Journal of Computer-Aided Molecular Design, 11 (1997) 357–368.

J-CAMD 402

Time-efficient flexible superposition of medium-sized molecules

Christian Lemmen* and Thomas Lengauer

German National Research Center for Information Technology (GMD), Institute for Algorithms and Scientific Computing (SCAI),
Schloss Birlinghoven, D-53754 Sankt Augustin, Germany

Received 28 January 1997
Accepted 3 March 1997

Keywords: Structural alignment; Molecular superposition; Molecular similarity; Molecular flexibility; Drug design

Summary

We present an efficient algorithm for the structural alignment of medium-sized organic molecules. The algorithm has been developed for applications in 3D QSAR and in receptor modeling. The method assumes one of the molecules, the *reference ligand*, to be presented in the conformation that it adopts inside the receptor pocket. The second molecule, the *test ligand*, is considered to be flexible, and is assumed to be given in an arbitrary low-energy conformation. Ligand flexibility is modeled by decomposing the test ligand into molecular fragments, such that ring systems are completely contained in a single fragment. Conformations of fragments and torsional angles of single bonds are taken from a small finite set, which depends on the fragment and bond, respectively. The algorithm superimposes a distinguished *base fragment* of the test ligand onto a suitable region of the reference ligand and then attaches the remaining fragments of the test ligand in a step-by-step fashion. During this process, a scoring function is optimized that encompasses bonding terms and terms accounting for steric overlap as well as for similarity of chemical properties of both ligands. The algorithm has been implemented in the FlexS system. To validate the quality of the produced results, we have selected a number of examples for which the mutual superposition of two ligands is experimentally given by the comparison of the binding geometries known from the crystal structures of their corresponding protein–ligand complexes. On more than two-thirds of the test examples the algorithm produces rms deviations of the predicted versus the observed conformation of the test ligand below 1.5 Å. The run time of the algorithm on a single problem instance is a few minutes on a common-day workstation. The overall goal of this research is to drastically reduce run times, while limiting the inaccuracies of the model and the computation to a tolerable level.

Introduction

A major goal in pharmaceutical research is to design molecules that interfere with specific biochemical pathways in living systems. A corresponding area in drug design aims at developing small organic molecules with a high affinity of binding towards a given receptor. If the three-dimensional structure of the receptor is known, standard rational drug design techniques are applicable. However, often structural knowledge of the system under consideration is lacking. In many such cases, we may know only a set of ligands, sometimes together with their measured biological activities towards the receptor. In this case, structure–activity relationship studies (3D QSAR, CoMFA) aim at correlating measured activities with structure-based properties of the ligands [1]. A central aspect in this context is the relative orientation of the ligands in 3D space. This is because 3D QSAR methods compute a correlation that depends on the relative differences between the considered chemical properties in particular areas in space.

If a proper superposition of a set of ligands is available, the relevant chemical features of the ligands can be readily extracted in order to derive a pharmacophore model that, in turn, can be used to search for possible inhibitors in a ligand database. In addition, QSAR studies may provide an estimate of the binding affinity of a novel ligand towards the receptor under consideration. And finally, in some cases, it may be possible to take the negative imprint of the set of superimposed ligands as a crude

*To whom correspondence should be addressed.

description of the binding pocket under consideration. This could be used as a starting point for a receptor modeling study.

In our approach, four major assumptions and simplifications are introduced. The overall reason is that we are content with approximate solutions and aim for a substantial reduction in computing time.

(1) The reference and test ligands are assumed to occupy maximally overlapping areas in space. In addition, it is assumed that in most parts both ligands interact with the same functional groups of the amino acids in the binding pocket.

(2) Only pairs of ligands are considered, i.e. no multiple superpositions of several ligands at a time are attempted directly.

(3) The number of degrees of freedom is reduced to the torsional degrees of freedom of the test ligand (including conformational flexibility of ring systems) in a discrete conformational model. Furthermore, six degrees of freedom of the overall translation and rotation are considered.

(4) All atoms of the reference ligand are kept fixed in space.

The following comments argue these simplifications and point to possible routes of weakening the constraints.

(a) Our application area of interest is the design of potent inhibitors for target receptors. Strong binding requires optimal space-filling of the binding pocket and an extended network of hydrogen bonds with the receptor, satisfying most acceptor and donor functions of the ligand. This partially justifies assumption 1.

(b) The run time of the presented algorithm is small enough to perform several runs with different conformations of the reference ligand, as well as to carry through pairwise comparisons among a larger set of ligands. These runs can be performed independently of each other and, in fact, in parallel. This partially justifies assumptions 2 and 4.

(c) As a postprocessing to the algorithm presented here, existing methods such as, for instance, TORSEAL [2] can be used for refining the superposition. This justifies the use of the approximate discrete model (assumption 3).

(d) As a matter of fact, the conformational space of the reference ligand can be severely restricted in several cases because, usually, the more rigid the molecules, the higher their binding affinity [3]. This partially justifies assumption 4.

Perhaps these assumptions appear rather limiting, especially since a potent reference ligand is a prerequisite for the approach. However, note that the reference ligand can be inferior with respect to various properties, such as bioavailability, cost of synthesis, or toxicity. In these cases, it is desirable to use mutual superpositions of the reference ligand with test ligands that are favorable in these respects, to detect novel candidates with improved properties.

Several superposition methods in the literature incorporate molecular flexibility to some extent. Some of these methods, however, need to be given (parts of) the pharmacophore that displays the commonalities of both ligands [2,4]. Other methods do not require such knowledge; however, often enough they treat both molecules as rigid [2,5]. Methods that handle molecular flexibility without extraneous knowledge of commonalities of both ligands are rare, but are in high demand [6,7].

The active analog approach introduced by Marshall et al. [8] provides the first rigorous treatment of the structural alignment problem; however, it is computationally quite expensive. The DISCO program by Martin et al. [9] can handle only a limited set of conformers and only a set of up to four pharmacophoric points on each molecule is allowed. The same observation applies to the AUTOFIT approach by Kato et al. [10]. The GASP program by Jones et al. is based on genetic algorithms. It can handle the flexibility of even a set of molecules simultaneously; however, it requires in the range of an hour of computing time for a sufficiently large set of GA runs. For a recent review of the existing approaches to molecular superposition, see Ref. 12.

FlexS explicitly takes into account the molecular flexibility of the test ligand and needs no predefined information on the pharmacophore shared by the reference and test ligands. The algorithm requires a few minutes of computing time on a common-day workstation. To estimate the quality of the produced alignments, we try to reproduce the binding geometry of ligands binding to a common protein receptor. As a figure-of-merit, the rms deviation between the crystallographically observed and the computed conformation and orientation of the test ligand is determined. This kind of stringent evaluation has not been published for any of the flexible superposition algorithms mentioned above. This makes the comparison with results from other methods difficult. Taking the sparse data available, the superpositions produced by FlexS seem to be of comparable accuracy; however, the required computing time is at least one order of magnitude shorter with respect to other approaches.

In the next section, we present our model and the superpositioning algorithm. Subsequently, we give results obtained with FlexS by applying our evaluation scheme to a set of ligand pairs. We conclude with a discussion of the strengths and weaknesses of our approach and an outlook to future work.

Methods

Modeling the physicochemical properties

We will use physicochemical properties of the ligands not only for scoring, but also for generating the solutions. Our model considers the flexibility of the test ligand. We

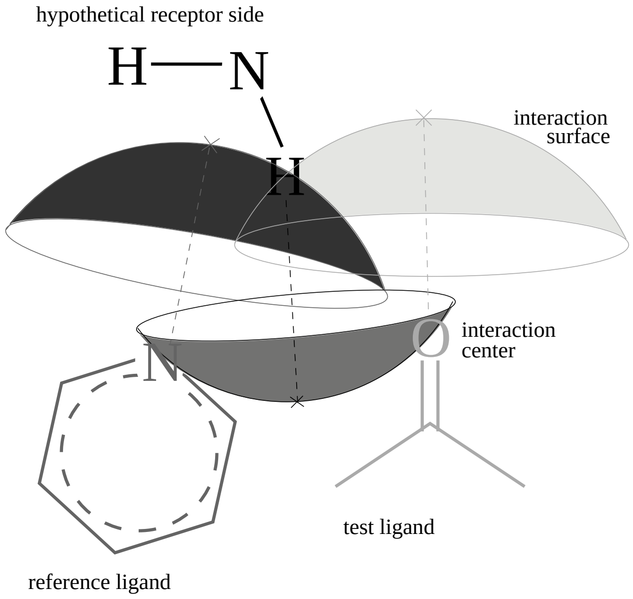

Fig. 1. The paired intermolecular interaction between a carbonyl oxygen and an aromatic nitrogen may be complemented by a donor function of the hypothetical receptor side (located inside the intersection area of the corresponding interaction surfaces), saturating both ligand acceptor functions simultaneously.

use two classes of contributions for scoring: *paired intermolecular interactions* and *overlap volumes*. Intermolecular interactions with a potential receptor atom that are plausible for both ligands are paired and contribute a term to the overall score. Density functions of various kinds, such as electrostatic potential, hydrophobicity, hydrogen-bonding donor and acceptor potentials, and van der Waals volume, are used in order to characterize the physicochemical properties of the ligands. We compute the spatial overlap of these density functions as an additional contribution to the score of the produced alignment.

The above contributions to the scoring function are divided into two groups, called *hard* and *soft* criteria. The hard criteria can be used to generate *placements* and to reject unsatisfactory ones. For example, we use the directional information of certain molecular interactions (e.g. hydrogen bonds) to generate plausible placements of parts of the molecules. The rationale behind this strategy is that interactions like hydrogen bonds provide the dominating force that fixes ligands in position inside the binding pocket of the receptor [13]. A minimum threshold for the *van der Waals overlap volume* serves as a criterion to reject unlikely placements. Keeping assumption 1 in mind, this appears appropriate. In contrast to the hard criteria, the soft criteria are used only for scoring and not for eliminating unlikely solutions. The soft criteria comprise the various overlap volumes and the scoring terms for the paired intermolecular interactions.

Conformational flexibility of the test ligand is modeled by a discrete set of molecular conformations. In particular, we attribute to each acyclic single bond a set of energetically favorable torsional angles by a fragment-based assignment process. This model of discrete conformational flexibility is very similar to the one used in the conformational search procedure MIMUMBA [14]. The QCPE program SCA is used to calculate a discrete set of low-energy conformations for ring systems [15]. The test ligand is decomposed into *fragments* that contain at most one acyclic single bond and always complete ring systems. The exception of this scheme is the so-called *base fragment* (see the section 'The FLEXS algorithm' below), which is allowed to contain a few rotatable bonds.

Paired intermolecular interactions

Our model for intermolecular interactions is based on distributions of nonbonded interactions derived from small-molecule crystal packings [16]. According to the spatial arrangement of these distributions, putative *interaction surfaces* are defined. They amount to sections of a spherical surface surrounding the functional group of interest. To each such *interaction center* a particular *interaction type* is attributed. The assignment of the type and geometry of the interactions to the functional groups of the ligands is done by a fragment-based assignment process again.

The criterion used for pairing potential molecular interactions of the two ligands in the superposition process is that, in principle, the involved functional groups in both ligands can interact with the same group of the receptor. Technically, this implies that the interaction types of the involved functional groups in both ligands

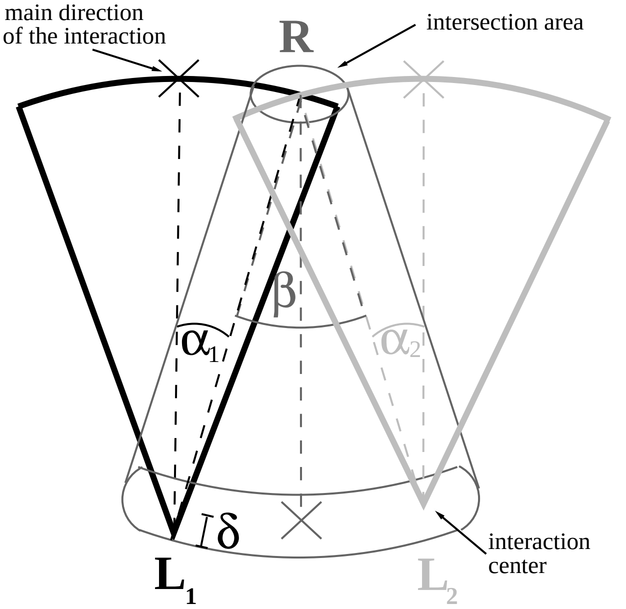

Fig. 2. Paired intermolecular interactions contribute a specific term to the scoring function. This term is weighted by length and angle deviations δ, α_1, α_2 and β from the ideal bond lengths and angles. L_1, L_2 and R denote the functional groups of two ligands and the hypothetical receptor, respectively.

are compatible, the interaction surfaces intersect in at least one point and, in addition, it must be possible to complement both interaction centers, one in each of the ligands, by a single counter group that belongs to the hypothetical receptor side (see Fig. 1).

Note that we superpose neither the interaction centers nor any prespecified points on the interaction surfaces. In the following, sets of paired intermolecular interactions are also called *matches*. To quantify the weight of a match, a scoring function is defined. In this scoring function, each paired intermolecular interaction is represented by a specific term. This term is weighted by penalty functions which account for deviations from ideal bond lengths and bond angles (see Fig. 2). Summing over the contributions of all matches results in the *match score*.

Directional hydrophobic interactions, of which e.g. aromatic ring systems and amide fragments are capable of, are treated separately. The centers of these interactions are required to be in close proximity, and the angle between the normal vectors to the planes, defined by the atoms of these fragments, has to be small (see Fig. 3).

Overlap volumes of different chemical properties

While the intermolecular interactions are the dominating force that immobilizes the ligand inside the binding pocket, they do not necessarily provide the major contributions to the binding affinity towards the receptor. The latter quantity is determined by a precisely matching complementarity of the different physicochemical properties of ligand and receptor [13] and various entropic contributions. The latter quantity correlates approximately with the size of the hydrophobic surface of a ligand, that gets buried upon binding. We assume for two ligands, which achieve a similar binding affinity, that their chemical fingerprints inside the receptor pocket are similar. Therefore, in addition to the match score described above, our scoring scheme considers a common occupancy in space of both ligands. To quantify this overlap, we express the four chemical properties partial charge, hydrophobicity, and hydrogen-bonding donor and acceptor potential in terms of Gaussian functions. The peaks of these functions are localized at positions where we expect these properties on the molecules. The actual spatial occupancy is described by hard spheres with appropriate van der Waals radii for each atom. The overlap volumes of these are taken into account during selection and scoring (see Fig. 4).

Gaussian functions are utilized since they allow for a sufficient smoothness of the chemical descriptors. In addition, for longer distances, these functions have the desirable behavior of a convex potential: it results in an increasing force as the peaks of two Gaussians come closer, and a weaker force at longer distances. We treat steric overlaps in terms of hard spheres instead of describing them by Gaussian potentials for reasons of computational requirements. Here we can reuse data that we computed earlier in the placement procedure during testing superpositions for sufficient steric overlap.

Partial charges must be assigned to the atoms by the user. The atom-based hydrophobicities are determined by a fragment-based assignment process. The corresponding fragment database is taken partially from Viswanadhan et al. [17] and extensions to this list by Klebe and Mietzner (personal communication). The putative hydrogen-bonding sites are computed as described in the previous section. The respective Gaussians are centered at the potential interaction centers. Gaussian functions and their overlap volumes have been applied previously to describe

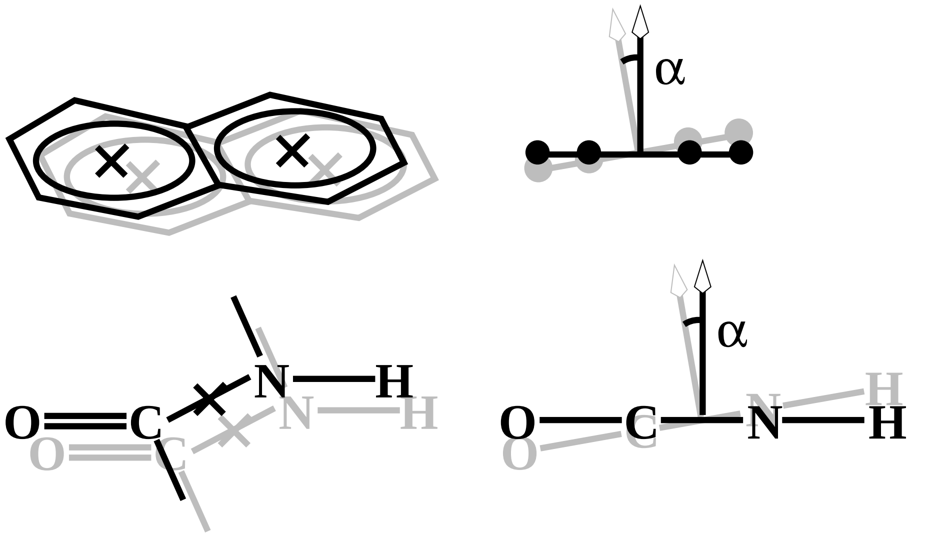

Fig. 3. Aromatic ring centers and amide centers are modeled as directional hydrophobic interaction partners. They are used to force the corresponding interaction centers to be in close proximity and the angle of the normal vectors to the planes (indicated by arrows) of these fragments to be small.

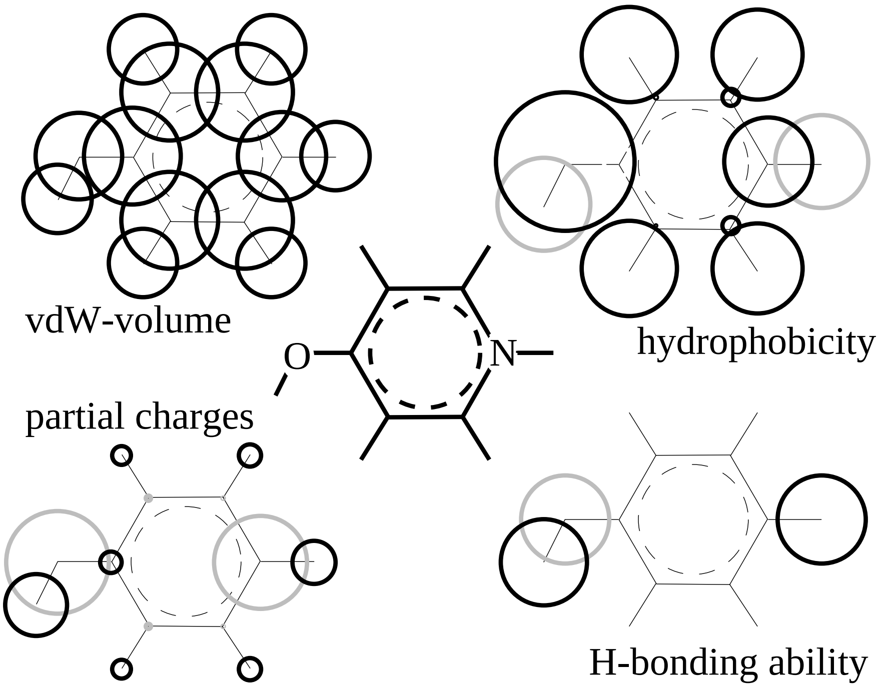

Fig. 4. Description of four chemical properties of the same molecule (center of figure). For measuring steric overlap we use hard spheres with van der Waals radii (top left). The other chemical properties of the molecule are described by Gaussian functions of varying heights. For 2D illustration, the contours shown in this scheme belong to fixed positive (black) and negative (grey) property values.

molecular properties and to score molecular similarity [2,5,18,19].

Input and static data

The input data require the reference ligand in a fixed conformation and the test ligand in an arbitrary low-energy conformation, both with hydrogens and precomputed partial charges. The format of the input files is either the mol or mol2 format of SYBYL [20]. A low-energy conformation is required for the test ligand, because bond lengths and bond angles are taken from the input structure. In addition, this eases the recognition of symmetries in the structure and simplifies the application of the fragment-based assignment process, that is used for labeling single bonds with the torsional angles belonging to them. The fragments stored in the appropriate database have idealized geometries with respect to bond lengths and bond angles. Fragments are successfully assigned to the input structure as long as the deviation of their geometries from the idealized entries in the database is sufficiently small (see Ref. 21 for details). In the case studies described below, we computed the partial charges of the atoms with the Gasteiger method [22], and minimized the structures with the SYBYL force field [20].

The physicochemical background knowledge of FLEXS as well as the parameters to control the program behavior are stored in a set of plain text files that can be easily modified by the user. Table 1 summarizes the essential user-defined parameters and their default values, which have been applied for all test cases described in the Results and Discussion section.

TABLE 1
LIST OF THE MOST IMPORTANT PARAMETERS AND VALUES USED IN THE SUPERPOSITION EXPERIMENTS

No.	Parameter	Estimated meaningful range	Default value
1	Maximum distance between two adjacent interaction points	[1.0,∞] Å	1.26 Å
2	Maximum bond length of a terminal atom, to be merged to the Gaussian representation of the adjacent atom	[0,1.5] Å	1.2 Å
3	Precision of lengths of query triangle edges used as a key in the query triangle list	[0,0.5] Å	0.1 Å
4	Range of lengths [d_{min},d_{max}] of the edges stored in the RL-table	[0,∞] Å	[1.5,10] Å
5	Tolerance for matching a triangle edge during the triangle query process	[0.3,0.7] Å	0.45 Å
6	Maximum rms deviation between two placements stored in the same cluster	[0.5,2.0] Å	1.0 Å
7	Threshold for minimum overlap volume (percentage of the actual test ligand volume)	[50%,100%]	70%
8	Number of solutions carried to the next step of the incremental construction procedure	[250,2000]	500

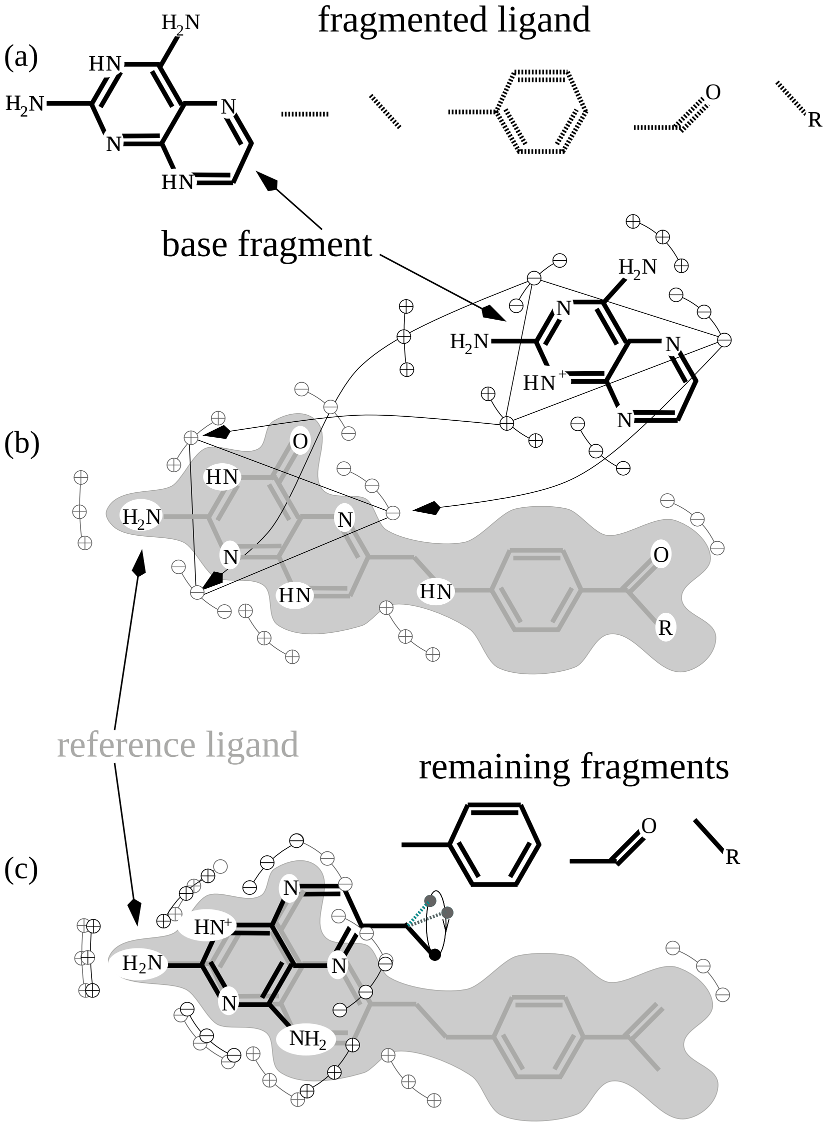

Fig. 5. The three phases of the algorithm: (a) fragmentation and determination of a base fragment, (b) placement of the base fragment, and (c) incremental construction of the entire test ligand.

The FlexS algorithm

The FlexS algorithm is a modification of the flexible docking algorithm in FlexX [23]. The main differences of both methods are the modified modeling of the physicochemical properties, the algorithmic treatment of the first placement phase (see below) and the scoring scheme which is applied. The run time of FlexS is about twice as high as that of FlexX and the results in docking are slightly more accurate. Both facts are balanced by the current parameter setting and result from the computationally higher demands in molecular superposition, especially the different needs in placing a fragment and scoring the alignments. The overall method consists of three phases (see Fig. 5).

In the first phase we select a special fragment, called the base fragment, which serves as an anchor for the subsequent superposition steps. In the second phase we place the base fragment onto the reference ligand. The third phase applies an iterative incremental construction procedure, in order to superimpose the entire test ligand onto the reference ligand. The scorings of the partial solutions mentioned above are heavily used in this phase. In each of the successive iterations (during phase three), we consider a fixed number of the highest scored placements from the previous iteration, and attach to them the next fragment in all its possible conformations. Similar fragment-based construction procedures have been used previously in de novo design [24–26] and in docking [27–29]. However, to our knowledge, they have not yet been applied to the superposition problem.

The basic ideas and the technical realization of the overall method, in the case of the docking problem, have already been presented in full detail [23]. Therefore, we concentrate here on the major modifications that have been introduced, in order to adapt the method to the superposition problem.

Placing the base fragment

As already motivated in the previous section, the main criterion for generating placements for the ligands is based on paired intermolecular interactions. The main criterion for pairing intermolecular interactions, in turn, is the intersection of the corresponding interaction surfaces. We formulate the fragment placement problem as a discrete combinatorial problem by approximating the interaction surfaces by sets of points (currently up to 10 points per interaction surface; see parameter 1 in Table 1). Then we search for nearly congruent triangles of such interaction points in both ligands. Each pair of nearly congruent triangles determines a unique transformation that superimposes one triangle in the first molecule onto the other triangle in the second molecule (minimizing the squared distances of the triangle corners). Through this operation a possible placement of the fragment under consideration is defined. The triangles for the reference ligand are stored in a triangle hash table (*RL-table*) in a preprocessing step. A query to this table, given a triangle Δ on the side of the test ligand (*query triangle*), results in a list of all triangles in the reference ligand that are nearly congruent to Δ.

Now the base placement consists mainly of an enumeration of all query triangles in the test ligand. During the enumeration, each query leads to a list of nearly congruent triangles in the RL-table. As described above, each pair consisting of the query triangle and a triangle in this list defines one placement of the base fragment over the reference ligand. The resulting list of placements is passed to the subsequent steps of the superposition procedure.

Clustering the query triangles

The number of possible triangles in a ligand grows with the cube of the number of interaction points. Therefore, we preprocess the query triangles in a clustering step as follows. At first, we label each query triangle by the types of its corners ($t(p_1)$, $t(p_2)$ and $t(p_3)$, corresponding to the type of interaction points p_1, p_2 and p_3) and the lengths of its sides ($l(p_1,p_2)$, $l(p_2,p_3)$ and $l(p_3,p_1)$, truncated to a user-specified precision; see parameter 3 in Table 1). To make this label unique, the entries of the label $[t(p_i), t(p_j), t(p_k), l(p_i,p_j), l(p_j,p_k), l(p_k,p_i)]$ are ordered such that $t(p_i) \leq t(p_j)$ and $t(p_j) \leq t(p_k)$ hold. If one of these inequalities is not strict, the order which sorts the edge-lengths increasingly is preferred (see Fig. 6 for an illustration).

All query triangles are then compiled in a list (the so-called *TL-list*), which is sorted lexicographically by the triangle labels. The reason for doing so is to obtain contiguous segments of triangles with identical labels (the so-called *L-segments*). What remains to be done is to query each triangle in the TL-list against the RL-table. In fact, we perform such queries only for the first triangle in each L-segment. The triangles which we retrieve from the RL-table are then simply mapped onto each triangle in the L-segment.

Still, in most cases, the number of query triangles is large, and for each query triangle a large list of nearly congruent triangles is retrieved. Normally, we produce between several hundred thousand up to millions of matches of triangles and, consequently, as many possible placements for the base fragment.

In a first step we reduce this enormous number by applying a fast compatibility test that rejects matches for which the additional criterion for pairing interactions (namely, that a hypothetical receptor group can saturate both of the interactions) is missing. Subsequently, van der Waals overlap volumes are computed to filter out unsatisfactory solutions. Since this step is computationally very demanding as well, we use upper and lower bounds and caching techniques to save computation time. Finally, we implemented an efficient on-line procedure in order to cluster similar placements. This step is described in detail in the following section.

On-line clustering of placements

On the one hand, the clustering of placements yields a reduction in the number of placements to a size that can be handled by the subsequent steps of the algorithm. On the other hand, the resulting clusters are able to represent

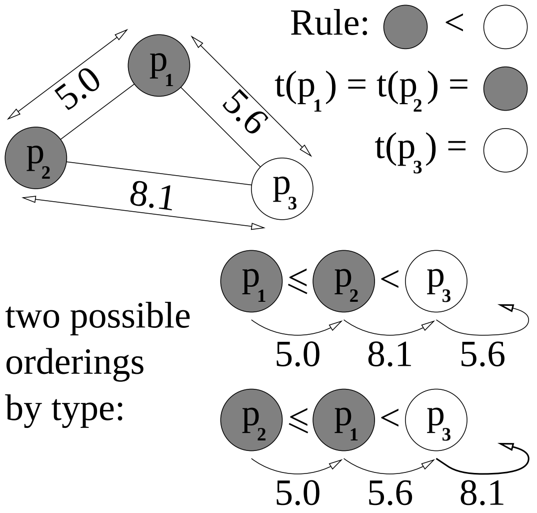

Fig. 6. Given the triangle formed by the interaction points p_1, p_2 and p_3 with interaction types $t(p_1) = t(p_2) \neq t(p_3)$ and edge-lengths $l(p_1,p_2) = 5.0$, $l(p_2,p_3) = 5.6$ and $l(p_3,p_1) = 8.1$. If we follow the rule given in the figure, to order the corners by type, two orderings are possible, because the inequality $t(p_1) \leq t(p_2)$ is not strict. Since $l(p_2,p_3) \leq l(p_3,p_1)$ and $l(p_1,p_3) \geq l(p_3,p_2)$ hold, the first alternative is selected, because it arranges the two lengths with variable position in the label (5.6 and 8.1 in the figure) in increasing order. The resulting label is $[t(p_2), t(p_1), t(p_3), l(p_2,p_1), l(p_1,p_3), l(p_3,p_1)]$.

TABLE 2
RESULTS OF THE SUPERPOSITION EXPERIMENTS

No.	Receptor	Reference ligand	Test ligand	Run time (min:s)				Accuracy (Å)			Partial placement	
				(a)	(b)	(c)	(d)	(a)	(b)	(c)		
1	Carboxypeptidase A	7cpa	1cbx	1	1:47	29	2:17	0.80	0.96	0.96	100	3
2		7cpa	2ctc	1	1:39	17	1:57	0.51	0.79	0.79	100	3
3		7cpa	3cpa	1	35	34	1:10	0.92	0.94	0.80	100	5
4		7cpa	6cpa	2	4:42	2:34	7:18	0.41	0.71	0.71	100	14
5	Endothiapepsin	2er7	5er2	9	7	12:48	13:04	0.99	1.74	1.49	88	35
6		2er7	5er1	5	7	4:06	4:18	0.99	2.13	1.11	42	19
7		2er7	4er1	7	7	5:48	6:02	0.95	1.90	1.26	60	25
8	DHFR	1dhf	4dfr	1	4:35	1:33	6:09	0.40	0.74	0.71	100	11
9	G-Phosphorylase B	4gpb	3gpb	0	19:24	1:05	20:29	0.26	0.45	0.45	100	7
10		4gpb	5gpb	0	21:02	40	21:42	0.53	0.62	0.62	100	7
11	Human rhinovirus	2r06	2r04	1	5	3:07	3:13	0.61	1.11	1.07	100	10
12		2r06	2r07	1	11	2:32	2:44	0.41	0.58	0.59	100	8
13		2r06	2rs5	1	5	2:36	2:42	0.65	0.78	0.77	100	8
14		2rm2	2rr1	1	5	3:22	3:28	0.66	0.62	0.53	100	10
15		2rm2	2rs1	1	6	3:15	3:22	0.52	0.50	0.50	100	10
16		2rm2	2rs3	1	5	3:25	3:31	0.50	0.52	0.50	100	11
17	Streptavidin	1srf	1srg	0	1	1	0:02	0.64	0.76	0.76	100	3
18		1srf	1srh	1	1	17	0:19	0.36	0.67	0.67	100	5
19		1srf	1sri	0	1	1	0:02	0.78	0.84	0.85	100	3
20		1srf	1srj	1	1	1	0:03	0.74	1.15	1.11	100	3
21	Thermolysin	5tmn	1tlp	2	2	3:25	3:29	0.79	1.48	1.48	100	15
22		5tmn	1tmn	1	2	1:16	1:19	0.46	1.72	1.15	35	14
23		5tmn	2tmn	1	6:20	49	7:10	0.49	0.69	0.70	100	5
24		5tmn	3tmn	1	2:14	48	3:03	0.66	0.95	0.96	100	6
25		5tmn	4tln	1	3:13	32	3:46	0.75	1.32	1.32	100	4
26		5tmn	4tmn	2	2	2:24	2:28	0.71	1.91	1.39	75	16
27		5tmn	5tln	1	2	1:02	1:05	1.92	1.92	1.92	0	12
28		5tmn	cbz	1	2:20	58	3:19	0.85	6.58	1.08	28	7
29		5tmn	ppp	1	2:16	39	2:56	0.77	1.66	0.98	57	7
30		5tmn	rthior	1	2	19	0:22	0.65	1.15	1.14	100	7
31		5tmn	thior	1	2	18	0:21	1.53	3.01	1.45	71	7
32	α-Thrombin	1dwd	1dwc	2	6:58	2:41	9:41	0.51	1.54	1.33	81	11
33		1dwd	3tapap	1	1:15	53	2:09	0.43	0.83	0.82	100	8
34		1dwd	4tapap	2	1:15	52	2:09	1.28	2.01	1.46	87	8
35	β-Trypsin	1tpp	3ptb	0	42	0	0:42	0.46	0.45	0.46	100	1

Columns 1–4 identify the test case. Ligands are indicated by the PDB code of the complex containing them. The run times were measured on a SUN UltraSparc workstation with 64 MB RAM; they are detailed as (a) I/O and preprocessing, (b) base placement, (c) incremental construction, and (d) total. The accuracy is given for (a) the base fragment and (b) the whole test ligand. The last three columns indicate the number of fragments attached in the construction procedure while preserving an accuracy of at most 1.5 Å rms deviation. The next to last column contains the percentage of fragments that have been attached up to this point, the last column contains the total number of fragments and the column 'Accuracy (c)' shows the accuracy at this stage of the process.

matches with more than three pairs of intermolecular interactions. This is the case if a placement of the base fragment realizes more than three paired intermolecular interactions. During base placement only triplets of paired intermolecular interactions are recognized. For similar placements (belonging to the same cluster) these triplets are merged and thus extended to possibly larger matches.

Since the number of placements before clustering can be as large as a hundred thousand, expensive methods, such as complete-linkage clustering, are not applicable here. The on-line clustering that we use is rendered efficiently in the following way.

The first computed placement p_0 is taken as a reference from now on. For every newly generated placement p_{new}, the rms deviation d_{new} from p_0 is determined in a first step. Then, it has to be decided whether there is a cluster represented by a placement p that is similar to p_{new}. As an indication of similarity, the rms distance of p and p_{new} has to be below a given threshold δ (see parameter 6 in Table 1). If this condition is fulfilled, we merge p and p_{new}. More precisely, we include the pairs of intermolecular interactions from p_{new} to the list of interaction pairs realized by p and we discard p_{new}. If no appropriate cluster is found, p_{new} is retained as the representative of a new cluster.

Because the rms distance condition fulfills the triangle inequality, the search for p can be restricted to clusters that have an rms distance d to the reference p_0 which falls in the range of $[d_{new} - \delta, d_{new} + \delta]$. In order to take advantage of this property, we sort all placements by their rms distance d to p_0. The sorted list is maintained as a leaf-chained search tree. In this tree, placements within the range $[d_{new} - \delta, d_{new} + \delta]$ form a contiguous segment inside the leaf-chain. The run time of the on-line clustering is proportional to $m.\max(n/\beta, \log n)$, where m is the number of solutions, n is the number of clusters generated, and $\beta = d_{max}/2\delta$ is the number of bins induced by a maximum distance d_{max} (see parameter 4 in Table 1) and the rms threshold δ.

The on-line clustering completes the base placement phase. The list of placements (cluster representatives) resulting from this step is passed to the subsequent third phase of the algorithm, in which the remaining fragments are attached in an iterative incremental manner.

Scoring placements

In addition to the match score, the scoring function for placements, as described above, includes terms for the overlap volumes of the Gaussian functions, describing the chemical properties of the molecules. Since the number of terms grows with the square of the number of Gaussians (i.e. approximately the number of atoms) and this evaluation has to be done for every partial solution, scoring is a time-consuming step in the superposition procedure. Therefore, we reduce the number of Gaussians used to describe the chemical properties in two ways. On the one hand, the fragment library can be designed in order to model the chemical properties at the desired level of detail. For example, the user can decide to describe an aromatic ring system by only three Gaussians. This would still define the plane of the ring system and may be sufficient to approximately describe the properties of this fragment. On the other hand, we developed an automatic merging procedure for Gaussians representing terminal atoms. It is applied whenever the distance to the adjacent atoms falls below a given threshold (parameter 2 in Table 1). In such a case, the corresponding Gaussians are replaced by one single Gaussian located at the center of gravity of the original Gaussians, and weighted by the sum of the volumes of these Gaussians. Currently, the user-defined threshold is set in a way that Gaussians for hydrogen atoms are merged with the Gaussians of adjacent atoms. This decreases the number of Gaussians by approximately a factor of 2. Accordingly, the run time necessary to determine the Gaussian overlap volumes is reduced by a factor of 4.

Results and Discussion

Our tests compare the alignment produced with FLEXS with the structural alignment extracted from the respective X-ray data. The quality of our results is measured in terms of the rms deviation of the predicted from the measured orientation and conformation of the test ligand. To compute the reference alignment, we take the protein–ligand complexes of both ligands and perform an rms fit of the C^{α} positions of the protein atoms. There is some uncertainty inherent in this superpositioning, because it does not appropriately consider any conformational changes resulting from a ligand-induced fit. The reduced resolution of protein structure determinations is another source of error in spatial position. These limitations should be kept in mind while inspecting the figures given below. The reference ligand is kept in the conformation and orientation produced by the above procedure. The test ligand is transferred to an arbitrary energetically minimized conformation.

TABLE 3
RESULT STATISTICS OF THE SUPERPOSITION EXPERIMENTS

Parameter	x = 1.0 Å	x = 1.5 Å	x = 2.0 Å
(a) Number of examples for which the base fragment has been placed with below x Å rms deviation	109 (82%)	126 (95%)	131 (98%)
(b) Mean rms deviation for the examples of (a)	0.64 Å	0.72 Å	0.76 Å
(c) Number of examples where the whole test ligand has been placed with below x Å rms deviation	64 (49%)	88 (67%)	107 (82%)
(d) Mean rms deviation for the examples of (c)	0.74 Å	0.87 Å	1.03 Å
(e) Mean number of fragments attached during the incremental buildup procedure, while preserving an rms deviation below x Å	5.1 (69%)	6.3 (85%)	6.9 (94%)
(f) Mean rms deviation at the respective stage of the construction process and for the examples of (e)	0.82 Å	0.97 Å	0.90 Å
(g) Mean number of fragments		7.4	
(h) Mean run time for the preprocessing step		0:01	
(i) Mean run time for the base placement		1:58	
(j) Mean run time for the incremental buildup		1:22	
(k) Mean run time total		3:21	

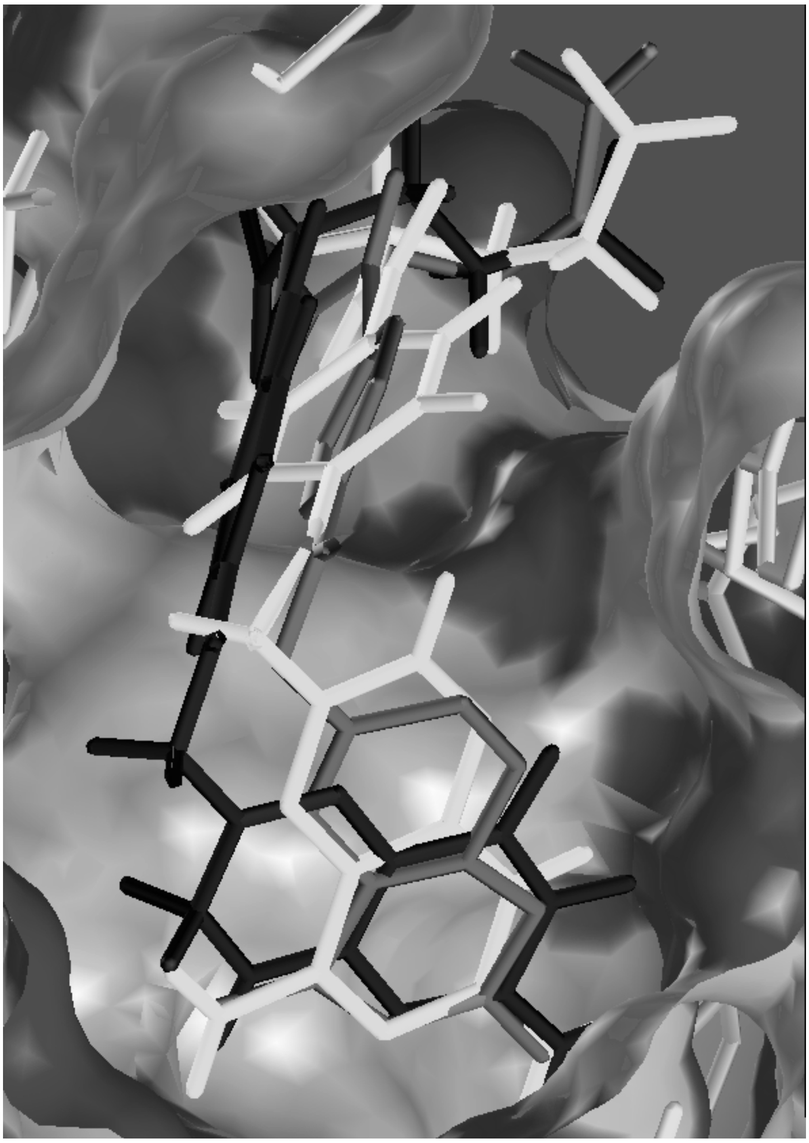

Fig. 7. Superposition of methotrexate as observed (grey) and computed by FLEXS (white) with dihydrofolate (black) inside the binding pocket of dihydrofolate reductase. (Picture created with SYBYL [20].)

Result statistics

Currently, our test set contains ligands of nine proteins: carboxypeptidase A, endothiapepsin, dihydrofolate reductase, glycogen phosphorylase B, human rhinovirus, streptavidin, thermolysin, α-thrombin and β-trypsin. In total, we have 45 ligands and 133 ligand pairs fulfilling requirement 1, stated in the Introduction.

The mean run time over all test cases is below 4 min per instance. The run time is spent to about equal parts on the base placement and on the complex construction. Only a minor fraction of the run time is spent on I/O and preprocessing. The mean rms deviation for the 107 examples, where we were able to reproduce the crystal solution with below 2 Å, is about 1 Å.

Table 2 presents the results of a subset of examples. This subset contains one reference ligand for every protein and all possible pairings with test ligands of this receptor. The statistics obtained for the entire test set, which are presented in Table 3, are almost the same for this subset. Supplementary material including the full-length table and 3D models of our results can be found on our webpage [*http://www.gmd.de/SCAI/alg/reliwe/reliwe_home.html*].

In 95% of the test cases we find a reasonable placement (by *reasonable* we denote a reproduction of experiment below 1.5 Å rms deviation) for the base fragment. For 82% of the examples we even get an accuracy of less than 1.0 Å rms deviation (see Table 3 for detailed test statistics). As the construction of the test ligand proceeds via attaching additional fragments in the third phase of the procedure, the reproduction of its relative orientation and conformation becomes increasingly difficult. Nevertheless, we reproduce the alignment found crystallographically with an rms deviation below 1.5 Å for 67% of the ligand pairs. For 49% of the examples, we were even able to predict the alignment with an accuracy of less than 1.0 Å rms deviation. These results show that, at least for medium-sized molecules, we suggest useful superpositions. For example, the mean number of fragments attached in the incremental construction procedure (preserving a reasonable position for the ligand) is 6.3 over all test cases and the mean rms deviation from the reference, over all these alignments, is about 1.0 Å. Taking the mean of 6.3 attached fragments, for some of the cases the conformational space then comprises already up to six million energetically favorable conformations, even with the constraints of our discrete conformational model. Sampling such a large space requires search techniques which go beyond a few random probes and perform a more global search as presented here.

Dihydrofolate reductase

Finally, we would like to comment on a representative test example, namely the superposition of dihydrofolate (PDB code: 1dhf) with methotrexate (PDB code: 4dfr), both binding to dihydrofolate reductase (see Fig. 7). Many of the existing superposition tools have been applied to this test case. It can be stated that, for this example, the correct superposition cannot be found by considering the bonding skeletons or steric aspects of the ligands alone, but it is largely determined by the involved intermolecular interactions [3].

The natural substrate dihydrofolate, which is the ionic form of folic acid, is taken as the reference ligand. The inhibitor methotrexate serves as the test ligand. It contains 10 rotatable bonds and two rigid ring systems. Therefore, it is decomposed into 11 fragments. Our discrete conformational model allows for more than 10^6 low-energy conformations of this molecule.

The heterocyclic portion of methotrexate serves as the base fragment. It is able to form 10 intermolecular interactions with its three hydrogen-bonding donor functions, five hydrogen-bonding acceptor functions and two centers of directional hydrophobic interactions. During base placement, 6143 triangles of interaction points are generated. With the query triangle clustering step, this number is reduced to 5580 triangles actually used for querying against the RL-table. About 271 000 pairings of test ligand triangles with reference ligand triangles and, thus, as many placements for the base fragment are determined during query triangle enumeration. The compatibility test for paired intermolecular interactions reduces this number to about 196 000 and, by applying the minimum threshold for the van der Waals overlap volume, this figure further reduces to about 77 000. The final clustering step of the base placement phase results in 1369 placements for the base fragment. The base placement is performed in 275 s and the lowest rms deviation from the alignment extracted from the X-ray data is 0.4 Å.

In the subsequent complex construction phase, the remaining 10 fragments are attached step by step. This final phase of the algorithm is performed in 93 s and results in 82 possible placements. Among these, the best solution by rms deviates from the reference alignment by 0.74 Å.

In total, the run time needed to carry out this example is about 6 min. Each of the paired intermolecular interactions realized by the favorable placement is found also in the crystallographically given alignment. The exception from this is the terminal carboxylate group which points to the solvent.

Conclusions and outlook

As our results show, the FLEXS algorithm is capable of quickly superimposing medium-sized molecules with reasonable accuracy. Currently, the major drawback of the algorithm is the inaccuracy of the ranking of solutions which prevents us from producing better results for large ligands with many conformational degrees of freedom. We intend to improve the ranking that our scoring function produces by carefully calibrating the coefficients of the various contributions to it. Another shortcoming of our method is the presently required rigidity of the reference ligand. We plan to incorporate some degree of flexibility for the reference ligand in our method.

Acknowledgements

We thank Matthias Rarey and Bernd Kramer for their help in modifying the docking tool FLEXX. We are grateful to our collaborating partners in the RELIWE project, especially G. Klebe and T. Mietzner from BASF, main laboratory, for preparing parts of the input and static data and for the many stimulating discussions that introduced us to the concepts of molecular similarity. The RELIWE project is partially funded by the German Federal Ministry for Education, Science, Research and Technology (BMBF) under Grant 01 IB 302 A.

References

1 Kubinyi, H. (Ed.) 3D QSAR in Drug Design: Theory, Methods and Applications, ESCOM, Leiden, The Netherlands, 1993.

2 Klebe, G., Mietzner, T. and Weber, F., J. Comput.-Aided Mol. Design, 8 (1994) 751.

3 Böhm, H.-J. and Klebe, G., Angew. Chem., Int. Ed. Engl., 35 (1996) 2588.

4 McMartin, C. and Bohacek, R.S., J. Comput.-Aided Mol. Design, 9 (1995) 237.

5 Kearsley, S.K. and Smith, G.M., Tetrahedron Comput. Methodol., 3 (1990) 615.

6 Klebe, G., Perspect. Drug Discov. Design, 3 (1996) 85.

7 Leach, A.R., In Dean, P.M. (Ed.) Molecular Similarity in Drug Design, Blackie Academic and Professional, London, U.K., 1995, pp. 57–88.

8 Marshall, G.R., Barry, C.D., Bosshard, H.D., Dammkoehler, R.D. and Dunn, D.A., In Olson, E.C. and Christoffersen, R.E. (Eds.) Computer-Assisted Drug Design, Vol. 112, American Chemical Society, Washington, DC, U.S.A., 1979, pp. 205–222.

9 Martin, Y.C., Bures, M.G., Danaher, E.A., DeLazzer, J., Lico, I. and Pavlik, P.A., J. Comput.-Aided Mol. Design, 7 (1993) 83.

10 Kato, Y., Inoue, A., Yamada, M., Tomioka, N. and Itai, A., J. Comput.-Aided Mol. Design, 6 (1992) 475.

11 Jones, G., Willett, P. and Glen, R.C., J. Comput.-Aided Mol. Design, 9 (1995) 532.

12 Klebe, G., In Kubinyi, H. (Ed.) 3D QSAR in Drug Design: Theory, Methods and Applications, ESCOM, Leiden, The Netherlands, 1993, pp. 173–199.

13 Dean, P.M., In Dean, P.M. (Ed.) Molecular Similarity in Drug Design, Blackie Academic and Professional, London, U.K., 1995, pp. 1–23.

14 Klebe, G. and Mietzner, T., J. Comput.-Aided Mol. Design, 8 (1994) 583.

15 Hoflack, J. and De Clercq, P.J., Tetrahedron, 44 (1988) 6667.

16 Klebe, G., J. Mol. Biol., 237 (1994) 221.

17 Viswanadhan, V.N., Ghose, A.K., Revankar, G.R. and Robins, R.K., J. Chem. Inf. Comput. Sci., 29 (1989) 163.

18 Good, A.C., Hodgkin, E.E. and Richards, W.G., J. Chem. Inf. Comput. Sci., 32 (1992) 188.

19 Grant, J.A., Gallardo, M.A. and Pickup, B.T., J. Comput. Chem., 17 (1996) 1653.

20 SYBYL, Tripos Associates Inc., St. Louis, MO, U.S.A., 1994.

21 Rarey, M., Wefing, S. and Lengauer, T., J. Comput.-Aided Mol. Design, 10 (1996) 41.

22 Gasteiger, J. and Marsili, M., Tetrahedron, 36 (1980) 3219.

23 Rarey, M., Kramer, B., Lengauer, T. and Klebe, G., J. Mol. Biol., 261 (1996) 470.

24 Rotstein, S.H. and Murcko, M.A., J. Comput.-Aided Mol. Design, 7 (1993) 23.

25 Böhm, H.-J., J. Comput.-Aided Mol. Design, 6 (1992) 61.

26 Moon, J.B. and Howe, W.J., Proteins, 11 (1991) 314.

27 Welch, W., Ruppert, J. and Jain, A.N., Chem. Biol., 3 (1996) 449.

28 Leach, A.R. and Kuntz, I.D., J. Comput. Chem., 13 (1992) 730.

29 DesJarlais, R.L., Sheridan, R.P., Dixon, J.S., Kuntz, I.D. and Venkataraghavan, R., J. Med. Chem., 29 (1986) 2149.