i is the distance of atom i from the location of the surface point.

Local lipophilicity was calculated as a molecular lipophilicity (pseudo) potential on the basis of atomic lipophilicity contributions[31,32] with a distance function defined earlier.[33]

$$\mathrm{MLP}_{HM} = \frac{\sum_{i=1}^{N} f_i \cdot g(d_i)}{\sum_{i=1}^{N} g(d_i)} \tag{14}$$

with

$$g(d_i) = \frac{e^{-ab} + 1}{e^{a(d_i - b)} + 1}$$

f_i is the partial lipophilicity of atom i. The function $g(d_i)$ contains the distance dependence. The constant values a and b were chosen as 1.5 and 4.0, respectively.

Topography representation was obtained from an STI[23] as a measure of convexity based on five basic shape descriptors defined by global curvatures.[21]

Topographical analysis of trypsin and trypsinogen

The structure of the digestive enzyme trypsin is very similar to that of its inactive precursor trypsinogen. However, some structural changes occur during zymogen activation by cleavage of the N-terminal hexapeptide,[34] which obviously affect substrate affinity.

Segmentation by means of topography was carried out for trypsin (pure) and trypsinogen (3D structure without the activation hexapeptide). Trypsin was automatically divided into 36 domains, while 60 domains for trypsinogen were found. However, disregarding insignificantly small domains (i.e., smaller than 25 Å^2 in size), there are 25 domains for trypsin and 30 for trypsinogen that show good correlation (Table 1).

For each of the two molecular surfaces, a large convex core domain was found containing more than 50% of the total surface area. Leaving these two domains unconsidered—as well as topographically insignificant ones—homologies and differences in shape can be examined easily (Color Plate 2).

Comparing the pocket-shaped segments of both molecules, the change in shape between trypsinogen and trypsin (shown previously[26]) can be seen easily (Color Plate 2b). Another obvious structural change reveals itself in the domain pair no. 6 (trypsin)–no. 10 (trypsinogen). This region lies near Ile 16, which is the N-terminal residue in trypsin after the removal of six preceding residues from the proenzyme trypsinogen. It is not surprising that essential structural changes take place at this site. In fact, segment no. 10 of trypsinogen, being built of residues Ile 16–Gly 19, Ile 138–Ser 146, Pro 152, Lys 156–Leu 158, Gly A188–Asp 194 and Val 213, Cys 220–Ala A221, can be correlated to the activation domain found by Huber and Bode,[34] which consists of four separate regions of the peptide chain: Ile 16–Gly 19, Gly 142–Pro 152, Gly A184–Gly 193 and Gly 216–Asn 223.

Segmentation by physical qualities

As selectivity in molecular recognition is not only due to topographical complementarity, but also to a combination of energetical and entropical effects, classification of protein surfaces by means of local electrostatic or hydrophobic properties may be helpful for structure-activity relationship (SAR) research. In Color Plate 3, the result of segmentation of the trypsin surface according to MEP and MLP is shown. Segmentation by MEP homogeneity leads to separate domains at the bottom of both the specificity and the Ca^{2+} binding pocket, showing negative potential with a strong gradient. Actually, the small spot representing the Ca^{2+} binding site shows the maximum absolute potential all over the trypsin surface. Seven residues between Glu 70 and Phe 82 contribute to this little surface spot with four O atoms. The electronegative spot at the bottom of the specificity pocket is also consistent with the specific cleaving of peptide bonds at the C side of Lys and Arg residues performed by trypsin. It is built mainly by residues Asp 189 and Ser 190 together with Gly 216, 219 and 226.

The analysis of the MLP results in several distinct lipophilic surface domains. Three of these are located in the vicinity of the active site, indicating the contribution of hydrophobic interactions to the substrate binding of trypsin. This suggestion is supported by the similarity found in the local lipophilicity pattern of the contact surface of trypsin and its inhibitor protein PTI.[33]

DISCUSSION

It has been demonstrated that fuzzy logic techniques can be successfully applied to the segmentation and classification of molecular surfaces. A rigorous segmentation can also be achieved in a straightforward manner on the basis of cuts through the triangular mesh along certain contour values defined for any position-dependent property that is assigned to the triangle node points. Following this approach, segments are defined by a crisp quality range, rather than homogeneity. One major disadvantage of the latter method is obviously the arbitrariness of the choice of contour values. When, for example, the MEP on the trypsin surface is chosen as the segmentation quality, the only contour value not based on a subjective choice is the neutral potential (MEP = 0.0 kcal/mol). Color Plate 4 illustrates the benefit of fuzzy segmentation over a crisp contour cut using this example.

When all surface parts with positive MEP are removed, the Ca^{2+} binding pocket (see also Color Plate 1, region (2)) almost exclusively remains (Color Plate 4a).

An examination of the trypsin/PTI contact region with the MEP definition used in this article revealed that the contact

Table 1. Topographical domains of trypsin and trypsinogen. Corresponding domains are oriented pairwise. The index numbers of the domains are related to the order of segment definition by the program MAFIA.

Trypsin		Trypsinogen		
Domain No.	Area [Å^2]	Domain No.	Area [Å^2]	Description
27	4506	30	4190	core domain
18	198	38	386	isolated convex region (See Color Plate 2c.)
6	149	10	309	cleft (trypsin), pocket (trypsinogen) (See Color Plate 2c.)
15 32	59 187	35	237	cleft, two segments in trypsin
14	202	14	201	embedded plateau
31	158	33	113	specificity pocket (See Color Plate 2b.)
3	151	6 9	126 61	large pocket, two segments in trypsinogen
30	107	2	135	declining cleft
12	126	5	126	cleft branching out
4	124	32	93	flat cleft
13	113	45 34	56 36	cleft leading into a pocket, two segments in trypsinogen
2	100	16	88	cleft
—	—	36	74	without correlation; side chain protruding from core domain
7	71	13	61	pocket
23	67	20	69	flat cavity
19	64	22	66	narrow cleft
5	60	31	62	flat cavity
29	60	15	58	Ca^{2+} binding pocket (See Color Plate 2b.)
20	59	1	51	flat cavity
10	48	24	39	narrow pocket
11	38	39	44	topologically insignificant
—	—	3	44	
18	43	21	28	
21	39	17	36	
24	36	26	24	
9	19	7	36	
25	33	27	34	
—	—	19	33	

surfaces of both molecules show good complementarity above/below an MEP value of about 150 kcal/mol.[23] With this result in mind, another contour cut was calculated showing only those surface regions with MEP values lower than 150 kcal/mol. As can be seen in Color Plate 4b, this procedure clearly cuts out the bottom of the trypsin specificity pocket (see also Color Plate 1, region (1)), leaving the Ca^{2+} binding pocket embedded in a large, continuous surface domain. Another contour cut at 200 kcal/mol delivers a completely different result, with a single segment covering the whole molecule with only some spots in between left out (Color Plate 4c).

The fuzzy segmentation, on the other hand, cleaves out both specificity and the Ca^{2+} binding pocket, even with a quite rigorous segment definition ($D_{LV} > 0.5$, Color Plate 4d). With a less-restrictive segmentation limit ($D_{LV} > 0.3$), additional segments are found with no significant difference between the results with emin,emax = ± 150 kcal/mol (Color Plate 4e) and emin,emax = ± 200 kcal/mol (Color Plate 4f). (For the definition of emin,emax, see Figure 2.)

It seems that the results derived from fuzzy logic analysis of large molecular surfaces can, in principle, also be obtained by crisp segmentation methods. The latter, however, require preceding individual investigation of the surface in question, while the fuzzy algorithm is much more tolerant concerning the input parameters. The new method presented in this article seems to be a valuable tool for a first examination of protein surfaces with little *a priori* information.

CONCLUSION AND OUTLOOK

In this paper, a method for the automatic segmentation of large molecular surfaces by means of physical and topographical properties is presented. The applicability of the method is demonstrated with some examples. In particular, it has been shown that the method can be used to extract those surface parts (from the complete surface) that are relevant to molecular recognition.

We are currently working on the extension of segmentation criteria to a combination of several qualities using compensatory fuzzy operators.

The color figures were generated using the molecular modeling program MOLCAD,[35] developed in the group of the authors and which is now part of the SYBYL software package of Tripos Associates, St. Louis, MO 63117 USA.

COMPUTATIONAL REMARKS

With neighborhood information for the surface points given by the triangular mesh, and qualities assigned to every point, the procedure is extremely fast. A complete segmentation of a protein surface (about 20,000 surface points) requires only 5 seconds CPU calculation time on a Silicon Graphics IRIS Indigo R4000 for each quality.

ACKNOWLEDGMENTS

This work was supported by the Fonds der Chemischen Industrie, Frankfurt. We would like to thank Thomas Goetze for helpful discussions and Harriet Seward for carefully reading the manuscript.

REFERENCES

1 Max, N. Computer representation of molecular surfaces. *J. Mol. Graphics* 1984, **2,** 8–13
2 Quarendon, P., Naylor, C.B. and Richards, W.G. Display of quantum mechanical properties on van der Waals surfaces. *J. Mol. Graphics* 1984, **2** (1) 4–7
3 Náray-Szabó, G. Electrostatic complementarity in molecular associations. *J. Mol. Graphics* 1989, **7** (2) 76–81
4 Sjoberg, P. The use of the electrostatic potential at the molecular surface in recognition interactions: Dibenzo-*p*-dioxins and related systems. *J. Mol. Graphics* 1990, **8,** 81–85
5 Lichtenthaler, F.W., Immel, S. and Kreis, U. Evolution of the structural representation of sucrose. *Starch/Stärke* 1991, **43** (4) 121–132
6 Bohacek, R.S. and McMartin, C. Definition and display of steric, hydrophobic, and hydrogen-bonding properties of ligand binding sites in proteins using Lee and Richards accessible surface: Validation of a high-resolution graphical tool for drug design. *J. Med. Chem.* 1992, **35,** 1671–1684
7 Chapman, M.S. Mapping the surface properties of macromolecules. *Protein Science* 1993, **2,** 459–469
8 Connolly, M.L. Measurement of protein surface shape by solid angles. *J. Mol. Graphics* 1986, **4,** (1) 3–6
9 Mezey, P.G. Global and local relative convexity and oriented relative convexity: Application to molecular shapes in external fields. *J. Math. Chem.* 1988, **2,** 325–346
10 Mezey, P. "Molecular Surfaces," In *Reviews in Computational Chemistry* (K. Lipkowitz and D. Boyd, Eds.) VCH, Weinheim (1990) 265–294
11 Cano, F.H. and Martínez-Ripoll, M. On shape. *J. Mol. Struct. (Theochem)* 1992, **258,** 139–158
12 Connolly, M.L. Shape complementarity at the hemoglobin $\alpha 1\beta 1$ subunit interface. *Biopolymers* 1986, **25** (7) 1229–1247
13 Wang, H. Grid-search molecular accessible surface algorithm for solving the protein docking problem. *J. Comp. Chem.* 1991, **12** (6) 746–750
14 Shoichet, B., Bodian, D. and Kuntz, I. Molecular docking using shape descriptors. *J. Comp. Chem.* 1992, **13** (3) 380–397
15 Brickmann, J., Goetze, T., Heiden, W., Moeckel, G., Reiling, S., Vollhardt, H. and Zachmann, C.-D. "Interactive visualization of molecular scenarios with the MOLCAD/SYBYL package," In *Visualization and Innovation in Data Visualization* (J.E. Bowie, Ed.) in press
16 Mittelhäußer, G. In "Segmentierung von MR-Volumendaten mit Bereichswachstumsverfahren," *Visualisierung in der Medizin* (proceedings) Freiburg, Germany March 10–11, 1993
17 Zadeh, L.A. Fuzzy sets. *Information and Control* 1965, **8,** 338–353
18 Schildt, H. *Artificial Intelligence Using C;* Osborne McGraw-Hill: Berkeley, 1987
19 Zimmermann, H.-J. *Fuzzy Set Theory—and Its Applications;* Kluwer: Boston, 1991
20 Zadeh, L.A. The concept of a linguistic variable and its application to approximate reasoning. *Memorandum ERL-M 411,* Berkeley, October 1973
21 Zachmann, C.-D., Heiden, W., Schlenkrich, M. and

Brickmann, J. Topological analysis of complex molecular surfaces. *J. Comp. Chem.* 1992, **13** (1) 76–84
22 Bellmann, R. and Zadeh, L.A. Abstraction and pattern classification. *J. Math. Anal. Applic.* 1970, **13,** 1–7
23 Heiden, W. *Methoden zur computerunterstützten Untersuchung selektiver Oberflächeneigenschaften von Proteinen.* Thesis, Technische Hochschule Darmstadt, 1993
24 Heiden, W., Goetze, T. and Brickmann, J. *Two new approaches to the quantification of molecular surface topography.* in preparation
25 Bernstein, F., Koetzle, T., Williams, G., Meyer, E. Jr., Brice, M., Rodgers, J., Kennard, O., Shimanouchi, T. and Tasumi, M. The Protein Data Bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol.* 1977, **112,** 535–542
26 Heiden, W., Schlenkrich, M. and Brickmann, J. Triangulation algorithms for the representation of molecular surface properties. *J. Comp.-aided Mol. Design* 1990, **4,** 255–269
27 Connolly, M.L. Solvent-accessible surfaces of proteins and nucleic acids. *Science* 1983, **221,** 709–713
28 Goetze, T. *Algorithmen zur Berechnung, Darstellung und Analyse molekularer Oberflächen.* Thesis, Darmstadt, in preparation
29 Heiden, W., Goetze, T. and Brickmann, J. Fast generation of molecular surfaces from 3D data fields with an enhanced "Marching cube" algorithm. *J. Comp. Chem.* 1993, **14** (2) 246–250
30 Bellido, M.N. and Rullmann, J.A.C. Atomic charge models for polypeptides derived from *ab initio* calculations. *J. Comp. Chem.* 1989, **10** (4) 479–487
31 Ghose, A. and Crippen, G. Atomic physicochemical parameters for three-dimensional structure-directed quantitative structure-activity relationships, I. Partition coefficients as a measure of hydrophobicity. *J. Comp. Chem.* 1986, **7** (4) 565–577
32 Viswanadhan, V.N., Ghose, A.K., Revankar, G.R. and Robins, R.K. Atomic physicochemical parameters for three-dimensional structure-directed quantitative structure-activity relationships, 4. Additional parameters for hydrophobic and dispersive interactions and their application for superposition of certain naturally occurring nucleoside antibiotics. *J. Chem. Inf. Comput. Sci.* 1989, **29,** 163–172
33 Heiden, W., Moeckel, G. and Brickmann, J. A new approach to analysis and display of local lipophilicity/hydrophilicity mapped on molecular surfaces. *J. Comp-aided Mol. Design* 1993, **7,** 503–514
34 Huber, R. and Bode, W. Structural basis of the activation and action of trypsin. *Acc. Chem. Research* 1978, **11,** 114–122
35 Waldherr-Teschner, M., Goetze, T., Heiden, W., Knoblauch, M., Vollhardt, H. and Brickmann, J. "MOLCAD—Computer-aided visualization and manipulation of models in molecular science." In *Second Eurographics Workshop on Visualization in Scientific Computing* (proceedings) Delft, Netherlands, April 22–24, 1991