communications in mathematical
match
and in computer chemistry

ISSN 0340-6253

no. 38, October 1998

MATCDY (38) 53-80 (1998)

Molecular Shape Recognition Using Fuzzy Logic Strategies

Thomas E. Exner[1] and Jürgen Brickmann[1,2]*

[1]Physical Chemistry I, Department of Physical Chemistry,

and [2]Darmstadt Center of Scientific Computing

Darmstadt University of Technology

Petersenstraße 20, D - 64287 Darmstadt, Germany

Fax:0-6151-164298, Tel:0-6151-162198

mail:brick@pc.chemie.tu-darmstadt.de

SUMMARY

An algorithm for the recognition of molecular shape is presented which is based on linguistically formulated shape descriptors. Fuzzy logic strategies are used in order to subdivide molecular surfaces into surface patches and to define scoring functions for the detection of shape complementarity. The strategy partly mimics the "eyeball" technique used by human "searchers". It is demonstrated with three examples (trypsin - PTI, HLE - ovomucoid inhibitor, α-chymotrypsin - ovomucoid inhibitor) that the proposed method can be very effectively used for the prediction of initial guesses for biomolecular complexes, particularly in those cases where the binding sites are not known.

Keywords: molecular recognition, molecular shape, fuzzy set theory, linguistic variables, shape complementarity

* corresponding author

I INTRODUCTION

The specific recognition of molecules by other molecules plays a key role in the biological activity of chemical catalysts and supermolecular chemistry. Biochemical specificity, for example, relies on the selective binding of molecules to a given protein in a well-defined orientation. Specific recognition forms the basis for computer aided strategies within the field of rational drug design. Only a few receptor sites are known today in full detail, even a smaller number of structures of biomolecule-ligand complexes (mostly protein-protein complexes) are available from experimental studies. It is a challenge to researchers from different fields to find model scenarios on one side, and conceptional and computational strategies for an effective simulation of the molecular recognition process on the other.

The general problem to be solved can be easily formulated as follows:

i) Find structures for complexes AB built from two molecular components A and B in solution for which the free energy $\Delta G_{ass}^{AB} = G_{AB}-G_A-G_B$ takes a (relative or absolute) minimum, and

ii) For a given receptor (say A) find those molecules B' for which $\Delta G_{ass}^{AB'} << \Delta G_{ass}^{AB''}$ holds for $\{B'\}\subset\{B\}$, $\{B''\}\subset\{B\}$ and $\{B'\}\cup\{B''\}=\{B\}$, wherein $\{B\}$ is the set of all molecules in consideration as potential docking partners and $\{B'\}$ are those which bind to A very specifically.

The solution of the general problem has two components, the *computational component* (related to the *computation* of free energy differences $\Delta G_{ass}^{AB'}$) and the *classification component* (the definition of the set B').

In principle, the solution of the computational problem is straightforward. Based on a local

force field for both the intramolecular interactions within the molecules A and B as well as all intermolecular interactions between A, B, and the solvent molecules, the ΔG_{ass}^{AB} values can be calculated at least in an approximate way using standard simulation techniques (molecular dynamics (MD) or Monte Carlo (MC)) or minimisation technologies. Unfortunately, the numerical effort for all of these techniques grows with N^{γ}, where N is the number of degrees of freedom which are taken into account within the model scenario and $\gamma \geq 4$. For example, if two molecules A and B are considered as rigid objects with a well-defined potential energy function $V(\mathbf{R}_{AB},\theta_{AB})$ wherein $\mathbf{R}_{AB}$ and θ_{AB} are the relative location and orientation respectively of molecule B with respect to A, the minimum search has to be performed in a six-dimensional space. The search cannot be performed analytically, i.e. one has to calculate the energy $V(\mathbf{R}_{AB},\theta_{AB})$ on a grid and to find the minima by numerical standard techniques. Already for medium size molecules one has to determine about $20^6 \approx 10^8$ potential values in a systematic screening procedure. Standard force fields contain several thousand two-body, three-body and four-body terms, i.e. for the interaction of two medium size proteins one has to calculate 10^5 - 10^6 interactions per grid point. These numbers show that a systematic screening of the configuration space is enormously time consuming even on massive parallel computer architecture. The situation becomes even worse when molecular flexibility is taken into account, i.e. when the two molecules are no longer treated as rigid bodies.

In recent years, considerable effort has been devoted to surmount the computational barrier, i.e. the design of computational procedures for the prediction of stable structures of molecule-ligand complexes (molecular docking).[1-13] In these methods simplified assumptions are made on the intermolecular recognition process. Simple representations of molecular surfaces are used in order to describe shape similarities and complementarities. Local positions for hydrogen bond donor or acceptor atoms, areas of hydrophobic interactions as well as charge

complementarities are taken into account for effective docking procedures. There is no doubt that all of these components contribute to the selectivity in the molecular recognition process. Nevertheless, hydrogen bonds, hydrophobic binding areas and strong electrostatic intermolecular interactions can only take place if surface atoms of the considered molecules come close enough, i.e. if the shapes of the molecules are approximately complementary to each other (at least in the configuration of the final complex AB). We will restrict the discussion here to the surface matching problem, since in a first trial we are mainly interested in finding out whether searching strategies of the human searcher can be transferred effectively into an algorithm. A systematic study of the other components within this framework (most energetic) will presented in a subsequent paper.

An effective search for complementarity of molecular surfaces is obviously a prerequisite for solving the classification problem. This problem deals with the question of how to define for a given molecule A and a reference set B the set B' of possible docking partners. This problem is strongly related to the question of molecular similarity or molecular complementarity of the molecule A to those from the set B' in the region of a receptor (if this is known). We are thus looking at the molecules from the point of view of a "molecular inspector" and trying to discover which may belong to a certain class of possible "keys" to fit some given "lock". The search becomes even be more complicated when the "lock" cannot be specified. 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. A variety of papers have been published in recent years dealing with the question of molecular surface complementarity and shape matching. [6,14-21]

A prominent example for a shape matching procedure is the DOCK program of Kuntz and coworkers [14] which has been one of the earliest docking methods frequently applied in

particular when the binding site of the protein is known. The algorithm calculates a "negative image" of the protein, consisting of a set of overlapping spheres. Groups of overlapping spheres are referred to as clusters. One or more of the clusters containing the largest number of spheres are then selected for docking.[15] Then orientations of the ligand are generated by matching subsets of ligand atoms onto subsets of protein sphere centers. The minimization of these orientations [16] and the docking of flexible ligands using a genetic algorithm is possible [17] in the latest versions of DOCK. The algorithm of Lenhof [6] is also based on a fitness function for evaluating the surface matching of a given conformation, defined as the weighted sum of a geometric and a chemical contact measure. The geometric contact measure describes the size of the contact area of two molecules, and depends on the atom pair building the van-der-Waals contact. Helmer-Citterich and Tramontano [18] introduced an algorithm where the geometric shape of the surfaces, represented by knobs and holes, is described by a 2-dimensional matrix. The search for complementary regions on the surfaces of the protein and the ligand in a given orientation can be reduced to the comparison of sub-matrices. The comparison has to be repeated for all possible relative orientations of the protein and of the ligand and results in a list of docking configurations. These configurations can be evaluated and scored according to a given criterion, e.g. surface area, electrostatic interaction or potential H-bond formation. Within the algorithms of Connolly [19] and Norel et al. [20,21] the surface is represented by critical points, describing knobs and holes. Connolly matches quartets of critical points of the two surfaces. This approach failed since some docking regions do not possess four knob and hole matches. Norel et al. match only two critical points. The surface normal vectors at these critical points serve as the third and fourth criterion for calculating the transformation.

In all of these papers molecular complementarity and matching are based on an "atomistic" point of view, where the word "atomistic" is used as a term to characterize a basic element of

a motif which should be recognized without any context to other "atoms". An example could be a proton donor group given by the local position and a unit vector for a favorite hydrogen bond direction or a surface atom characterized by its position and the radius for the onset of repulsion (van der Waals radius). The recognition motif is then composed out of atomistic elements in a certain geometric arrangement. Similarity and complementarity are quantified using standard (Euclidean) measures.

In this paper we follow in principle a similar strategy, however with one substantial difference: the elements for the composition of a pattern are given in a linguistic way. The motivation for such a scenario is based on the strategies of a human pattern-searcher. There is no doubt that the most effective search procedure - for those instances in which it can be applied - is still the "eyeball" technique used by human "searchers". It is easy to see by inspection that a regularly shaped object (the key) "probably fits" into a rigid surface of complementary shape (the lock).

A variety of papers deals with the question of how to transform the molecular scenario into a representation for which the "eyeball" technique can still be used, i.e. in which human pattern recognition abilities can be successfully applied. New instruments of man-machine communication in molecular science have been developed based on the concept of molecular surfaces. These surfaces are envisioned as the interface between different molecules or between a molecule and its solvent. The visualization and interactive treatment of molecules have been very successfully used in order to study the complementarity problem, which forms the fundament of specific molecular recognition.

However, it is well known that the "eyeball" technique has a variety of limitations. These are significant in all those cases when there is no way of transforming the scenario into a

representation where the human senses are able to recognize data or features. Another limitation is related to the large numbers of objects within a search. If one has to check all molecules stored in a structural database (10^4-10^5 molecular structures) in order to find those molecules which, in principle, can be considered as possible "keys" for a given receptor (set B', see above), the "eyeball" technique will no longer be applicable, simply for pragmatic reasons. Such a search can be done only using the increasing power of modern computational technology.

How can strategies based on human recognition be used for the development of algorithms which can be applied in molecular recognition processes, at least in a pre-selective manner. It will be demonstrated in this paper that fuzzy set theory offers some promise for a solution. Fuzzy set theory has already been successfully applied in different areas of pattern recognition and at different stages of the recognition process.[27] We shall demonstrate in particular that the concept of linguistic variables can prove very useful in molecular similarity search processes and in the study of complimentarity.

This article is organized as follows: In the second section the basic principles of fuzzy set theory and fuzzy logic are reviewed. In section III the procedure for generating the molecular surfaces and calculating the topographical properties of these molecular surfaces is reported. Then the matching algorithm is discussed in detail. The results obtained for three protein-ligand complexes are presented and discussed in the "Results" section. The last section provides some concluding remarks and a short preview of future developments.

II BASIC PRINCIPLES OF FUZZY LOGIC

The concept of fuzzy logic was introduced in 1965 by Zadeh.[29] By now, fuzzy set theory has many applications in many different fields. Because the field is quite complex and under

permanent development, the basics of fuzzy logic can not be discussed fully in this paper. Here, we only present the concepts we actually use. We refer to the literature [27] for more detailed information.

2.1 Fuzzy Sets

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

$$A = \{(x, \mu_A(x)) \mid x \in X\} \qquad (1)$$

Fuzzy logic allows almost all types of functions for membership definition. The crisp set of elements that belong to the fuzzy set A at least to the degree α is called the α-level set:

$$A^{\alpha} = \{x \in X \mid \mu_A(x) \ge \alpha\} \qquad (2)$$

2.2 Linguistic Variables

One of the basic tools for fuzzy logic is based on the concept of linguistic variables (LV's), whose values are not numbers but words of a natural or artificial language. LV's 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 a 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:

$$L = \{A_1, \ldots\ldots, A_n\} \qquad (3)$$

2.3 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. Also the decision itself shall again lead to a crisp value (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 LV 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. The details are discussed as far as necessary in the application section. For further details see Zimmermann.[27]

III FUZZY SHAPE DESCRIPTION BASED ON THE MOLECULAR SURFACE

In this paper the treatment of shape similarity and complementarity is based on linguistic variables describing the shape properties of both docking molecules. To represent the molecular shape the concept of molecular surfaces is used. The preparation of molecular surfaces follows a well-known process. Therefore only a short description is given here.

The atomic coordinates of the complexes and of the unbounded components are taken from the Brookhaven Protein Data Bank (PDB). The receptor (labeled R) and the ligand (labeled L) of the complexes stored in a common PDB file are separated. Only a few entries of the PDB provide

hydrogen position. Therefore the missing atomic coordinates are added using a standard procedure implemented in the program SYBYL (v.6.2, Tripos Associates Inc., St. Louis, MO).

For both molecules R and L a *solvent accessible surface* is generated using the MS algorithm proposed by Connolly.[31] This algorithm is based on the idea of rolling a test sphere along a CPK model of the molecules and produces a set of points representing the molecular surface. We used a probe sphere with radius 1.4 Å for a water molecule throughout the calculations. The set of points of the molecular surface is triangulated for a better graphic representation of the molecular surface.[32] For each surface point a normal vector is calculated.

3.1 Surface Topography

The topographical properties of a solvent accessible surface [31] can be quantified mathematically by the two canonical curvatures at each surface point. Since our interest is focused on binding sites of protein surfaces, we use here the definition of *global curvatures* introduced by Zachmann et al.[33] The global curvatures may be interpreted as average curvatures of the corresponding surface region and are denoted as C_1 and C_2 for the larger and smaller global curvature, respectively. All surface points can then be classified according to the signs of the corresponding curvatures:

1. two positive curvatures, i.e., the surface point belongs to a concave surface region.
2. one positive, one negative curvature, i.e., the surface point belongs to a saddle-type region.
3. two negative curvatures, i.e., the surface point belongs to a convex region.

These curvatures can be mapped to a single quantity describing the degree of convexity, increasing continuously through five basic shape descriptors (0 - 4) plus a flatness value (-1) if

the two curvatures are equal to zero.[34] This quantity is labeled *surface topography index* (T). T is calculated as follows:

$$T = (C_1 - C_2)/C_1 \quad \text{if } C_1 > 0 \text{ and } C_2 > 0, \quad |C_1| \geq |C_2|$$

$$T = 1 + (1 - (C_1 + C_2)/C_1) \quad \text{if } C_1 > 0 \text{ and } C_2 \leq 0, \quad |C_1| > |C_2|$$

$$T = 2 + (C_1 + C_2)/C_2 \quad \text{if } C_1 > 0 \text{ and } C_2 \leq 0, \quad |C_1| \leq |C_2|$$

$$T = 3 + (1 - (C_1 + C_2)/C_2) \quad \text{if } C_1 \leq 0 \text{ and } C_2 < 0, \quad |C_1| \leq |C_2|$$

$$T = -1 \quad \text{if } C_1 = C_2 = 0$$

T can be selected as the basic variable for the definition of a linguistic variable L_T termed *topography*.[34] This linguistic variable is defined by six classes, denoted as *bag, cleft, saddle, ridge, knob and plateau.*

$$L_T = \left\{ \begin{array}{l} (T, \mu_{bag}(T)); \\ (T, \mu_{cleft}(T)); \\ (T, \mu_{saddle}(T)); \\ (T, \mu_{ridge}(T)); \\ (T, \mu_{knob}(T)); \\ \left(\max(|C_1|, |C_2|), \mu_{plateau}(C_1, C_2)\right) \end{array} \right\} \qquad (4)$$

The membership function of the class *plateau* is calculated on the bases of both global curvatures themselves, because T doesn't include any information on the values of the global curvatures. All membership functions of the linguistic variable L_T are shown in Figure 1.

3.2 Segmentation of Triangulated Surfaces

A method to subdivide molecular surfaces into discrete domains has been introduced in an earlier work.[34] This approach using linguistic variables calculates the boundary of a surface domain around a certain reference point. The dissimilarity D_{LV} defined below is used as the criterion for determinating surface domain boundaries.

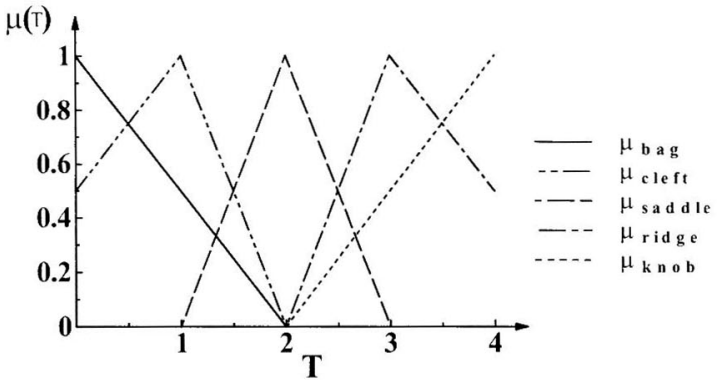

Figure 1: Linguistic variable topography L_T

(a) membership functions of the classes *bag, cleft, saddle, ridge and nob*

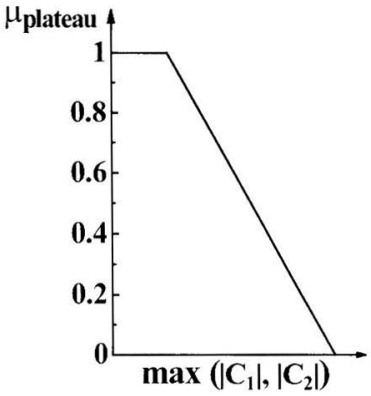

Figure 1: Linguistic variable topography L_T

(b) membership function of the additional class *plateau* derived from the maximum of absolute global curvatures

$$D_{LV}(A,B) = \frac{\sum_{i=1}^{n} w_i \cdot \left|\mu_{A_i}(x) - \mu_{B_i}(x)\right|}{\sum_{i=1}^{n} w_i \cdot \left(\mu_{A_i}(x) + \mu_{B_i}(x)\right)} \qquad (5)$$

with

A, B: linguistic variables of the same type

w_i: weighting factor for class i

n: number of classes in A, B

Sequentially working its way through all triangle node points of the molecular surface, the method achieves complete segmentation of the triangulated surface. The size of the resulting domains was chosen in the range between 200 and 800 triangle node points, corresponding to a surface area between 30 and 150 $Å^2$, respectively (see Heiden et al.[34] for more details).

IV MATCHING OF MOLECULAR SURFACES

4.1 Complementarity of Surface Domains

Within the matching algorithm we present here, each domain is represented by a reference point. This point is defined by the center of gravity of the surface points of the corresponding domain. A surface normal vector is assigned to each domain by calculating the average of the normal vectors of the surface points belonging to this domain. The averages of the surface topography indices and the global curvatures as well as the sizes of the domains are calculated to characterize the shape of the domains. These values are then used to define the degree of shape complementarity of the receptor R and ligand L.

For the matching procedure the surface of R is transformed to a complementary image R'. R' is

built up by domains, which have the same absolute global curvatures as the original domains of R but with the plus and minus signs reversed. The T-value of R' can be calculated as:

$$T(R') = 4 - T(R) \tag{6}$$

with

R: domain of the receptor

R': domain of the negative image of R

The values of the membership functions are calculated for each domain of R'. These values are compared with the membership values calculated for each ligand domain L. The set of domains of R' and L under consideration in one algorithm step are labeled as *main domains*. Every domain having a common border with the main domain is called its *neighbor domain*. If the areas of the main domains of R' and L differ significantly, points of a neighbor domain are added to the smaller domain. The neighbor domain with the best similarity to the smaller main domain (smallest value of the dissimilarity function D_{LV}) is chosen. The decision is made on the basis of two different concepts:

1. In the first trial the dissimilarity function introduced by Heiden et al.[34] (equation 5) is used. The complementarity of a receptor with a ligand domain is defined as the fuzzy complement of the values of the dissimilarity function of the ligand and the complementary image:

$$Comp_{LV}(R,L) = \neg D_{LV}(R',L) \tag{7}$$

with

$$\mu_{Comp_{LV}}(R,L) = 1 - \mu_{D_{LV}}(R',L)$$

R: domain of the receptor

R': domain of the complementary image of R

L: domain of the ligand

$Comp_{LV}$: complementarity of R and L

D_{LV}: value of the dissimilarity function of R' and L

$\neg D_{LV}$: Complement of D_{LV}

2. In a more sophisticated procedure the values of the global curvatures are compared. Two linguistic variables L_{C1} and L_{C2} are defined to quantify the complementarity of the global curvatures of R and L. L_{C1} and L_{C2} are classified by only one fuzzy set and the quotient and difference of the global curvatures are used as definition spaces, respectively:

$$L_{C1}(R,L) = \{\{|C(\alpha)|/|C(\beta)|,\ \mu_{C1}(C(\alpha),C(\beta))\}\} \qquad (8)$$

$$L_{C2}(R,L) = \{\{|C(\alpha)|-|C(\beta)|,\ \mu_{C2}(C(\alpha),C(\beta))\}\} \qquad (9)$$

with

α: convex domain of R or L

β: concave domain of R or L

The membership functions μ_{C1} and μ_{C2} are shown in figures 2 and 3, respectively.

Main domains of the receptor R and the ligand L are identified as complementary if the "weighted average" of $Comp_{LV}$, L_{C1} and L_{C2} is a member of the α-level set with $\alpha = 0.99$. The weighted average is defined by equation 10. L_{C1} and L_{C2} must be calculated for the larger and smaller global curvature of the receptor ligand combination:

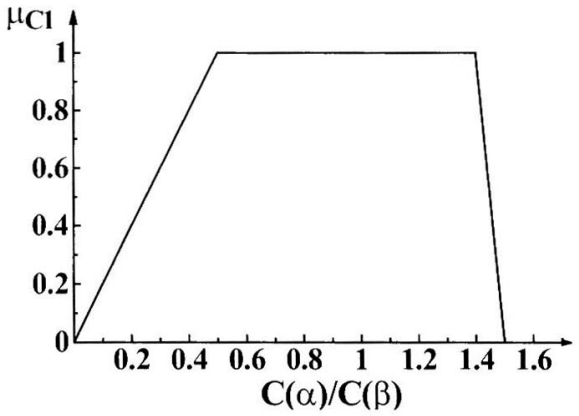

Figure 2: Membership function of the linguistic variable L_{C1}

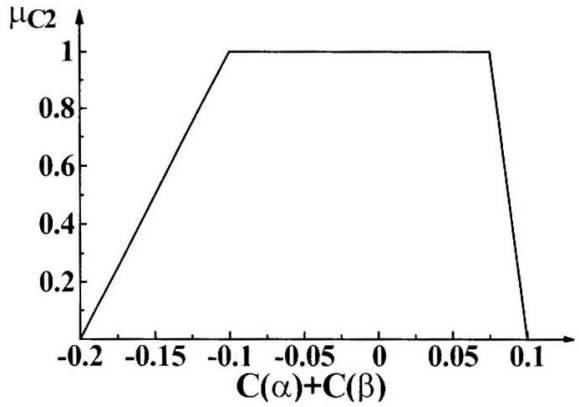

Figure 3: Membership function of the linguistic variable L_{C2}

$$\mu_{aver} = \frac{\sum_{i=1}^{N} w_i \cdot \mu_i}{\sum_{i=1}^{N} w_i} \qquad (10)$$

with

μ_i: value of the membership function of $Comp_{LV}$, L_{C1} and L_{C2}

w_i: weighting factor

N: number of linguistic variables

The complementary domains can be ranked by the value of their weighted average.

4.2 Matching Algorithm

The reference points of the main domain of the receptor and of the ligand are matched, and the ligand is rotated until the normal vectors of the main domains are antiparallel. The position of the centers of gravity of the neighbor domains of R and L are projected on one plane defined by the reference point and the normal vector of the main domain. The angles between the projections of the neighbor domains are calculated to characterize the relative position of the neighbor domains (figure 4).

The ligand is rotated in discrete steps of 45° around an axis defined by the normal vector of the main domain. In this way 8 complex configurations are produced for each pair of main domains. The angles between each projection of the neighbor domains of R and each projection of the neighbor domains of L are calculated for these configurations. For each neighbor domain of R the neighbor domain of L making the smallest angle with the R domain is determined. This is done until each neighbor domain of R is combined with a domain of L. These combinations are called neighbor domain combinations and are compared in the same manner as the main domains. The values of the weighted average (equation 10) are computed for each neighbor domain combination and the average is built over these values. The docking configuration with the largest average is retained as a possible docking structure of the protein ligand complex.

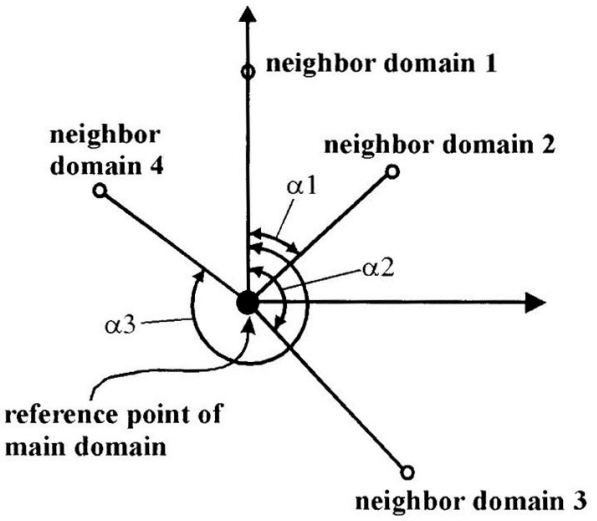

Figure 4: Calculation of the angles between the neighbor domains

V RESULTS

In order to demonstrate that the linguistic formulation of recognition patterns can indeed be used for the molecular shape matching problem we tested the algorithm with the structures of three complexes of serin proteases and their inhibitors. The selected complexes were trypsin - pancreatic trypsin inhibitor (PTI), α-chymotrypsin - ovomucoid inhibitor and human leukocyte elastase - ovomucoid inhibitor. In the case of the trypsin - PTI complex both the separated components of the known complex structure and the components determined separately by x-ray investigations were used. In the other two examples only the structures of the complexes were used.

The structures of the protein ligand complexes were taken from the Brookhaven Protein Data Bank. We used the entries 2PTC for the trypsin - PTI-, 1CHO for the α-chymotrypsin - ovomucoid inhibitor and 1PPF for the HLE - ovomucoid inhibitor-complex. For the unbounded components of the trypsin - PTI-complex the entries 2PTN and 4PTI were used for trypsin and PTI respectively. The missing hydrogen atoms were added by the SYBYL program (v.6.2, Tripos Associates Inc., St. Louis, MO) and the solvent accessible surfaces were generated.

The values of the topographical properties of the surfaces were calculated and a segmentation of the surfaces was carried out as described above. The numbers of domains, which were generated, are shown in table 1. Figure 5 shows five exemplary domains of surface of PTI.

Table 1: Number of domains of the protein-ligand-complexes

Complex	**PDB-file**	**number of domains**
Trypsin	2PTC	86
PTI	2PTC	35
Trypsin	2PTN	109
PTI	4PTI	44
HLE	1PPF	150
Ovomucoid	1PPF	46
Chymotrypsin	1CHO	120
Ovomucoid	1CHO	44

The matching procedure described here selects less than 100 possible docking configurations for each protein ligand pair out of a set of several ten thousand of trials. For each complex the number of docking configurations is listed in table 2. The root-mean-square (rms) derivation of one of these configurations is less than 10 Å from the original x-ray structure. The values of the weighted averages (cf. equation 10) and of the rms derivations are also listed in table 2. In the case of the HLE - ovomucoid inhibitor-complex, the rms value is less than 3 Å. The structures determined by x-ray analysis and by the matching algorithm are drawn in figure 6, showing a good agreement between predicted and experimental structure. In figure 7 the structure of the trypsin – PTI complex and the predicted structure of the unbounded components are compared. In this case, the docking areas are predicted correctly (amino acid Lysin 15 of PTI), but

conformational changes of the ligand must be taken into account by further investigations.

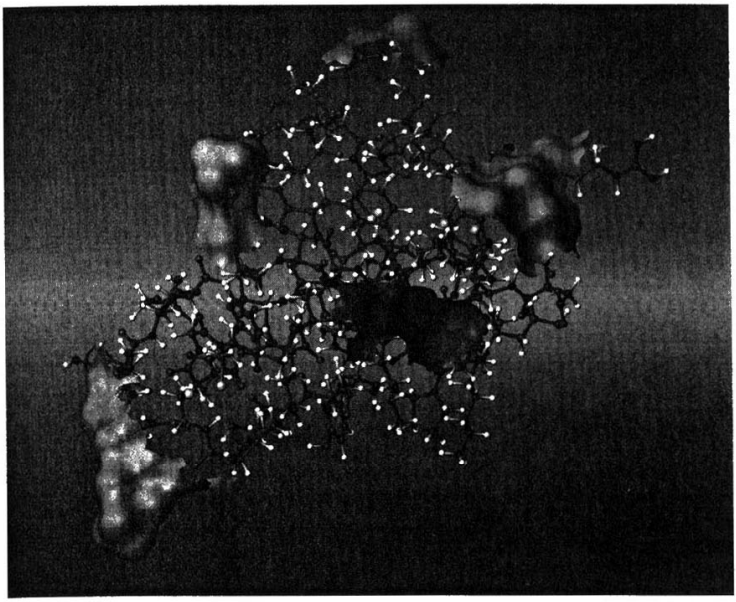

Figure 5: Segmentation of the surface of PTI (solvent accessible surface)
Five representative domains are shown. The domain in the lower left corner corresponds to the binding region.

The configurations suggested by our matching program can be used in a minimization procedure to generate a reliable prediction of the structure of the complex. Within many minimization procedures,[35-37] a large number of random docking positions are produced if the active side of the receptor is not known. These docking positions are then optimized to find the optimal

docking position. The preliminary docking configurations produced by the method we presented here

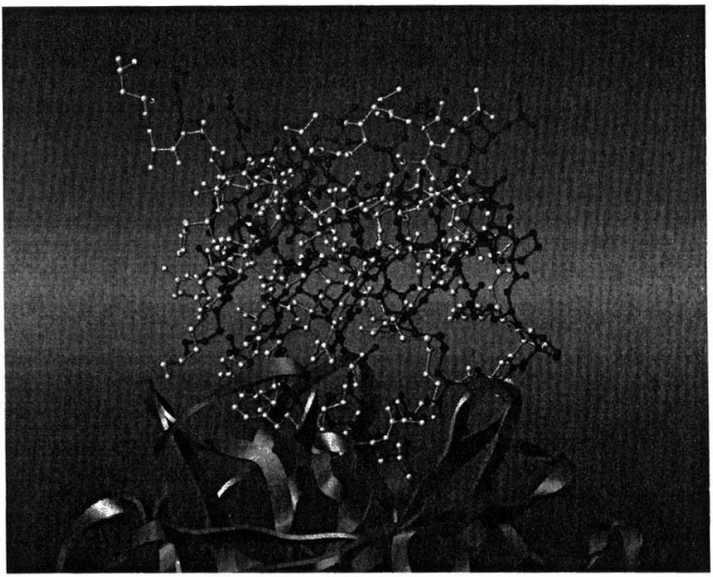

Figure 6: Docking configuration of the HLE - ovomucoid inhibitor-complex
The backbone of HLE from PDB-file 1PPF is shown in ribbon representation
(black) configuration of the ovomucoid inhibitor form PDB file 1PPF
(light gray) predicted configuration of ovomucoid inhibitor

could also be used with these algorithms. The advantage of this procedure is that only a small number of configurations must be tested since the configurations are preoptimized, allowing a quick investigation of the best docking configuration.

Figure 7: Docking configuration of the trypsin - PTI complex

The backbone of trypsin is shown in ribbon representation, PTI as balls and sticks and the binding region of PTI (amino acid Lysin 15) as CPK model.

(a): predicted configuration of the unbounded components from the PDB-files 2PTN and 4PTI

(b): configuration from PDB-file 2PTC

Table 2: Matching results

Complex	**weighted average**	**rms-values**	**number of docking configurations** $\mu_{aver} \geq 0.9$	**number of possible combinations**
Trypsin PTI	0,9947	10,8	28	15050
Trypsin PTI (unbounded)	0,9964		50	23890
HLE Ovomucoid	0,9992	2.5	91	34500
Chymotrypsin Ovomucoid	0,9964	6,5	80	26400

The data in table 3 clearly show that the algorithm is able to produce a small number of possible docking configurations in a very short time. Therefore the algorithm can also be used as a first step for the prediction of protein-ligand configurations in cases where the active side of the receptor is not known.

Table 3: Performance on a Silicon Graphics INDIGO R4400

Complex	CPU time (min:sec)	memory (MB)
Trypsin PTI 1	26:30	14,2
HLE Ovomucoid	56:25	18,2
Chymotrypsin Ovomucoid	41:20	18,6

VI CONCLUSION

In this paper a formalism is presented for the representation and classification of elements of a molecular surface within a scheme based on fuzzy set theory. Surface patches are described therein with the aid of linguistic variables. It is demonstrated that this scenario is well suited for generating similarity and complementarity motifs, and that fuzzy logic treatments can be used in the prediction of structures of biomolecular complexes at least to produce first guesses for a more detailed molecule-molecule matching. A large number of algorithms for effective matching of molecules has been published recently, based both on rigid structures [3,6,8,13,16] and flexible molecular structures.[1,2,5,10,37-40]. We do not consider the present paper as a continuation along the lines of these works. It is an initial trial to transform the strategies of a human searcher (following an "eyeball" procedure) to a computer-driven algorithm, which mimics the pattern recognition ability of the human searcher. The incorporation of properties in addition to shape properties into our linguistically-controlled strategy is underway.

REFERENCES

1. Vieth, M.; Hirst, J.D.; Dominy, B.N.; Daigler, H.; Brooks III, C.L. *J. Comp. Chem.* **1998,** *19,* 1623.
2. Morris, G.M.; Goodsell, D.S.; Halliday, R.S.; Huey, R.; Hart, W.E.; Belew, R.K.; Olson, A.J. *J. Comp. Chem.* **1998,** *19,* 1639.
3. Gabb, H.A.; Jackson, R.M.; Sternberg, M.J.E. *J. Mol. Biol.* **1997,** *272,* 106.
4. Arteca, G.A. *in Rewiews in Computational Chemistry,* **1996.**
5. Lengauer, T.; Rarey, M.; Kramer, B. *J. Comput. -Aided Mol. Des.* **1997,** *11,* 369.
6. Lenhof, H.-P. *First Annual International Conference on Computational Molecular Biology* **1997,** 1.
7. Klebe, G.; Mietzner, T. *J. Comput. -Aided Mol. Des.* **1996,** *8,* 583.
8. Meyer, M.; Wilson, P.; Schomburg, D. *J. Mol. Biol.* **1996,** *264,* 199.
9. Sobolev, V.; Wade, R.C.; Vriend, G.; Edelman, M. *Proteins* **1996,** *25,* 120.
10. Welch, W.; Ruppert, J.; Jain, A.N. *Chem. Biol.* **1996,** *3,* 449.
11. Jones, G.; Willett, P.; Glen, R.C. *J. Mol. Biol.* **1995,** *245,* 43.
12. Norel, R.; Lin, S.L.; Wolfson, H.J.; Nussinov, R. *J. Mol. Biol.* **1995,** *252,* 263.
13. Oshiro, C.M.; Kuntz, I.D. *Proteins* **1998,** *30,* 321.
14. Kuntz, I.D.; Blaney, J.M.; Oatley, S.J.; Langridge, R.; Ferrin, T.E. *J. Mol. Biol.* **1982,** *161,* 269.
15. Meng, E.C.; Shoichet, B.K.; Kuntz, I.D. *J. Comp. Chem.* **1992,** *13,* 505.
16. Gschwend, D.A.; Kuntz, I.D. *J. Comput. -Aided Mol. Des.* **1996,** *10,* 123.
17. Oshiro, C.M.; Kuntz, I.D.; Dixon, J.S. *J. Comput. -Aided Mol. Des.* **1995,** *9,* 113.
18. Helmer-Citterich, M.; Tramontano, A. *J. Mol. Biol.* **1994,** *235,* 1021.
19. Connolly, M.L. *Biopolymers* **1986,** *25,* 1229.
20. Lin, S.L.; Nussinov, R.; Fischer, D.; Wolfson, H.J. *Proteins* **1994,** *18,* 94.
21. Norel, R.; Fischer, D.; Wolfson, H.J.; Nussinov, R. *Protein Eng.* **1994,** *7,* 39.

22. Mudersbach, K.; private communications

23. Brickmann, J.; Goetze, T.; Heiden, W.; Moeckel, G.; Reiling, S.; Vollhardt, H.; and Zachmann, C.-D. in Data Visualization in Molecular Science "Tools for Insight and Innovation"; Bowie, J.E.; Ed.; Addison-Wesley Publishing Company Inc., Reading, Mass., **1995**, 83.

24. Brickmann, J.; Heiden, W.; Vollhardt, H.; and Zachmann,C.-D. in "Proceedings of the 28th Annual Hawaii International Conference on System Sciences"; Hunter, L.; Shriver, B.D.; Eds., IEEE Computer Society press, Los Alamitos, CA., **1995**; 273.

25. Brickmann, J.; Moeckel, G.; Exner, T.; and Keil, M. in Proceedings of the 1997 Chemical Information Conference, **1997**, 107.

26. Brickmann, J.; Exner, T.; Keil, M.; Marhöfer, R.; and Moeckel, G. in "Encyclopedia of Computational Chemistry", P. von Ragué Schleyer, P.R. Schreiner, Eds., J.Wiley, 1998, in press

27. Zimmermann, H.-J. *Fuzzy Set Theory and Its Applications,* Kluwer, Boston, **1991.**

28. Rouvray, D.H.(Ed.) Fuzzy Logic in Chemistry, Academic Press, San Diego, **1997**

29. Zadeh, L.A. *Information and Control* **1965,** *8,* 338.

30. Brookhaven Protein Data Bank; http://www.pdb.bnl.gov

31. Connolly, M.L. *Science* **1983,** *221,* 709.

32. Heiden, W.; Schlenkrich, M.; Brickmann, J. *J. Comput. -Aided Mol. Des.* **1990,** *4,* 255.

33. Zachmann, C.-D.; Heiden, W.; Schlenkrich, M.; Brickmann, J. *J. Comp. Chem.* **1992,** *1,* 76.

34. Heiden, W.; Brickmann, J. *J. Mol. Graphics* **1994,** *12,* 106.

35. Caflisch, A.; Niederer, P.; Anliker, M. *Proteins* **1992,** *13,* 223.

36. Hart, T.N.; Read, R.J. *Proteins* **1992,** *13,* 206.

37. Morris, G.M.; Goodsell, D.S.; Huey, R.; Olson, A.J. *J. Comput. -Aided Mol. Des.* **1996,** *10,* 293.

38. Jones, G.; Willett, P.; Glen, R.C.; Leach, A.R.; Taylor, R. *J. Mol. Biol.* **1997,** *267,* 727.

39. Makino, S.; Kuntz, I.D. *J. Comp. Chem.* **1997,** *18,* 1812.

40. Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. *J. Mol. Biol.* **1996,** *261,* 470.