J Mol Model (2001) 7:370–373
DOI 10.1007/s008940100050

ORIGINAL PAPER

Thomas E. Exner · Jürgen Brickmann

The identification of complementarity of molecular surfaces using fuzzy set theory

Received: 22 December 2000 / Accepted: 26 July 2001 / Published online: 29 September 2001

Abstract Fuzzy logic-based algorithms for the quantitative treatment of complementarity of molecular surfaces are presented. The identification of complementary surface patches can be considered as a first step for the solution of molecular docking problems. Based on these initial guesses, docking structures can be further optimized by standard technologies. In this work a simple downhill simplex method for the optimization is used. The algorithms are applied to various biomolecular complexes. For all these complexes, at least one structure was found to be in very good agreement with the experimental data.

Keywords Complementarity of molecular surfaces · Molecular docking · *Fuzzy set theory* · Surface segmentation

Introduction

Molecular similarity and molecular complementarity play an important role in many branches of molecular science such as computer-aided drug design, biological activity, guest–host interaction in supramolecular chemistry, and nanotechnology. Like many other successfully used concepts in chemistry (hydrophobicity, acidity, etc.) the terms "similarity" and "complementarity" cannot be defined uniquely. This paper deals with molecular complementarity based on the concept of molecular surfaces. Our aim is the manifestation of algorithms for the treatment of molecular docking problems, which are solely based on the surface concept, i.e. which do not need the atomic resolution of the molecular scenario any longer. Such a strategy becomes increasingly important if the number of atoms in the system is very large as in large biomolecular complexes or for nanotechnology problems. We apply our strategy to some biomolecular complexes for which experimental data are available for comparison.

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 enzyme–inhibitor complexes. Most of these algorithms use starting configurations, where the inhibitor is close to the active site of the protein. This is, however, only possible when the receptor site and/or the structure of the inhibitor is known. Predictions of the complex structures with no such information are very time-consuming because of the large number of degrees of freedom, which have to be taken into account. The aim of the present work is to find out first guesses for complex structures of two biomolecules without any information on the possible binding sites. These first guesses will then be used in an optimization procedure resulting in energetically favorable binding geometries. Because of the large quantity of docking algorithms introduced until now, only those methods based on a similar concept will be considered here.

Solutions to the molecular docking problem have used approaches based upon the chemistry and geometry of macromolecules to reduce the solution space of the problem. Nussinov and coworkers [1, 2, 3] use a reduction of molecular surface representation by identifying discrete points on the molecular surfaces of two proteins having specific local shape features, e.g. knob, hole, or saddle-type shapes. The conformational space search is constrained to conformations defined by alignment of shape-congruent points using efficient geometric hashing techniques. [4, 5] For the prediction of flexible molecules the method was extended by representing the molecules by rigid parts, which are connected by rotary joints. [6, 7] In the approach of Hendrix and Kuntz [8] the solid angle, introduced by Connolly [9] and calculated on each point of a solvent accessible surface, [10, 11,

T.E. Exner · J. Brickmann (✉)
Department of Physical Chemistry,
Darmstadt University of Technology, 64287 Darmstadt, Germany
e-mail: brick@pc.chemie.tu-darmstadt.de

J. Brickmann
Darmstadt Center of Scientific Computing, 64287 Darmstadt, Germany

12] is applied to define shape regions as clusters of adjoining points with similar shape features. Molecules are then docked using the DOCK program [13, 14, 15, 16] and implementing the solid angle values of the regions as a shape-based filter.

In this short communication only a resumé of the algorithms is presented and one example for the application is given. The complete strategy and the full formalism will be published elsewhere as a series of papers. [17, 18, 19]

Some basic relations from fuzzy set theory

In contrast to the methods mentioned above, *fuzzy set theory* introduced by Zadeh [20] is used in this work to find out the initial guesses of docking structures. The aim is to transfer strategies of human ability for pattern recognition into mathematical algorithms, which can be applied to the molecular recognition problem. 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 element x in the definition space X and the second μ is 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.

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

One of the basic tools in fuzzy set theory 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 speaking a linguistic variable **L**, classified by n fuzzy sets A_i, can be defined as: [21, 22, 23]

$$\mathbf{L} = \{A_1, ..., A_n\} \tag{2}$$

Fuzzification of molecular properties

In this work, the triangulated Connolly surface is used for the description of the molecular model. [11, 24] On each point defining these surfaces, molecular properties like the *shape index* and the *curvedness* [25, 26] according to the global curvatures, [27] the electrostatic potential, the local lipophilicity, and the ability to build hydrogen bonds are calculated. LVs are defined for all these properties. A similar approach to that of Heiden et al. [28] was used. The LV *topography* according to the shape index is shown as an example in Fig.1.

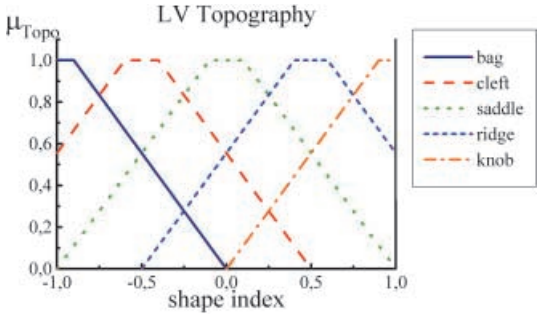

Fig. 1 *Linguistic variable topography* describing the surface shape. The basic variable *shape index* covers the range from –1 for concave through 0 for saddle type to 1 for convex regions. The classes bag, cleft, saddle, ridge, and knob are defined by piecewise linear membership functions

Segmentation of molecular surfaces

Using the LVs of the molecular properties, the similarity of two surface points can be quantified by a similarity measure $S_{LV}(\mathbf{A}, x_A, x_B)$, which is defined as the complement of the *dissimilarity function* $D_{LV}(\mathbf{A}, x_A, x_B)$ proposed by Heiden et al. [28] Therein, two values of the same basic variable, e.g. the shape index, are compared by a weighted sum of the difference of corresponding membership function values. For simplicity the weighting factors for each class of the LV is set to 1.

$$S_{LV}(\mathbf{A}, x_A, x_B) = 1 - D_{LV}(\mathbf{A}, x_A, x_B) \tag{3}$$

with

$$D_{LV}(\mathbf{A}, x_A, x_B) = \frac{\sum_{i=1}^{n} w_i \cdot |\mu_{Ai}(x_A) - \mu_{Ai}(x_B)|}{\sum_{i=1}^{n} w_i \cdot (\mu_{Ai}(x_A) + \mu_{Ai}(x_B))} \tag{4}$$

and

A: LV of corresponding type

$$\mathbf{A} = \{(x, \mu_{Ai}(x)), ..., (x, \mu_{Ai}(x))\}$$

x_A, x_B: values of the basic variable of surface points A and B, respectively

w_i: weighting factors of class i, $0 \le w_i \le 1$

n: number of classes of LV **A**

The molecular surface is subdivided into surface patches (domains), which can be classified according to the linguistic variable, i.e. which can be termed as bag, cleft, saddle, ridge, or knob. In a first step, such points are identified which represent a local maximum of the membership of a certain class. In the second step, neighboring points are added to the domain until the similarity (Eq. 3) becomes smaller than a given threshold value. The details of the procedure are described elsewhere in a full paper. [17] Two typical domains of the methotrexate molecule are shown in Fig.2.

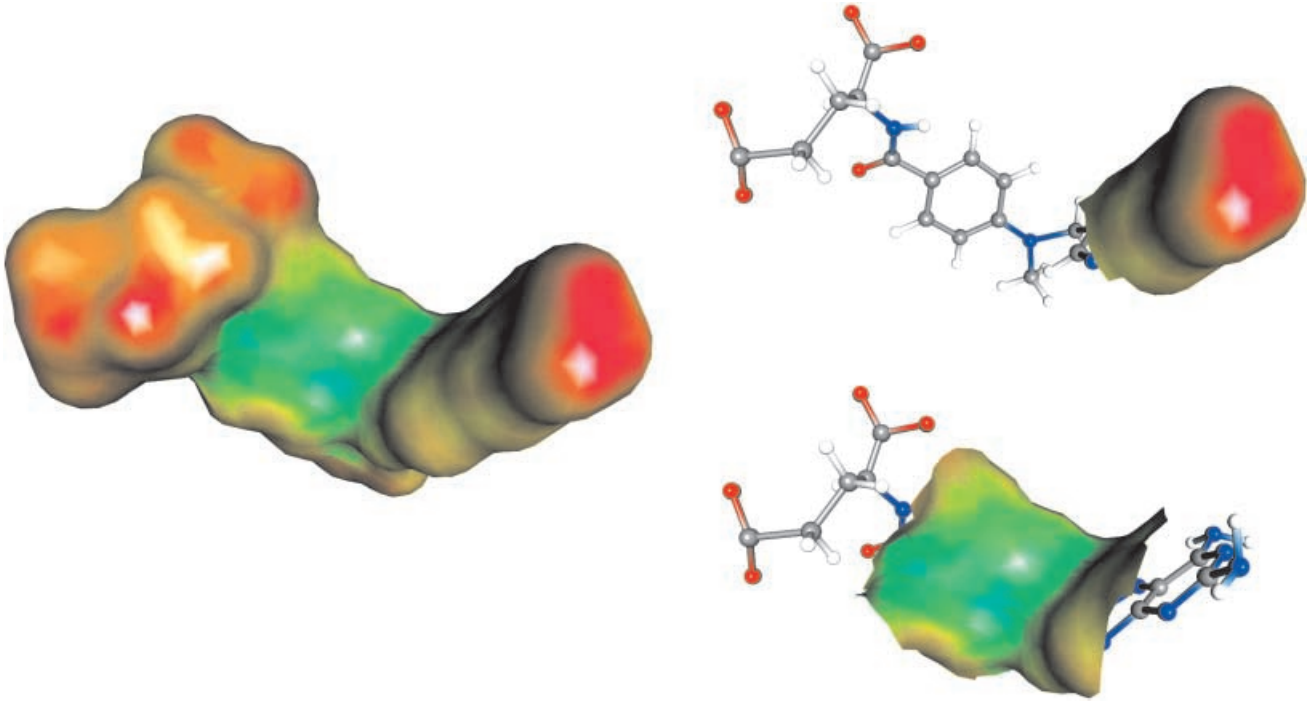

Fig. 2 Two typical surface domains generated with the *linguistic variable topography*. On the left-hand side the hole molecular surface is shown color coded according to the shape index. The *red* color signifies convex, and the *blue* color concave regions. On the upper right-hand side a very convex region is defined as a domain. This domain is characterized as a knob. The domain on the lower side is a saddle-type region

Matching of surface domains

For the matching algorithm described here, the domains are characterized by a central surface point, an average surface normal and the average values of the molecular properties. Domains segmented according to the shape index, the electrostatic potential and the local lipophilicity are used. Complementary domains are identified using the average molecular properties and the dissimilarity function (Eq. 4) as fuzzy complementarity measure. The central points of these complementary domains as well as regions where hydrogen bonds can be built are used as critical points in the geometric hashing algorithm introduced by Schwartz and coworkers. [4, 5]

In the first step of the geometric hashing algorithm, a subset of all central points describing complementary domains of the first molecule that matches a subset of corresponding central points of the second molecule is searched. For these subsets the transformation that best superimposes the complementary critical points is computed using the least-squares-fit algorithm of Ferro and Hermanns. [29] In the second step, the seed matches are extended by searching for additional matching pairs of complementary domains not included in the subsets so far. This procedure is carried out for all possible subsets and the resulting structures are ranked according to the number of matching pairs. All structures with more than ten matching pairs are defined as initial guesses for the complex structure. The consideration of the flexibility of molecules is partially included in the algorithm by the fuzzy representation of the molecular surface properties and the allowance of inexact matches of critical points.

The initial guesses are then clustered to reduce the number of complex structures, which are considered further on. Structures with similar transformations are combined to an average structure. Finally the remaining structures are optimized with the downhill simplex method [30] and the energy function of Gehlhaar et al. [31] as scoring function. Within the downhill simplex algorithm only the rotational and translational degrees of freedom of the hole molecule are considered. The flexibility of the molecules is not treated yet. To identify the most reasonable binding sides, the resulting complex structures are ranked according to the energy. Finally the root mean squares deviations according to the crystallographic structure are calculated.

Results and discussion

We have applied the procedures described above to 35 biochemical complexes taken from the *protein data bank*. [32] In addition to 28 enzyme–inhibitor complexes, three protein dimers and four antigen–antibody complexes were considered. For all these complexes, structures were generated, which could be optimized with reasonable computational effort to the crystallographic structures within 2.5 Å rms-deviation. We did not make any comparison to standard docking procedures up to now because we know that at the present stage our new method cannot compete with these procedures in speed and accuracy. The aim of this extended abstract is to in-

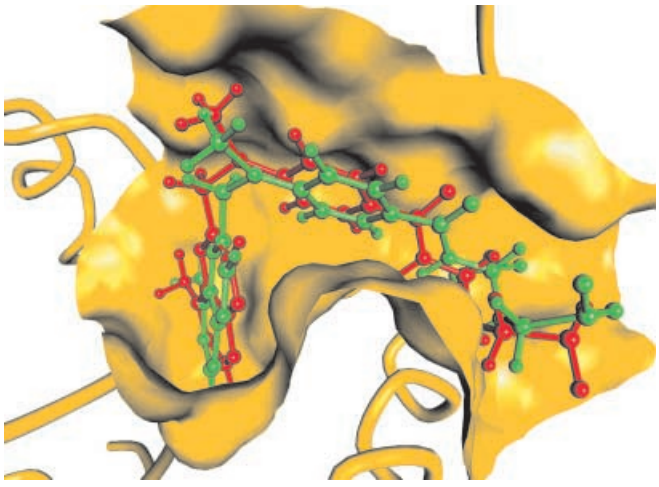

Fig. 3 Complex structure of dihydrofolate reductase with the inhibitor methotrexate as predicted by the fuzzy docking algorithm. The surface of the active site of the enzyme as well as the backbone in ribbon representation are shown in *yellow*. The inhibitor is shown as a balls-and-sticks model. The *green* model is the crystallographic structure taken from the pdb entry 4dfr. The orientation of the inhibitor predicted by the proposed algorithm, shown as a *red* model, differs only slightly with an rms-deviation of 1.056 Å

troduce the use of fuzzy set theory in the molecular docking problem and to show that the results obtained so far are very promising. Thus, we hope that our ideas (with further improvements) can contribute to the progress towards the automatic identification of complex structures from the 3D data.

A detailed analysis of the results and the scaling properties of the method will be presented in Exner et al. [18] In this extended abstract, we demonstrate with the results of the complex of dehydrofolate reductase and the inhibitor methotrexate as an example that the proposed method is able to predict the binding sites and potential structures of biomolecular complexes. The predicted structure is shown in Fig. 3 and differs only slightly from the crystallographic one. The energy calculated for this structure by the simple scoring function was ranked first and, thus, the structure can be easily found out of the other proposed structures. For a further improvement of the results, a more sophisticated optimization procedure should be used and the flexibility of molecules should be handled explicitly during the optimization procedure.

References

1. Lin SL, Nussinov R, Fischer D, Wolfson HJ (1994) Proteins 18:94–101
2. Fischer D, Lin SL, Wolfson HJ, Nussinov R (1995) J Mol Biol 248:459–477
3. Lin SL, Nussinov R (1996) J Mol Graphics 14:78–90
4. Lamdan Y, Schwartz JT, Wolfson HJ (1990) IEEE Trans Robotics Automation 6:578–589
5. Schwartz JT, Sharir M (1987) Int J Robotics Res 6:29–44
6. Sandak B, Nussinov R, Wolfson HJ, (1995) CABIOS 11:87–99
7. Sandak B, Wolfson HJ, Nussinov R (1998) Proteins 32:159–174
8. Hendrix DK, Kuntz ID (1998) Surface solid angle-based site points for molecular docking. In: Altman RB, Dunker AK, Hunter L, Klein TE (eds) Proceedings of the Pacific Symposium on Biocomputing '98. World Scientific, Singapore, pp 317–326
9. Connolly ML (1986) J Mol Graphics 4:3–6
10. Richards FM (1977) Annu Rev Biophys Bioeng 6:151–176
11. Connolly ML (1983) Science 221:709–713
12. Connolly ML (1983) J Appl Crystallogr 16:548–558
13. Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Ferrin TE (1982) J Mol Biol 161:269–288
14. Meng EC, Shoichet BK, Kuntz ID (1992) J Comput Chem 13:505–524
15. Oshiro CM, Kuntz ID, Dixon JS (1995) J Comput-Aided Mol Des 9:113–130
16. Gschwend DA, Kuntz ID (1996) J Comput-Aided Mol Des 10:123–132
17. Exner TE, Keil M, Brickmann J (2001) in preparation
18. Exner TE, Keil M, Brickmann J (2001) in preparation
19. Keil M, Exner TE, Brickmann J (2001) in preparation
20. Zadeh LA (1965) Information Control 8:338–353
21. Zimmermann HJ (1991) Fuzzy set theory and its applications. Kluwer, Boston, Mass.
22. Gottwald S (1993) Fuzzy sets and fuzzy logic: the foundation of application, from a mathematical point of view. Vieweg, Braunschweig
23. Rouvray DH (ed) (1997) Fuzzy logic in chemistry. Academic Press, San Diego, Calif.
24. Heiden W, Schlenkrich M, Brickmann J (1990) J Comput-Aided Mol Des 4:255–269
25. Duncan BS, Olson AJ (1993) Biopolymers 33:219–229
26. Duncan BS, Olson AJ (1993) Biopolymers 33:231–238
27. Zachmann C-D, Heiden W, Schlenkrich M, Brickmann J (1992) J Comput Chem. 1:76–84
28. Heiden W, Brickmann J (1994) J Mol Graphics 12:106–115
29. Ferro DR, Hermans J (1977) Acta Crystallogr, Sect A 33:345–347
30. Nelder JA, Mead R (1965) Comput J 7:308–313
31. Gehlhaar DK, Verkhivker GM, Rejto PA, Sherman CJ, Fogel DB, Fogel LJ, Freer ST (1995) Chem Biol 2:317–324
32. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) Nucleic Acids Res 28:235–232