FLEXS: A Method for Fast Flexible Ligand Superposition

Christian Lemmen,*,† Thomas Lengauer,† and Gerhard Klebe‡

Institute for Algorithms and Scientific Computing (SCAI), German National Research Center for Information Technology (GMD), Schloβ Birlinghoven, 53754 Sankt Augustin, Germany, and Department of Pharmaceutical Chemistry, Philipps-University of Marburg, Marbacher Weg 6, 35032 Marburg, Germany

Received May 11, 1998

If no structural information about a particular target protein is available, methods of rational drug design try to superimpose putative ligands with a given reference, e.g., an endogenous ligand. The goal of such structural alignments is, on the one hand, to approximate the binding geometry and, on the other hand, to provide a relative ranking of the ligands with respect to their similarity. An accurate superposition is the prerequisite of subsequent exploitation of ligand data by either 3D QSAR analyses, pharmacophore hypotheses, or receptor modeling. We present the automatic method FLEXS for structurally superimposing pairs of ligands, approximating their putative binding site geometry. One of the ligands is treated as flexible, while the other one, used as a reference, is kept rigid. FLEXS is an incremental construction procedure. The molecules to be superimposed are partitioned into fragments. Starting with placements of a selected anchor fragment, computed by two alternative approaches, the remaining fragments are added iteratively. At each step, flexibility is considered by allowing the respective added fragment to adopt a discrete set of conformations. The mean computing time per test case is about 1:30 min on a common-day workstation. FLEXS is fast enough to be used as a tool for virtual ligand screening. A database of typical drug molecules has been screened for potential fibrinogen receptor antagonists. FLEXS is capable of retrieving all ligands assigned to platelet aggregation properties among the first 20 hits. Furthermore, the program suggests additional interesting candidates, likely to be active at the same receptor. FLEXS proves to be superior to commonly used retrieval techniques based on 2D fingerprint similarities. The accuracy of computed superpositions determines the relevance of subsequently performed ligand analyses. In order to validate the quality of FLEXS alignments, we attempted to reproduce a set of 284 mutual superpositions derived from experimental data on 76 protein−ligand complexes of 14 proteins. The ligands considered cover the whole range of drug-size molecules from 18 to 158 atoms (PDB codes: 3ptb, 2er7). The performance of the algorithm critically depends on the sizes of the molecules to be superimposed. The limitations are clearly demonstrated with large peptidic inhibitors in the HIV and the endothiapepsin data set. Problems also occur in the presence of multiple binding modes (e.g., elastase and human rhinovirus). The most convincing results are achieved with small- and medium-sized molecules (as, e.g., the ligands of trypsin, thrombin, and dihydrofolate reductase). In more than half of the entire test set, we achieve rms deviations between computed and observed alignment of below 1.5 Å. This underlines the reliability of FLEXS-generated alignments.

Introduction

In drug design, often enough, no structural information about a particular receptor protein, of therapeutic interest, is available. However, in many such cases, a considerable number of different ligands are known together with their measured binding affinities toward the receptor under consideration. In such a situation, a set of plausible relative superpositions of different ligands, approximating their putative binding geometry, is highly desired. Accordingly, generating structural superpositions or alignments of ligands is usually the method of choice to prepare data for the subsequent techniques that analyze the similarity or diversity of the 3D structure of these ligands. For example, structural superpositions are prerequisite to perform 3D-QSAR studies, to derive sophisticated pharmacophore models, or to embark into receptor modeling.

In general, drug-size molecules possess several rotatable bonds. Accordingly, they can adopt many low-energy conformations, one of which is likely to be present inside the receptor binding pocket. As a consequence, the consideration of conformational flexibility is of utmost importance. However, the problem of estimating the relevant bioactive conformation is extremely difficult, and real-life applications would be impossible to perform if a reference ligand of limited conformational flexibility is not available or if a set of distinct ligands showing conformational rigidity in complementary parts of their molecular skeletons is not known.

Usually, in the absence of a structurally known receptor, modeling techniques cannot achieve a signifi-

* Corresponding author. Phone: +49/2241/142481. Fax: +49/2241/142656. E-mail: Lemmen@gmd.de.
† GMD.
‡ Philipps-University of Marburg.

10.1021/jm981037l CCC: $15.00
Published on Web 10/14/1998

cance and accuracy that are comparable to recent docking methods.[1−3] The produced alignment can only exploit the ligand information, whereas docking methods use both receptor and ligand information.

An ideal reference ligand that is relatively rigid and shows high spatial occupancy and tight binding can still be an inferior drug molecule due to various reasons (e.g., toxicity, bioavailability, metabolism). In such cases, the search for new structurally diverse ligands is an important task. Thus, even in the case of the rather stringent prerequisite about an already known propitious reference ligand, there can be a definite need for alignment techniques.

Various methods for ligand superposition have been reported in the literature. Many of them treat the ligands as rigid,[4−6] or consider a limited set of rigid conformers in a sequential manner.[7−9] Others require predefined relationships between functional groups, assumed to be similar, in the molecules,[10−12] or allow for only a limited number of functional groups, to be able to test every possible matching or partial matching of groups separately.[13,14] Other approaches are targeted at very fast search, and accordingly, only approximate alignments are provided during the search through large databases.[15−17]

To the best of our knowledge, the program GASP[18] is currently the only method available that can handle molecular flexibility of several ligands without relying on predefined correspondences between groups in the superimposed molecules. The program does not seriously limit the number of functional groups or possible conformers. The runtime of GASP is in the range of 1 h of computing time per problem instance. Overviews on ligand superposition methods can be found in recent reviews.[19,20]

We developed the tool FLEXS for superimposing pairs of ligands, one of which is treated as flexible (*test ligand*) and is placed onto the other ligand which is treated as a rigid structure (*reference ligand*). The runtime of FLEXS lies within a few minutes on a common-day workstation for a single-problem instance. Such short computing times enable the use of FLEXS in various ways. First, it can be used to screen databases of considerable size. Second, it can be applied interactively to analyze a single superposition in detail. Third, different conformers of a partially flexible reference structure can be handled in a sequential manner, each as a reference to be aligned with a set of test ligands. Since the method can easily be parallelized on the data level, further large-scale applications are conceivable.

In the following section, we give a brief overview of the overall superpositioning strategy of FLEXS and details on the models used. Then we provide and discuss results of two kinds of evaluations of FLEXS. A description of the algorithms and their implementation is appended in the algorithmic section. Wherever appropriate, we provide pointers into this section for the interested reader. Additional data on all the superposition experiments, including structural formulas and 3D models, can be found on our web page (http://cartan.gmd.de/reliwe/superpose_res.html).

Theory and Methods

Overall Superposition Strategy. The overall strategy of our flexible superposition method is an iterative incremental construction procedure. This type of approach has already shown to be successful in flexible ligand docking[1,2,21] and in de novo design.[22−24]

The superposition method has been conceptualized as an extension and variation of the docking program FLEXX.[1] In essence, the strategy of FLEXS is to decompose the flexible ligand into small and relatively rigid portions (*fragments*), to start the placement with a user-defined anchor fragment (*base fragment* and *base placement*, respectively), and, subsequently, to add the remaining fragments in a stepwise manner (*incremental construction*) taking the conformational degrees of freedom into account. The key algorithms and the physicochemical models involved have been described elsewhere, together with some preliminary results.[25] We extended our approach in three ways. First, we added a fast rigid-body placement procedure called RIGFIT using Fourier space methods.[26] This procedure can be applied as an alternative base placement method which is especially well-suited for placing relatively large base fragments with only few functional groups that are suited for directional interactions. Second, we extended the incremental construction procedure in such a way that the sequence in which fragments are added is selected dynamically depending on the actual placement. This extension turns out to be effective in cases where the flexible test ligand partially extends beyond the reference ligand. Third, we performed a systematic parameter study in order to improve the runtime and the quality of the achieved superpositions. We did so by optimizing the settings of several adjustable parameters in the scoring function calculated for each actually obtained alignment.[27] The improvements have been validated using an enlarged test set of 284 examples in order to broaden the scope of our tool.

Physicochemical Model. We briefly summarize our approach to handling conformational flexibility, modeling putative intermolecular interactions with a possible receptor, and spatially distributing various localized physicochemical properties across the molecules.

Molecular flexibility is handled using discrete sets of torsional angles for each rotatable bond[28] and sets of distinct ring conformations. The latter sets are computed by either SCA[29] or CORINA.[30]

Intermolecular interactions are divided into *highly directional* (hydrogen bonds, salt bridges) and *less directional* (lipophilic interactions). Highly directional interactions are modeled in terms of *interaction centers* (key atom in the functional group of interest) and *interaction geometries* (geometrical description of the position where a countergroup would be expected[31]). The comparison of crystallographically observed binding geometries of structurally diverse ligands in a common receptor site reveals that the countergroups superpose. However, the interaction centers in the ligands do not necessarily coincide. Figure 1 illustrates this situation for two thrombin ligands.

Less directional interactions, such as interactions between aromatic rings or aromatic moieties and amide bonds, are handled differently. Crystal data show that, to a first approximation, in these cases already the respective functional groups are in close proximity. In addition, if available, also the predominant directions of the corresponding interaction geometries are usually quite similar.

These observations motivated the introduction of the concept of *paired intermolecular interactions*.[25] The interaction geometries around functional groups are approximated by sets of discrete *interaction points*. This treatment allows us to apply discrete combinatorial procedures for the placement of molecular fragments.

In order to distribute physicochemical properties, such as local hydrophobicity, partial atomic charges, and H-bonding potential across the molecules, the respective densities are approximated by sets of Gaussian functions (*Gaussians*, for short). The center of a Gaussian is positioned in the region of space where the respective property is expected (see Figure 2). By default this is an atom center; however, the user can modify this default in various ways.[25]

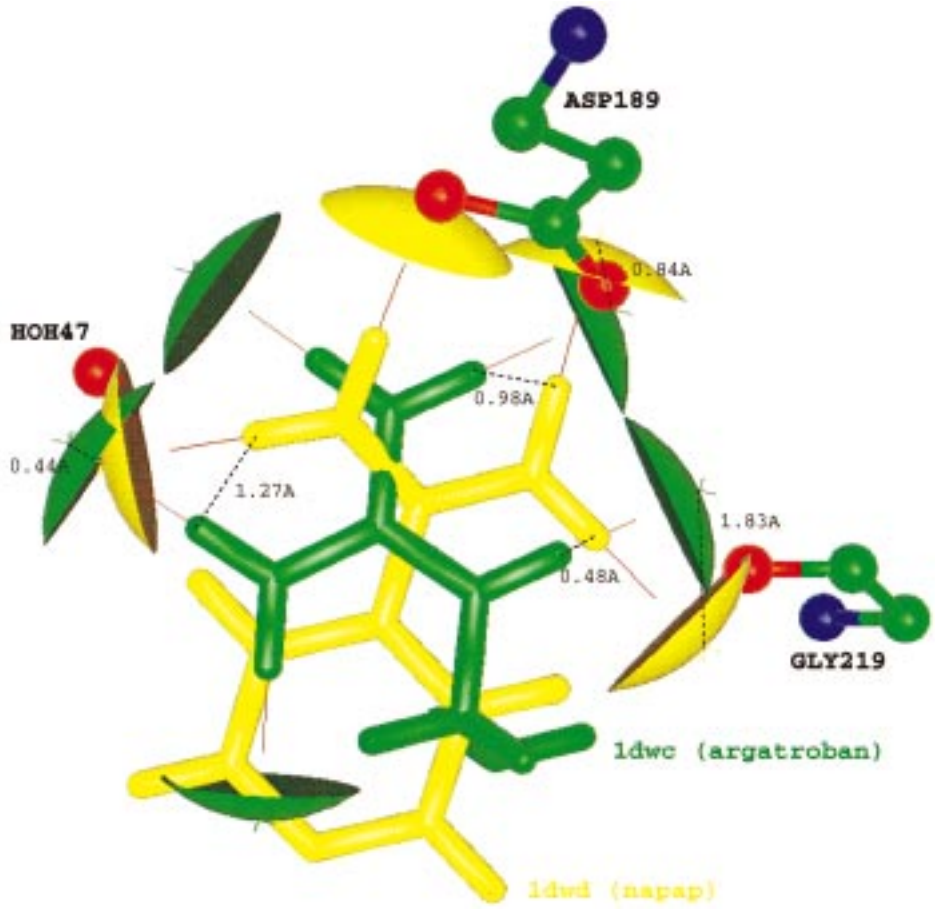

Figure 1. Alignment of the benzamidino group of napap (yellow) and the guanidinium group of argatroban (green) as derived by the superposition of corresponding C_α-positions of thrombin. The ligands are shown together with their interaction geometries, main directions, and site points for the main directions (cap, line, and cross in the same color). Two interacting groups of the protein and one structural water molecule are also provided. Neither the interaction centers nor the main directions coincide (distances are indicated by dashed black lines and measured in Å). However, in all cases, the corresponding protein countergroup falls close to the intersection of the involved interaction geometries.

Two parameters define the shape of a Gaussian function $g = ae^{-bx^2}$, the height a and the width b. The height of a Gaussian, modeling the 'electron density', is given by the number of electrons of the corresponding atom. The height for the property 'partial charge' is given by the partial charge of the atom. The height for Gaussians modeling the 'H-bonding potential' equals 1.0 for atoms that are able to form H-bonds as a donor or acceptor function, respectively. The assignment of hydrophobicities is discussed separately below. The width b is constant for all Gaussians. Empirical testing showed that a value of 1.2 is a reasonable choice.

In contrast to our previous hydrophobicity model, we replaced the fragment-based assignment[32] by a simple rule-based assignment. Our superposition approach requires localized property values. However, those found in the literature are optimized to add up to the global property log p. As a consequence, we simply take the absolute value of the partial charge of an atom a, $|\mathrm{chg}(a)|$, and classify atoms according to the following scheme into hydrophobic, hydrophilic, and ambiguous.

$$\text{if } |\mathrm{chg}(a)| > \mathrm{th}_1(\mathrm{type}(a)) \Rightarrow \text{hydrophilic}$$

$$\text{else if } |\mathrm{chg}(a)| < \mathrm{th}_2(\mathrm{type}(a)) \Rightarrow \text{hydrophobic}$$

$$\text{else} \Rightarrow \text{ambiguous}$$

Table 1 shows the classification of atom types and the corresponding threshold values which have been derived by visual inspection of several ligand molecules. The heights of the corresponding Gaussians are 1.0 for hydrophobic atoms, −1.0 for hydrophilic atoms, and 0.0 for ambiguous atoms. The improvement in our results with respect to molecular superpositioning can be seen in Table 3. We intend to further improve these results using a more sophisticated classification.

Table 1. Hydrophobicity Classification Scheme[a]

type(a)	th_1	th_2
H	0.1	0.06
C, N, O, F, B	0.2	0.1
P, Cl, Br, J, S	∞	0.1

[a] Type(a) indicates the element of the corresponding atom a, and th_1 and th_2 give the upper and lower thresholds, respectively.

Results and Discussion

As mentioned in the Introduction, structural superposition is the method of choice if the 3D structure of the target protein is unknown. However, in this case it is difficult to anticipate whether a computed mutual superposition of two structurally diverse ligands provides any useful information on the actual binding site geometry of the structurally unknown binding pocket. Instead, we demonstrate the potential of our superposition method by means of two applications. The first will show that it can be used for virtual screening in order to detect high-affinity ligands as alternative lead structures from a database of chemically characterized

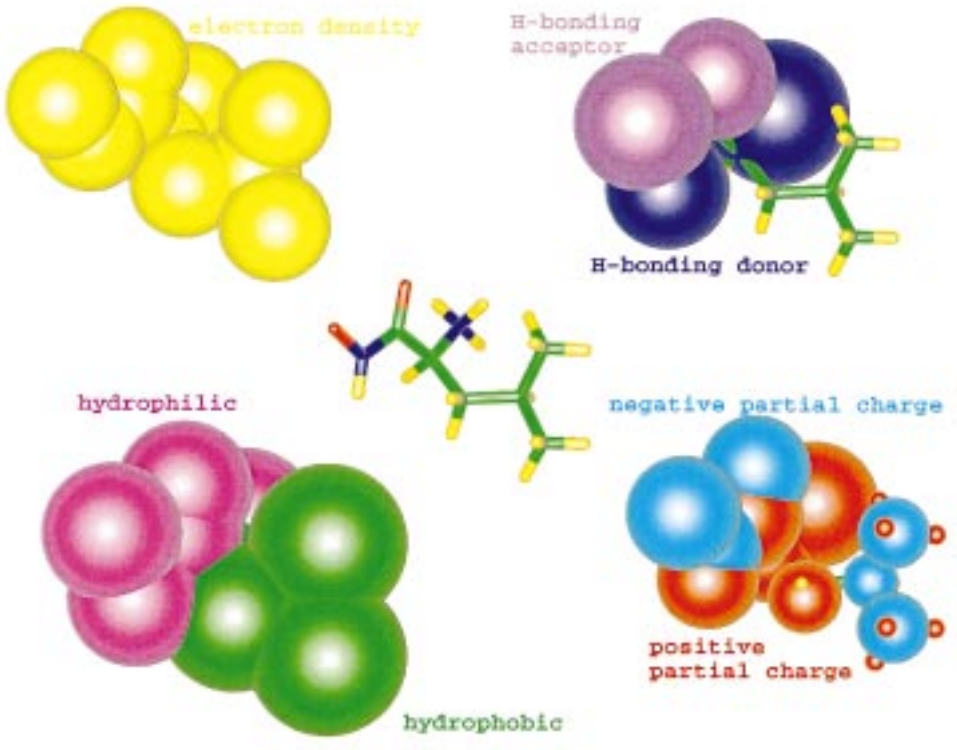

Figure 2. L-Leucylhydroxylamine (extracted from the complex with thermolysin; PDB code: 4tln) is shown with its Gaussian representation of different qualities. The isocontour surfaces are given by appropriate threshold values in order to display an illustrative level of the respective Gaussian functions.

compounds. Prerequisite for this application is the availability of a potent reference ligand with unambiguously defined bioactive conformation. The second application serves as an elaborate validation with respect to the accuracy in predicting ligand binding modes. For this purpose we try to reproduce the crystallographically observed binding geometry of various examples with ligands of different sizes. This study demonstrates the scope and limitations of the present approach.

Virtual Screening for Fibrinogen Receptor Antagonists. For this case study, the search for fibrinogen receptor antagonists, no information about the 3D geometry of the target protein is available. However, detailed studies of rigid cyclic peptides have revealed the putative arrangement of the pharmacophoric groups at the recognition site of the protein.[49] The essential epitope comprises an arginine, a glycine, and an aspartate (RGD) followed by a phenylalanine. For our case study we selected the cyclic peptide (Figure 3, structure rgd-1) as the reference.

As a test data set for the search of novel compounds, we used a collection first compiled by Briem and Kuntz.[50] This set comprises 136 PAF antagonists, 114 HMG-CoA inhibitors, 40 ACE inhibitors, 49 thromboxan A2 receptor antagonists, 52 5HT3 receptor ligands, and 581 randomly selected compounds from the MDDR database.[51] In addition we included the tripeptide RGD and 11 chemically distinct fibrinogen receptor antagonists (Figure 3 structures rgd-2−rgd-12), yielding a total set of 984 compounds. Each structure in the dataset is labeled by its activity class and the index in Briem's data set.[50] An arbitrary conformation of each fibrinogen ligand has been generated by the SYBYL molecular modeling package.[52] After adding Gasteiger−Marsili charges,[53] the entire set of structures has been minimized with the Maximin force field within SYBYL.[52]

The analysis was split into two steps. At first, we applied a rather crude but fast filter. Since the guanidinium and the carboxylate group of the RGD peptide are the essential pharmacophoric groups, their properties were selected for the initial step. We extracted the amidinium and the carboxylate fragment, each with a terminal CH_2 group, from the reference RGD peptide. Partial charges were directly transferred from the peptide onto the two fragments. In the following, each of the two fragments was aligned separately with all 984 entries from the database using the RIGFIT method (cf. Numerical Placement Procedure, Algorithmic Section). The individual scores and the obtained alignment positions for the top ranking superpositions with respect to the two pharmacophoric fragments were stored. Then, the two independent scores were added for all 984 placements and sorted. In the following, only the 50 solutions possessing the highest combined scores were further considered for detailed alignment. The prefiltering using RIGFIT required 33:41 min, that is, about 1 s per test compound.

In the second, more elaborate, superposition step the entire 3D structure of the cyclic peptide was used as the reference. To unambiguously assign an appropriate base fragment in each of the 50 remaining test ligands, the results of the previous RIGFIT alignment of the amidinium fragment were evaluated. Those atoms of each database entry that were aligned next to the amidinium fragment were selected as base fragment for the subsequent FLEXS superposition.

In our opinion, this strategy has advantages over a pure assignment in terms of functional group topologies

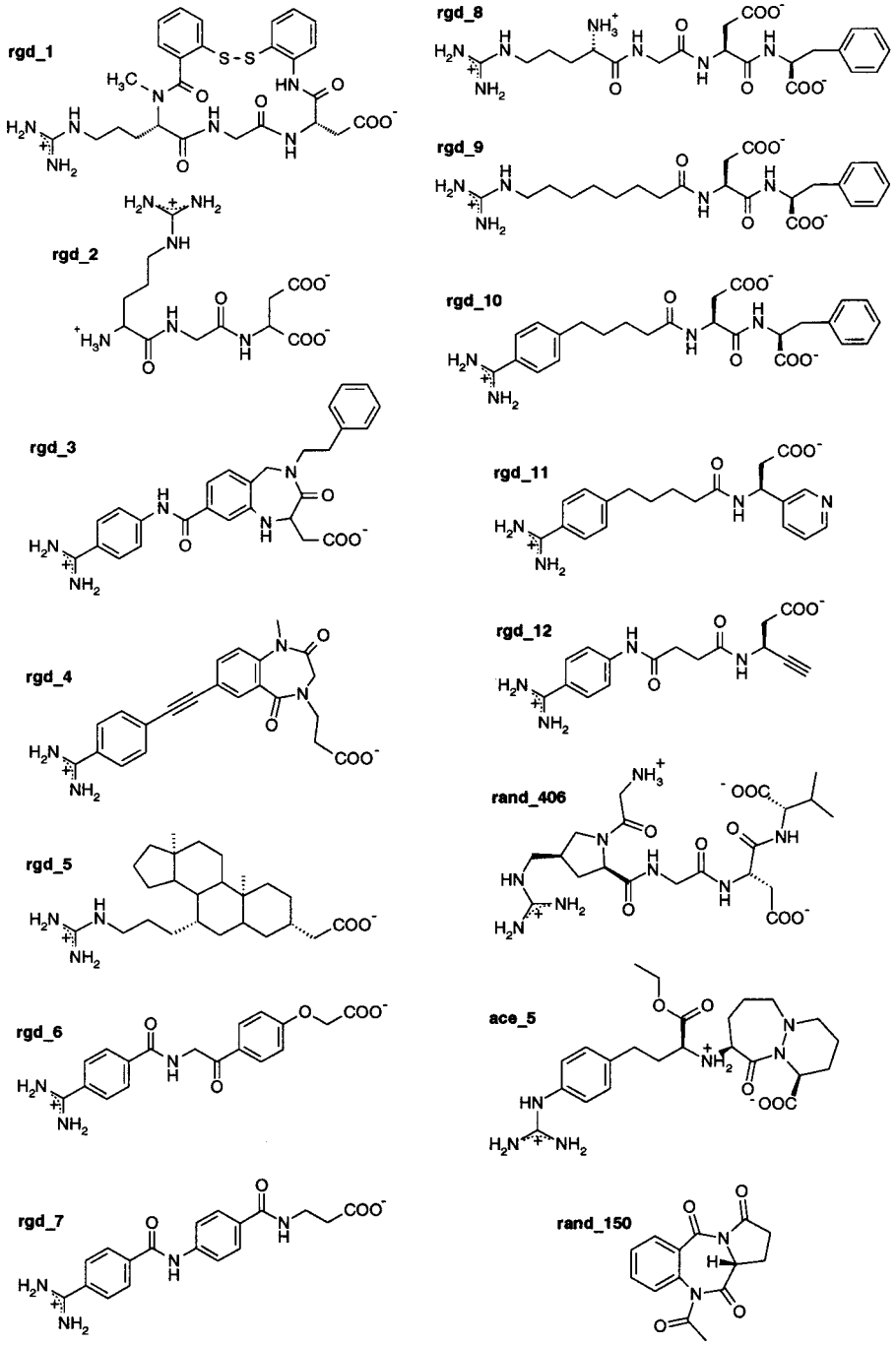

Figure 3. Chemical formulas of 15 examples out of the 984 compounds in the data set. The cyclic peptide rgd-1 is chosen as the rigid reference ligand. Structure rgd-2 depicts the tripeptide RGD. rgd-3−rgd-12 are 10 distinct, known fibrinogen receptor antagonists. The structures rand-406 and ace-5 depict two interesting additional hits. rand-406 is labeled as 'platelet aggregation properties' in the MDDR. rand-150 is the first DAYLIGHT hit.

based on 2D connectivities. For example, in the present case, one could have selected the fragment comprising a positively charged nitrogen. Using the information of the previous RIGFIT alignment allows us to focus on molecular similarity in terms of spatial physicochemical properties. The subsequent flexible alignment of the 50 preselected database entries with FLEXS followed the usual protocol. The entire procedure required 103:28 min, that is, about 2 min per test structure. For each case, only the best-scoring solution was considered. Interestingly, all compounds known to be potent RGD antagonists are included among the 20 best ranked solutions. Furthermore, the second best solution (see Figure 3, structure rand-406, alignment in Figure 4) originated from the portion of the database that was randomly selected from the MDDR. As subsequently checked, MDDR assigns 'platelet aggregation properties' to this compound. Another interesting solution belongs to the ACE inhibitors (Figure 3, structure ace-5). Its superposition with the reference is shown in Figure 4.

The example demonstrates that the FLEXS approach is capable of retrieving molecules likely to be good candidates for binding to a particular target. It can be asked, however, whether the computational effort spent on this computer screening is justified. Much faster methods for clustering molecules in terms of molecular

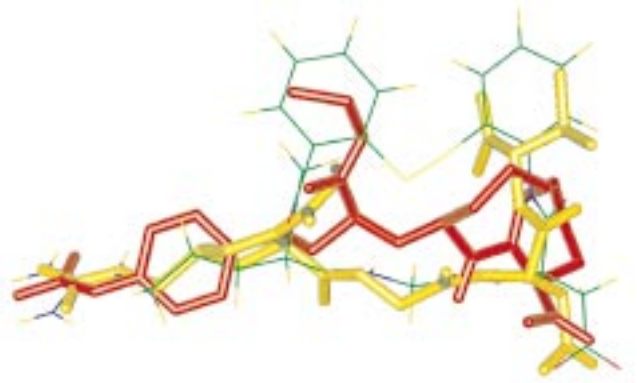

Figure 4. Flexible fit of two ligands not belonging to the set of known fibrinogen receptor antagonists (rand-406, yellow; ace-5, red) onto the cyclic reference ligand (rgd-1, type colored, lines).

Table 2. RGD Screen Results[a]

compound identifier	RIGFIT score	RIGFIT rank	FLEXS score	FLEXS rank	DAYLIGHT rank
rgd-1	351.6	14	1217.4	1	
rand-406	332.3	32	980.8	2	54
rgd-8	337.8	24	953.1	3	11
rand-545	364.2	10	934.6	4	147
rgd-10	364.0	11	901.6	5	21
rgd-3	397.5	1	885.0	6	8
rgd-9	324.9	44	860.2	7	25
rand-221	324.7	45	792.2	8	306
rand-527	365.0	9	770.1	9	299
rand-398	367.4	8	743.7	10	47
ace-5	329.437	37	727.2	11	38
rgd-2	336.6	26	715.1	12	83
rgd-11	372.1	6	712.5	13	182
rgd-5	377.4	3	709.5	14	847
rgd-4	376.8	5	687.2	15	6
rgd-12	377.3	4	685.0	16	142
rgd-7	348.6	18	678.7	17	12
rand-183	355.8	13	668.8	18	43
rgd-6	340.9	22	665.6	19	309
rand-356	351.3	15	644.1	20	59
ace-10	335.3	30	636.1	21	165
rand-559	322.0	48	625.9	22	356
rand-267	325.4	42	510.5	23	148
paf-13	323.6	46	496.3	24	418
rand-517	329.8	36	494.4	25	607

[a] The scores and ranks of 25 structures after the two separate screening phases, carried out with RIGFIT and FLEXS, respectively, are given in columns 2−5. The last column provides the corresponding ranks for the DAYLIGHT fingerprints. Rows are sorted by the FLEXS score. The compound identifier nomenclature in the first column is explained in the text. Only the top 25 FLEXS hits are provided.

similarity are available. Among them are approaches based on the most popular fingerprint descriptors suggested in the DAYLIGHT software.[54] Similarity with a given reference is determined by comparing the assigned bit strings according to the Tanimoto coefficient.

In Table 2 the scorings of the 25 best solutions of our approach are compared to the ranking based on DAYLIGHT fingerprints. Out of the 11 ligands known to antagonize the fibrinogen receptor, only 4 are found among the 20 best-scored cases. The remaining ones are scattered up to solution 847. Furthermore, in Figure 3, the chemical formula of structure rand-150, the best solution from the fingerprint method, is listed. This molecule contains a benzodiazepine moiety. The DAYLIGHT fingerprint approach is performed within seconds; however, the FLEXS screening appears to suggest more reliable and relevant solutions. Accordingly, the additional computational effort of about 2 h appears clearly justified.

Reproduction of Known Binding Site Geometries. The previous application was performed to demonstrate the relevance and practical impact of our method for the lead discovery process. The following validation study intends to explore the scope and accuracy limits of FLEXS.

We selected a set of 14 proteins for which several protein−ligand complexes have been determined crystallographically. We mutually superimposed these complexes by minimizing the positional differences between the backbone C_α atoms. Then we extracted the ligands from the complexes keeping their obtained relative orientation in space. This defines our *reference alignment*. Note that the validation process mentioned above involves inherent accuracy limits of about 0.7 Å. These uncertainties are due to experimental inaccuracies or ligand-induced structural modifications of the protein backbone.[7]

In all test cases, we superimposed pairs of ligands. The reference ligand was taken in its crystallographically determined binding conformation, and the test ligand was transformed to an arbitrary orientation and conformation. The computed alignment of the test ligand was compared to its reference alignment. As an objective measure for the goodness-of-fit, we used the rms deviation (rmsd, for short) in atomic coordinates excluding hydrogens. However, rmsd values can be misleading, especially if they adopt larger values. A particular rmsd value can result either from a moderate overall fit or from a convincing fit in some structural portions and an obvious misfit in another part due to conformational and orientational differences.

As a consequence, we inspected cases with an rmsd beyond 1.5 Å visually. Less clear-cut criteria such as goodness-of-fit in some substructural parts or correct orientations of the key functional groups involved in directional interactions with the protein were considered.

Our analysis starts with those examples for which both ligands possess fairly rigid skeletons. These examples mainly test the quality of the initial placement phase by either the RIGFIT or the paired interaction approach (cf. Two Base Placement Methods, Algorithmic Section). This group of examples comprises steroid-type ligands binding to an immunoglobulin, aromatic azo compounds ligating streptavidin, particularly small ligands inhibiting trypsin, and sugar-type molecules observed as ligands of glycogen phosphorylase and concanavalin. As another extreme for testing conformational flexibility, we considered large peptide-type ligands. In this group, ligands binding to endothiapepsin, HIV-protease, elastase, thermolysin, and carboxypeptidase are discussed. Finally, we extended our test to ligands of considerable size, containing a large variety of organic building blocks besides peptidic portions. In this series, examples of ligands binding to thrombin, dihydrofolate reductase, human rhinovirus, and fructose bisphosphatase are discussed.

For each protein example to be considered, we give a cross-table containing four values in each entry. The first value gives the best rmsd obtained (in Å) between

Table 3. Statistics of FLEXS Performance for Different Variants of the Algorithms[a]

			ϕ rmsd (Å)				N_{placem} (%) with accuracy x		
no.	test scenario	ϕ rt (s)	1st	top 10	all	% fpl	x < 1.0 Å	x < 1.5 Å	x < 2.0 Å
1	variable sequence construction	88	1.20	1.14	1.05	83.0	79 (28%)	126 (44%)	169 (60%)
2	predefined sequence construction	64	1.20	1.13	1.05	78.6	81 (29%)	125 (44%)	167 (59%)
3	rigid-body superposition	10	0.78	0.84	0.84		161 (57%)	230 (81%)	249 (88%)
4	default parameter set	88	1.18	1.08	1.06	81.2	71 (25%)	112 (39%)	143 (50%)
5	old hydrophobicity values	90	1.20	1.10	1.03	80.7	72 (25%)	118 (42%)	150 (53%)
6	CORINA ring conformations	75	1.18	1.14	1.06	80.6	746 (26%)	120 (42%)	159 (56%)

[a] No., line number (used as reference in the text); ϕ rt, mean runtime; ϕ rmsd, mean rms deviations (considering all placements below 2.0 Å) for the highest ranking placement, the best of the top 10 ranking placements, and the best of all placements generated; % fpl, mean percentage of fragments placed with an rmsd < 1.5 Å during complex construction; N_{placem} (%), number and percentage of placements predicted with rms deviations smaller than 1.0, 1.5, and 2.0 Å in the first 10 places.

Table 4. Statistics of FLEXS Performance for Different Groups of Ligands[a]

			ϕ rmsd (Å)				N_{placem} (%) with accuracy x		
test scenario	N_{expl}	ϕ rt (s)	1st	top 10	all	% fpl	x < 1.0 Å	x < 1.5 Å	x < 2.0 Å
ϕ ligands with few conformations	71	35	1.11	1.09	1.02	97.1	34 (48%)	5 (31%)	65 (92%)
immunoglobulin ligands	16	49	1.08	1.48	1.52	87.5	1 (6%)	5 (31%)	11 (69%)
streptavidin ligands	20	2	1.09	0.93	0.89	100.0	13 (65%)	20 (100%)	20 (100%)
trypsin ligands	27	13	1.19	1.10	0.93	100.0	12 (44%)	21 (78%)	26 (96%)
glycogen phosphorylase ligands	6	166	0.70	0.66	0.66	100.0	6 (100%)	6 (100%)	6 (100%)
concanavalin ligands	2	152	1.66	0.84	0.51	100.0	2 (100%)	2 (100%)	2 (100%)
ϕ large peptidic ligands	141	114	1.38	1.34	1.21	73.6	23 (16%)	45 (32%)	65 (46%)
endothiapepsin ligands	7	275	1.72	1.53	1.51	79.8	0 (0%)	3 (43%)	4 (57%)
HIV-protease ligands	41	233	1.70	1.73	1.66	46.3	0 (0%)	2 (5%)	10 (24%)
elastase ligands	14	58	1.29	1.07	0.92	73.9	2 (14%)	6 (43%)	6 (43%)
thermolysin ligands	65	48	1.21	1.19	1.02	84.3	14 (22%)	21 (32%)	31 (48%)
carboxypeptidase A ligands	14	44	1.16	1.06	0.90	100.0	7 (50%)	13 (93%)	14 (100%)
ϕ medium-sized molecules	72	89	1.14	0.98	0.89	93.6	22 (31%)	27 (38%)	39 (54%)
thrombin ligands	12	98	1.65	1.52	1.37	80.6	22 (31%)	4 (33%)	11 (92%)
dihydrofolate reductase ligands	2	153	1.56	1.53	1.42	100.0	0 (0%)	1 (50%)	2 (100%)
human rhinovirus ligands	56	84	1.01	0.84	0.76	97.4	20 (36%)	21 (38%)	24 (43%)
fructose bisphosphatase ligands	2	96	1.43	1.18	1.18	57.1	1 (50%)	1 (50%)	2 (100%)

[a] Same nomenclature as in Table 3. Additionally the number of examples tested N_{expl} is provided.

observed and computed orientation of the test ligand among the 10 best-scored solutions (referred to as BEST-10, the rank is given in parentheses). From a practical point of view, in modeling applications only perhaps 10 distinct solutions are considered. In the second row, we give the lowest rmsd found among all of the suggested solutions (referred to as BEST-ALL). The third value (in %) refers to the maximum percentage of the fragments of the test ligand being placed with an rmsd better than 1.5 Å during the complex construction phase (referred to as PARTIAL-GOOD). Frequently, reasonable partial placements can be obtained even if the final (complete) placement deviates substantially (cf. Variable Sequence Construction, Algorithmic Section). The fourth row gives the rmsd obtained for the test ligand being rigid-body-fitted in its crystallographically given conformation according to the RIGFIT approach (referred to as RIGID). The various ligands are labeled by the PDB codes assigned to the corresponding protein−ligand complexes. An additional set of structures, obtained quite recently, is indicated by their labels used during the CASP2 competition.[45] Each table entry refers to the superposition of the ligands denoted by the names in the corresponding rows and columns. Reference ligands (RL) are listed vertically, whereas the test ligands (TL) are arranged horizontally. The self-fit is found along the diagonals. It should be noted that the self-fit values are not considered in any of the statistical analyses, neither mentioned in the text nor displayed in Tables 3 and 4.

The tables are not symmetrical since the obtained results depend on which ligand is used as the reference. Those examples for which no relevant superposition can be expected are crossed out in the tables. These are cases for which either the mutual overlap of two ligands in the observed relative orientation is small or a large ligand is attempted to be superimposed with a substantially smaller one. Both cases cannot be expected to reveal relevant and significant results. They clearly indicate the limitations of approaches that rely purely on a comparison of ligand structures. To apply an objective selection criterion for the test cases, the following general rule has been applied:

$$\text{if } ((\text{vol(RL)} \cap \text{vol(TL)})/\text{vol(TL)} < 0.6$$

$$\vee\ N_{atom}(\text{TL}) \notin (\text{vol(RL)} \cap \text{vol(TL)}) > 10)$$

$$\Longrightarrow \text{reject example}$$

In valid examples the volume portion of the test ligand intersecting with the reference ligand has to be at least 60%. Additionally, the number of atoms in the test ligand located outside the intersection volume is limited to 10. The selection is carried out based on observed relative orientations ignoring hydrogen atoms.

In order to allow for a quick overview, the table entries are shaded according to the best rmsd found among the first 10 solutions (line 1 of the entry) as follows:

rmsd < 1.0 Å ⟹ white

rmsd < 1.5 Å ⟹ light gray

rmsd < 2.0 Å ⟹ medium gray

rmsd > 2.0 Å ⟹ dark gray

Examples for which FLEXS is not able to produce an alignment are indicated by a dash. rms deviations > 10 Å are indicated as such, and the rank is omitted for these cases. In our experience, an rmsd below 1.5 Å indicates the reproduction of the observed alignment closely approximating the experimentally given situation. If the deviation is between 1.5 Å and 2.0 Å, the overall orientation and conformation of a ligand is correctly reproduced. Above 2.0 Å rmsd, major differences occur, at least in some substructural portion of the test ligand.

FLEXS computes several solutions. However, with respect to practical modeling applications, it is very important that solutions that are similar to the experimentally detected superposition rank among the best candidates. A statistical analysis of the entire set of 284 superposition pairs shows that in about 40% of the cases a solution resembling experiment (rmsd < 1.5 Å) is found among the three best-scoring solutions. Detailed statistics from the application of different algorithms or different parameter settings to the entire test set are given in Table 3. Results obtained by the variable versus the predefined sequence construction strategy can be compared in the first two rows (cf. Variable Sequence Construction, Algorithmic Section). The quality of placements obtained by applying RIGFIT to the crystallographically obtained conformations is listed in the third row (cf. Numerical Placement Procedure, Algorithmic Section). The first and fourth rows, respectively, allow for a comparison of the results obtained with and without calibration of the empirical scoring function (cf. Scoring Scheme, Algorithmic Section). With default parameters (row 4), every coefficient ΔS_x in the linear scoring scheme equals 1.0. A comparison of row 1 with row 5 indicates the improvements achieved by the novel hydrophobicity classification. Rows 1 and 6 compare the results using SCA or CORINA, respectively, for generating discrete ring conformations. Table 4 displays the FLEXS results on different subsets of the test suite. It is grouped into three sections, according to the discussion. Mean values are provided within each group.

All computations have been performed on a single processor SUN-Ultra-30 Workstation with 128MB of main memory and 296-MHz clock speed. We used a single-parameter setting (cf. Table 20, below), except for the run explicitly marked by 'default parameter set' (row 4 in Table 3). Computing times for I/O and ligand preparation have been excluded from the tables. However, these steps require at most 2 s per test case.

Superposition of Ligands with Few Conformational Degrees of Freedom. All ligands considered in this part of the study have a limited number of low-energy conformations. Accordingly, most of the examples perform well in the alignment procedure. Since for a series of the cases the number of functional groups is not large enough for the combinatorial placement procedure, mostly the RIGFIT method operating in Fourier space is applied (cf. Variable Sequence Construction, Algorithmic Section). The average runtime of the flexible superposition within this subset is 35 s.

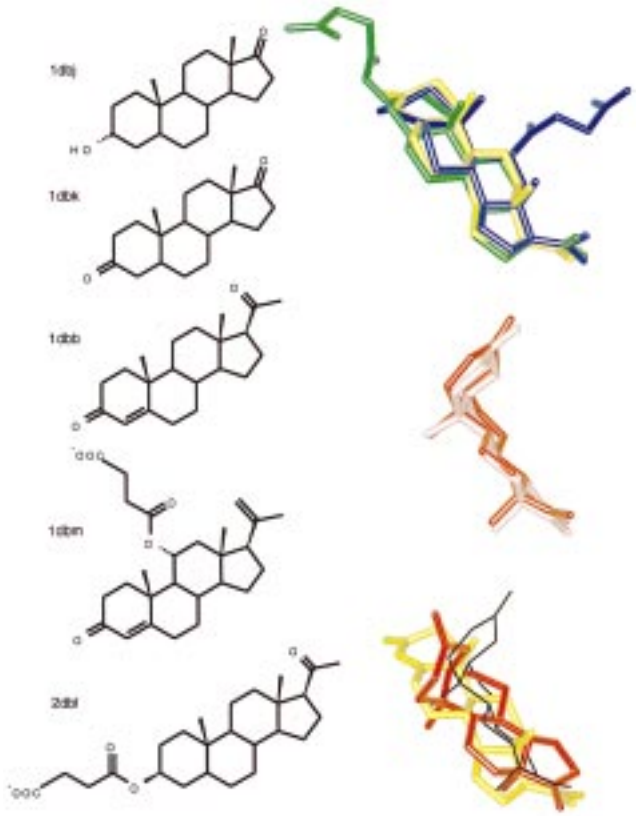

Figure 5. Five ligands binding to immunoglobulin are displayed in three separate images, all in the same orientation with regard to the protein. The upper and central parts show the experimentally observed ligand geometries grouped into two sets (1dbm, blue; 2dbl, green; 1dbb, yellow; and 1dbj, pink; 1dbk, red). The lower part shows the FLEXS-superposition (rank 3, 1.6 Å rmsd) for the alignment of 1dbk (red) onto 1dbb (yellow). In addition, the observed relative orientation of 1dbk (black, lines) is provided. Despite the fact that reference and test ligand adopt significantly different overall shapes, the calculated alignment approximates the experimental solution fairly well.

1. Immunoglobulins. The present set of ligands binding to the F_{ab} fragment of a monoclonal antibody comprises two cholic acid-type (1dbj, 1dbk) and three steroid-type (1dbm, 2dbl, 1dbb) molecules. Figure 5 shows the observed ligand geometries of these examples.

For the sake of clarity, the two subsets are displayed separately. The fit of the two cholic acid derivatives among each other performs well (1dbj/1dbk). The steroid progesterone (1dbb) is substantially smaller than 1dbm which is substituted by a large ester group at the 11-position. The other steroid 2dbl is a monoester of succinic acid substituting the 5-OH group. The alignment among the three deviating steroids turns out reasonably well, especially if the smaller progesterone is fitted onto the other two larger derivatives. The mutual alignment of the steroids to cholic acid derivatives is less convincing, possibly due to the different stereochemistry at the A-to-B ring fusion. This gives rise to a quite different overall shape of the ligands (flat extended versus curved). In the lower part of Figure 5, the predicted superposition of 1dbk onto 1dbb (rank 3,

Table 5. Immunoglobulin[a]

RL \ TL	1dbj	1dbk	1dbm	2dbl	1dbb
1dbj	1.33 (1) 1.33 (1) 100.0% 0.0 (1)	0.41 (1) 0.41 (1) 100.0% 0.35 (2)			
1dbk	1.46 (1) 1.46 (1) 100.0% 0.36 (1)	0.44 (1) 0.44 (1) 100.0% 0.0 (1)			
1dbm	1.70 (4) 1.70 (4) 0.0% 1.38 (7)	1.77 (4) 1.77 (4) 0.0% 1.46 (5)	1.53 (1) 1.49 (64) 100.0% 0.0 (1)	1.59 (7) 1.59 (7) 33.3% 0.59 (2)	1.20 (4) 1.20 (4) 100.0% 0.40 (2)
2dbl	1.95 (10) 1.63 (20) 0.0% 1.65 (5)	1.72 (4) 1.72 (4) 0.0% 1.51 (2)		1.48 (3) 1.48 (3) 100.0% 0.0 (1)	1.17 (4) 1.17 (4) 100.0% 0.41 (3)
1dbb		1.68 (3) 1.68 (3) 0.0% 1.57 (3)		1.36 (9) 1.34 (18) 100.0% 0.61 (1)	1.09 (3) 1.09 (3) 100.0% 0.0 (1)

[a] Sequence of values within each entry: BEST-10, BEST-ALL, PARTIAL-GOOD, RIGID.

Table 6. Streptavidin[a]

RL \ TL	1srf	1srg	1srh	1sri	1srj
1srf	0.56 (1) 0.56 (1) 100.0% 0.0 (1)	0.66 (2) 0.66 (2) 100.0% 0.81 (1)	0.65 (3) 0.60 (68) 100.0% 0.49 (1)	0.90 (2) 0.90 (2) 100.0% 1.10 (1)	1.11 (5) 1.11 (5) 100.0% 1.02 (1)
1srg	0.99 (2) 0.99 (2) 100.0% 0.83 (1)	0.26 (1) 0.26 (1) 100.0% 0.0 (1)	0.88 (5) 0.79 (27) 100.0% 0.72 (1)	0.65 (7) 0.65 (7) 100.0% 0.91 (1)	0.74 (1) 0.74 (1) 100.0% 0.51 (1)
1srh	0.64 (1) 0.64 (1) 100.0% 0.47 (1)	0.74 (3) 0.74 (3) 100.0% 0.74 (1)	0.49 (4) 0.49 (4) 100.0% 0.0 (1)	0.62 (4) 0.62 (4) 100.0% 0.82 (1)	1.09 (2) 1.09 (2) 100.0% 1.06 (1)
1sri	1.35 (7) 1.00 (23) 100.0% 1.23 (1)	0.95 (3) 0.95 (3) 100.0% 0.90 (1)	0.92 (10) 0.92 (10) 100.0% 0.88 (1)	0.59 (1) 0.59 (1) 100.0% 0.0 (1)	1.11 (6) 1.11 (6) 100.0% 1.27 (1)
1srj	1.31 (7) 1.31 (7) 100.0% 1.03 (1)	0.63 (5) 0.63 (5) 100.0% 0.50 (1)	1.28 (1) 1.10 (163) 100.0% 1.06 (1)	1.32 (5) 1.32 (5) 100.0% 1.26 (1)	0.26 (1) 0.26 (1) 100.0% 0.0 (1)

[a] Sequence of values within each entry: BEST-10, BEST-ALL, PARTIAL-GOOD, RIGID.

1.6 Å rmsd) is shown. All considered immunoglobulin ligands are quite rigid and hydrophobic. They only show a limited number of functional groups potentially involved in highly directional interactions. Accordingly, FLEXS selects the RIGFIT algorithm to place the base fragment (Table 5).

2. Streptavidin. The structural deviation among the five azo compounds binding to streptavidin is relatively small. All variations occur at the iminophenol moiety. The mutual alignment is convincing, although a different number of water molecules is involved in ligand binding. The largest deviations occur with the superpositions involving 1srj that bears a naphthol moiety instead of a substituted phenolic portion (Table 6).

3. Trypsin. Five of the six ligands binding to trypsin display different primary amines with a terminal phenyl group. The spacers between the amino and the aromatic group differ in chain length and composition. The sixth ligand (3ptb) is benzamidine. Most of the superpositions work reasonably well. Since the considered ligands only possess a limited number of functional groups capable of performing highly directional interactions, the RIGFIT algorithm is used for base fragment placement (Table 7).

4. Glycogen Phosphorylase. All four ligands correspond to phosphorylated glycose-type derivatives. The sugar moiety in 1gpy occupies a different region in the active site than in the three other derivatives. Accordingly, a meaningful superposition cannot be expected if

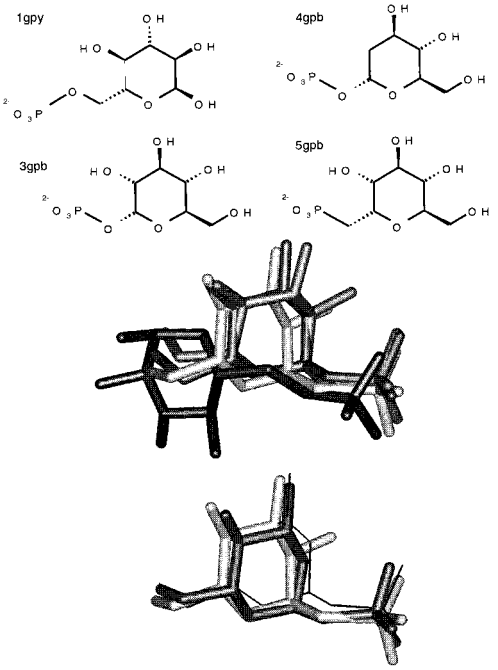

Figure 6. (Top) Four ligands binding to glycogen phosphorylase (1gpy, black; 3gpb, medium gray; 4gpb, dark gray; 5gpb, light gray) in their observed binding orientation and conformation. It is obvious that the position of the sugar moiety in 1gpy deviates significantly from that in the other ligands. Thus, the rejection of this ligand from the analysis appears appropriate. (Bottom) Top ranking prediction (0.51 Å rmsd, same orientation as above with regard to the protein) of 4gpb (dark gray) onto 5gpb (light gray). For comparison purposes, the observed geometry of 4gpb (black, lines) is included.

Table 7. Trypsin[a]

RL \ TL	1tnh	1tni	1tnj	1tnk	1tnl	3ptb
1tnh	0.16 (4) 0.16 (4) 100.0% 0.0 (1)		0.96 (6) 0.96 (6) 100.0% 1.46 (1)	1.06 (8) 1.06 (8) 100.0% 1.69 (1)	1.15 (10) 1.16 (50) 100.0% 2.19 (1)	0.34 (4) 0.34 (4) 100.0% 0.40 (1)
1tni	1.37 (4) 1.37 (4) 100.0% 1.55 (3)	0.94 (2) 0.94 (2) 100.0% 0.0 (1)		1.95 (1) 1.79 (304) 0.0% 1.37 (1)	1.89 (10) 0.79 (22) 100.0% 2.03 (1)	1.39 (3) 1.39 (3) 100.0% 1.42 (1)
1tnj	1.24 (6) 1.24 (6) 100.0% 1.50 (1)	1.41 (7) 1.00 (47) 100.0% 0.71 (1)	0.65 (1) 0.65 (1) 100.0% 0.0 (2)	0.95 (4) 0.95 (4) 100.0% 0.50 (1)	0.49 (10) 0.49 (10) 100.0% 0.64 (1)	1.01 (2) 1.01 (2) 100.0% 1.06 (4)
1tnk	1.01 (1) 1.01 (1) 100.0% 1.53 (2)		0.88 (10) 0.88 (10) 100.0% 0.50 (1)	0.76 (4) 0.73 (37) 100.0% 0.0 (1)	0.45 (9) 0.32 (21) 100.0% 0.43 (1)	0.83 (2) 0.83 (2) 100.0% 0.92 (2)
1tnl	0.90 (7) 0.90 (7) 100.0% 2.41 (1)	1.91 (6) 0.88 (106) 100.0% 1.83 (2)	0.97 (5) 0.97 (5) 100.0% 0.63 (1)	0.96 (2) 0.96 (2) 100.0% 0.45 (1)	0.16 (1) 0.16 (1) 100.0% 0.0 (1)	0.72 (2) 0.72 (2) 100.0% 0.81 (1)
3ptb	0.40 (9) 0.40 (9) 100.0% 0.35 (3)		0.71 (8) 0.71 (14) 100.0% 1.49 (3)	1.21 (5) 1.10 (56) 100.0% 0.99 (1)	1.03 (1) 1.03 (1) 100.0% 0.88 (6)	0.31 (2) 0.28 (27) 100.0% 0.0 (1)

[a] Sequence of values within each entry: BEST-10, BEST-ALL, PARTIAL-GOOD, RIGID.

1gpy is used as the reference. Among the remaining three closely related ligands, the alignment is reproduced accurately. Figure 6 shows the observed binding geometries of this set of ligands and the fit of 4gpb onto 5gpb.

The glycogen phosphorylase ligands require comparatively long computing times. Due to the symmetry of the PO_3 group which serves as the base fragment in all cases, the combinatorial placement generates many similar positions for many nearly equivalent triangles. This results in hundred thousands of placements for the base fragment, most of which are merged during clustering. Note that the combinatorial placement

Table 8. Glycogen Phosphorylase[a]

RL \ TL	1gpy	3gpb	4gpb	5gpb
1gpy	0.68 (1) 0.68 (1) 100.0% 0.05 (1)			
3gpb		0.46 (2) 0.46 (2) 100.0% 0.0 (1)	0.66 (9) 0.66 (9) 100.0% 0.36 (1)	0.90 (7) 0.90 (7) 100.0% 0.72 (1)
4gpb		0.56 (1) 0.56 (1) 100.0% 0.34 (2)	0.33 (1) 0.33 (1) 100.0% 0.0 (1)	0.63 (6) 0.63 (6) 100.0% 0.47 (2)
5gpb		0.68 (1) 0.68 (1) 100.0% 0.72 (3)	0.51 (1) 0.51 (1) 100.0% 0.45 (1)	0.52 (1) 0.52 (1) 100.0% 0.0 (1)

[a] Sequence of values within each entry: BEST-10, BEST-ALL, PARTIAL-GOOD, RIGID.

Table 9. Concanavalin[a]

RL \ TL	5cna	t0013
5cna	0.25 (1) 0.25 (1) 100.0% 0.0 (1)	0.72 (5) 0.58 (14) 100.0% 1.38 (1)
t0013	0.97 (10) 0.43 (97) 100.0% 1.47 (2)	0.41 (1) 0.41 (1) 100.0% 0.0 (1)

[a] Sequence of values within each entry: BEST-10, BEST-ALL, PARTIAL-GOOD, RIGID.

avoids substructure comparison and therefore ignores equivalences on this level (Table 8).

5. Concanavalin. Even though, the two sugar-type ligands of concanavalin (a mannose and an arabinose derivative) show different ring sizes, they can be aligned convincingly by FLEXS (Table 9).

Superposition of Large Peptidic Ligands. To test the scope and limitation of handling flexibility during the alignment procedure, different sets of peptidic ligands were studied. The molecules under investigation span the range from di- to heptapeptides. Accordingly, the number of rotatable bonds amounts up to 35. In such cases, a combinatorial search can easily produce several billions of possible conformers. This is clearly beyond the scope of any presently available superposition technique. In addition to the size complexity, it can easily happen that side-chain portions are superimposed onto backbone portions of the reference during the incremental construction. After such an incorrect placement, the algorithm cannot compute a reasonable alignment, since backtracking is not implemented. Applying RIGFIT to the experimentally observed conformations results in reasonable alignments for most of the examples.

1. Endothiapepsin. The five peptidic inhibitors of endothiapepsin are quite large and have between 87 (5er1) and 152 (2er7) atoms and up to 35 rotatable bonds. Starting for both with the receptor-bound conformation, RIGFIT operates convincingly. Taking 2er7, the largest ligand in the data set, as a reference, the remaining inhibitors can be flexibly superimposed with rms deviations below 2.0 Å. In all cases, a section of the backbone has been used as the base fragment. The flexible alignment of 2er7 onto the smaller ligands as reference fails mainly due to size differences. Using 5er2, the second largest inhibitor (144 atoms) in the set, as a reference of the fit yields satisfactory results only for 4er2. The large rmsd obtained for the superposition of 5er1 onto 5er2 can be explained by the substantial

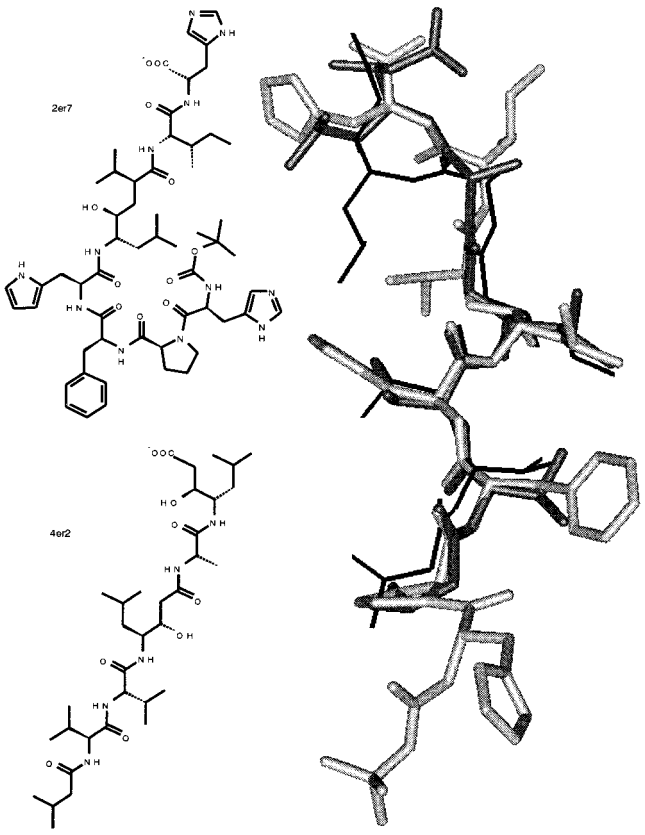

Figure 7. Superposition of two endothiapepsin ligands: 4er2 (dark gray; rank 3, 2.01 Å rmsd) onto the reference ligand 2er7 (light gray) is shown together with the observed binding geometry of 4er2 (black, lines). Despite 2 Å rmsd, the overall fit is still convincing, except at the C-terminus.

Table 10. Endothiapepsin[a]

RL \ TL	2er7	4er1	4er2	5er1	5er2
2er7	0.55 (1) 0.55 (1) 100.0% 0.01 (1)	1.48 (140) 100.0%		1.47 (5) 1.41 (62) 100.0% 9.63 (1)	1.41 (7) 1.39 (29) 100.0% 0.37 (1)
4er1		0.92 (1) 0.92 (1) 100.0% 0.0 (1)			
4er2			0.53 (2) 0.53 (2) 100.0% 0.0 (1)		
5er1				0.76 (9) 0.69 (11) 100.0% 0.0 (1)	
5er2			1.46 (8) 1.37 (29) 100.0% 0.20 (1)		0.89 (1) 0.89 (1) 100.0% 0.08 (1)

[a] Sequence of values within each entry: BEST-10, BEST-ALL, PARTIAL-GOOD, RIGID.

size difference between the two inhibitors. 5er1 is fitted with reverse orientation onto 5er2 resulting in a large rmsd. Because of the size of the ligands, even at larger rms deviations, reasonable alignments are produced (cf. example 4er2/2er7, rank 3, 2.01 Å rmsd, illustrated in Figure 7) (Table 10).

2. HIV-Protease. As in the previous case, large inhibitors were considered in this test example comprising between 88 (4phv) and 134 (8hvp) atoms. Furthermore, the C_2 symmetry of the enzyme binding site imposes additional complications. Only a few of the ligand topologies exhibit C_2 symmetry (e.g., 1hos, 4phv, 9hvp). Applying the RIGFIT method based on the crystallographically given conformations reveals quite con-

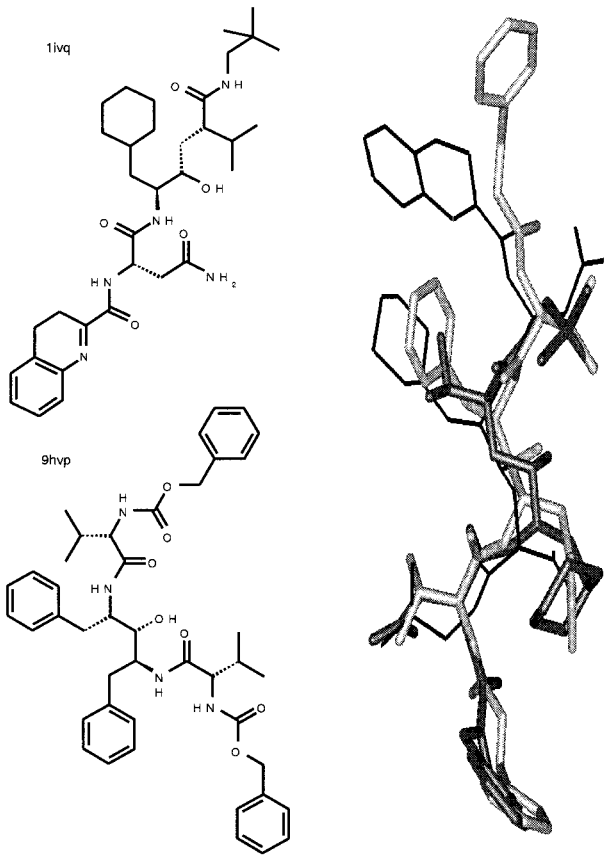

Figure 8. During alignment FLEXS selects a 'symmetry-related' alignment for the two HIV-protease inhibitors (9hvp, light gray; 1ivq, black, lines). The calculated reverse orientation of 1ivq (dark gray) has an rmsd of 11.05 Å.

vincing results. Taking into account that the 'symmetry-related' alignment always results in rms deviations of approximately 10−12 Å, a variety of test cases can still be considered reasonably placed (cf. example 1ivq/9hvp, rank 1, 11.05 Å rmsd, illustrated in Figure 8). The size of the molecules considered either in the HIV or endothiapepsin data set demonstrates the limitation of our flexible alignment strategy (Table 11).

3. Elastase. The seven ligands inhibiting elastase are all of the size of tripeptides. Whereas 1ela to 1ele are of peptidic nature, t0035 and t0036 contain different heterocycles derived from thiophene. This data set represents an excellent example for alternative binding modes adopted by structurally related inhibitors.[46] The three inhibitors 1ela, 1eld, and 1ele occupy the different specificity pockets S1−S4 of the serine protease with structurally comparable molecular portions. Accordingly, the flexible superposition method can reproduce these examples reasonably well. For the remaining pairs less convincing results are obtained. However, it can be questioned whether relevant alignments can be expected at all if the experiment shows that alternative binding modes are adopted by structurally related ligands.[47] This example points out the limitation of flexible superposition techniques in general (Table 12).

4. Thermolysin. The 12 inhibitors binding to thermolysin vary substantially in size (4tln, 24 atoms; 1tlp, 69 atoms) and chemical composition. Inhibitors of comparable size adopting similar binding modes such as 1tlp/1tmn or thior/rthior can be fitted successfully. However, in cases with alternative binding modes,[48] e.g., cbz/ppp, the alignment procedure fails. Some of the smaller inhibitors (2tmn, 3tmn, 4tln) occupy only part of the volume accommodated by the substantially larger ligands (1tlp, 1tmn, 4tmn, 5tmn). In most cases, superpositions of the smaller compounds onto references, substantially larger in size, work reasonably well. However, 5tln which occupies the volume of 4tmn only in part (S1′ and S2′ pockets), cannot be aligned with the latter ligand as the reference. Interestingly, the fit is successful using the given binding site conformations of both ligands. Most of the examples in this data set exhibit in their crystallographically observed orientation a relative intersection volume close to the requested minimum of 60%. On the one hand, this demonstrates the limitations of our approach. On the other hand, for these cases the advantages of our novel placement strategy become obvious (cf. Variable Sequence Construction, Figure 14, Algorithmic Section) (Table 13).

Table 11. HIV-Protease[a]

TL \ RL	1hiv	1hos	1ivp	1ivq	2mip	4hvp	4phv	5hvp	8hvp	9hvp
1hiv	0.76 (7) 0.76 (7) 100.0% 0.0 (1)			1.15 (5) 0.95 (28) 100.0% 0.63 (1)						
1hos			1.46 (3) 1.44 (69) 70.8% 0.72 (1)							1.93 (7) 1.88 (26) 85.5% 0.45 (1)
1ivp		1.99 (2) 1.99 (23) 59.1% 0.65 (2)	0.58 (2) 0.58 (2) 100.0% 0.0 (1)				1.48 (5) 1.48 (5) 100.0% 0.66 (1)			
1ivq				0.62 (1) 0.62 (1) 100.0% 0.0 (1)						
2mip		1.85 (4) 1.84 (22) 67.6% 1.00 (1)		1.73 (7) 1.43 (213) 81.2% 1.09 (1)	0.69 (1) 0.66 (18) 100.0% 0.0 (1)					
4hvp						0.99 (1) 0.99 (32) 100.0% 0.0 (1)				1.75 (3) 1.68 (68) 95.5% 0.40 (1)
4phv							1.31 (3) 1.06 (12) 100.0% 0.03 (1)			
5hvp								0.57 (1) 0.57 (1) 100.0% 0.0 (1)		
8hvp									0.58 (10) 0.57 (22) 100.0% 0.02 (1)	
9hvp							1.79 (7) 1.57 (255) 94.1% 1.08 (2)			0.87 (1) 0.87 (1) 100.0% 0.06 (1)

[a] Sequence of values within each entry: BEST-10, BEST-ALL, PARTIAL-GOOD, RIGID.

Table 12. Elastase[a]

TL \ RL	1ela	1eld	1ele	1elb	1elc	t0035	t0036
1ela	0.62 (9) 0.62 (9) 100.0% 0.0 (1)	1.09 (2) 0.91 (90) 100.0% 0.99 (1)	0.69 (2) 0.65 (168) 100.0% 0.33 (1)				
1eld	1.01 (1) 1.01 (1) 100.0% 0.59 (1)	0.93 (6) 0.86 (100) 100.0% 0.80 (1)	0.98 (4) 0.73 (94) 100.0% 0.55 (1)				
1ele	1.33 (6) 0.96 (44) 100.0% 0.31 (1)	1.35 (5) 1.28 (57) 100.0% 1.01 (1)	0.37 (3) 0.37 (3) 100.0% 0.08 (1)				
1elb				0.78 (3) 0.78 (3) 100.0% 0.0 (1)			
1elc					0.99 (7) 0.94 (14) 100.0% 0.0 (1)		
t0035						0.74 (1) 0.74 (1) 100.0% 0.0 (1)	
t0036							0.56 (2) 0.56 (2) 100.0% 0.0 (1)

[a] Sequence of values within each entry: BEST-10, BEST-ALL, PARTIAL-GOOD, RIGID.

Table 13. Thermolysin[a]

TL / RL	1tlp	1tmn	2tmn	3tmn	4tln	4tmn	5tln	5tmn	thior	rthior	cbz	ppp
1tlp	1.15 (2) 0.86 (18) 100.0% 0.0 (1)	1.41 (6) 1.24 (71) 100.0% 0.78 (1)	0.79 (6) 0.79 (6) 100.0% 0.70 (1)	1.13 (4) 0.95 (17) 100.0% 0.93 (1)	1.67 (2) 1.43 (31) 100.0% 1.05 (1)				1.03 (10) 0.81 (65) 100.0% 0.79 (1)			
1tmn	1.58 (3) 1.34 (47) 100.0% 0.78 (1)	0.73 (1) 0.72 (17) 100.0% 0.0 (1)	0.63 (2) 0.63 (2) 100.0% 0.27 (1)	0.98 (3) 0.78 (304) 100.0% 0.86 (1)	1.77 (1) 1.72 (283) 0.0% 1.30 (1)				0.76 (6) 0.75 (44) 100.0% 0.72 (1)			
2tmn			0.37 (1) 0.37 (1) 100.0% 0.0 (1)		1.82 (3) 1.25 (31) 0.0% 1.19 (1)				0.92 (10) 0.92 (10) 100.0% 0.92 (1)			
3tmn				1.25 (9) 1.15 (12) 100.0% 0.0 (1)								
4tln			1.71 (1) 0.97 (316) 100.0% 1.08 (1)		0.94 (10) 0.82 (26) 100.0% 0.0 (1)							
4tmn		1.67 (9) 1.61 (220) 92.3% 1.43 (1)	0.72 (2) 0.72 (2) 100.0% 0.31 (1)		1.22 (3) 1.16 (158) 66.7% 1.05 (2)	0.53 (1) 0.53 (1) 100.0% 0.0 (1)				1.45 (7) 0.99 (105) 100.0% 1.05 (1)		1.17 (10) 1.07 (22) 100.0% 0.83 (1)
5tln			1.02 (1) 0.91 (126) 100.0% 1.24 (1)		1.76 (6) 0.83 (304) 100.0% 1.33 (1)		0.69 (3) 0.64 (13) 100.0% 0.0 (1)		0.96 (1) 0.96 (1) 100.0% 1.00 (2)	0.83 (7) 0.83 (7) 100.0% 0.99 (1)		
5tmn			0.51 (2) 0.51 (2) 100.0% 0.37 (1)	0.75 (7) 0.75 (7) 100.0% 0.38 (1)	1.37 (1) 0.85 (97) 100.0% 1.16 (2)			0.60 (1) 0.60 (1) 100.0% 0.09 (1)	0.81 (3) 0.73 (39) 100.0% 1.76 (1)	0.85 (2) 0.70 (13) 100.0% 1.33 (1)		
thior							1.73 (6) 1.34 (37) 100.0% 1.04 (1)		0.44 (1) 0.44 (1) 100.0% 0.05 (1)	0.71 (4) 0.71 (4) 100.0% 0.29 (1)		
rthior									0.59 (1) 0.59 (1) 100.0% 0.29 (1)	0.52 (1) 0.52 (1) 100.0% 0.0 (1)		
cbz											1.28 (3) 1.19 (16) 100.0% 0.0 (1)	
ppp												0.63 (1) 0.63 (1) 100.0% 0.05 (1)

[a] Sequence of values within each entry: BEST-10, BEST-ALL, PARTIAL-GOOD, RIGID.

Table 14. Carboxypeptidase A[a]

TL / RL	1cbx	2ctc	3cpa	6cpa	7cpa
1cbx	0.44 (1) 0.44 (1) 100.0% 0.0 (1)	0.55 (2) 0.55 (2) 100.0% 1.94 (2)	1.42 (3) 1.38 (236) 100.0% 1.19 (2)		
2ctc	1.08 (1) 1.04 (37) 100.0% 0.64 (1)	0.32 (1) 0.32 (1) 100.0% 0.0 (1)	1.42 (6) 1.12 (39) 100.0% 1.12 (1)		
3cpa	0.98 (1) 0.76 (78) 100.0% 1.01 (3)	0.96 (10) 0.84 (33) 100.0% 1.09 (1)	0.36 (1) 0.35 (30) 100.0% 0.0 (1)		
6cpa	0.78 (7) 0.78 (7) 100.0% 0.25 (1)	0.77 (3) 0.77 (3) 100.0% 0.48 (1)	1.15 (9) 0.95 (185) 100.0% 1.07 (1)	1.31 (8) 1.31 (8) 100.0% 0.03 (1)	1.84 (1) 0.93 (19) 100.0% 0.27 (1)
7cpa	0.88 (1) 0.88 (1) 100.0% 0.25 (1)	1.01 (3) 1.01 (3) 100.0% 0.63 (1)	1.30 (2) 0.86 (228) 100.0% 1.21 (1)	0.75 (4) 0.69 (13) 100.0% 0.25 (1)	1.25 (2) 1.25 (2) 100.0% 0.03 (1)

[a] Sequence of values within each entry: BEST-10, BEST-ALL, PARTIAL-GOOD, RIGID.

5. Carboxypeptidase A. The five inhibitors binding to carboxypeptidase A span a range between 21 and 74 atoms. However, the ligands are quite similar in chemical composition. Accordingly, across the whole data set convincing mutual alignments can be achieved. This example demonstrates under which conditions the flexible alignment approach can be expected to yield reliable results (Table 14).

Superposition of Medium-Sized Molecules. This subset of examples contains a variety of structures with sizes and flexibilities in the range of usual drug molecules. Most of these ligands are far too flexible to be handled by rigid superposition methods and require an approach such as FLEXS.

1. Thrombin. The four thrombin inhibitors are quite similar in size (59−71 atoms). However, this data set also has several challenges. First of all, the functional groups forming a salt bridge to Asp 189 in the specificity pocket of the enzyme deviate substantially in orientation from a simple atom-by-atom superposition. This is due to deviating topologies in the benzamidine or guanidine fragments (cf. Figure 1, above). A second complication arises from the folded conformation adopted by the two hydrophobic side chains of the ligands at the binding site. Nevertheless, quite convincing superpositions are generated by FLEXS. The alignment of argatroban (1dwd) onto 4tapap shows some problems both when using RIGFIT based on given conformations and when performing a completely flexible fitting (Table 15).

2. Dihydrofolate Reductase. The two rather large ligands dihydrofolate (1dhf) and methotrexate (4dfr) are convincingly fitted by the flexible alignment procedure. The actual binding modes of the two fused heterocycles in both ligands deviate by a ring flip of 180°. This is accurately reproduced by FLEXS as illustrated in Figure 9 (Table 16).

3. Human Rhinovirus. Eight antiviral compounds binding to the coating protein of human rhinovirus have been considered. All ligands are composed of two heterocycles at both terminal ends and an extended aromatic/aliphatic spacer between these moieties. This example is another interesting case of alternative binding modes. The data set is grouped into two subsets both showing quite similar binding modes among each other. The modes between the two sets differ by a reversed orientation of the entire molecules. This behavior can be rationalized via a symmetrical pattern of physicochemical properties along the long molecular

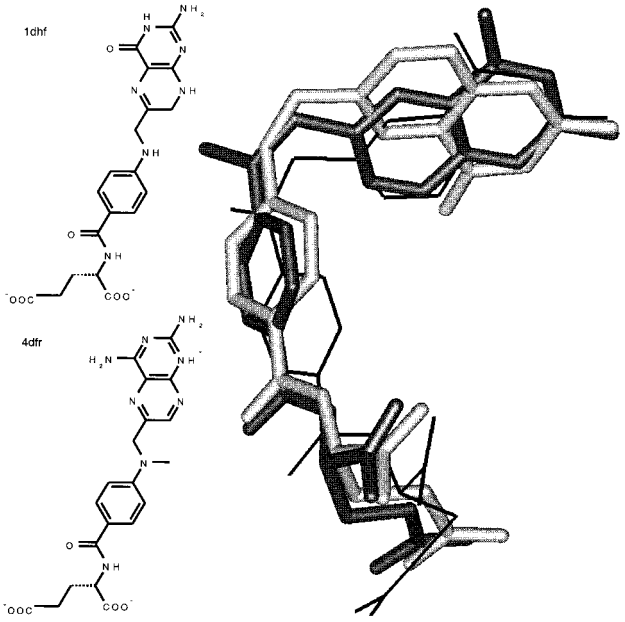

Figure 9. Dihydrofolate reductase ligands with a flexible alignment (dark gray) of methotrexate (4dfr, observed geometry, black, lines) onto dihydrofolate (1dhf, light gray). The prediction at rank 4 deviates by 1.39 Å from the observed geometry. Deviations are mainly found in the central part, supposedly due to the poor overlap between reference and test ligands in this region.

Table 15. Thrombin[a]

TL \ RL	1dwc	1dwd	3tapap	4tapap
1dwc	0.60 (1) 0.60 (1) 100.0% 0.0 (1)	1.55 (3) 1.44 (27) 100.0% 1.25 (1)	1.83 (16) 1.83 (16) 12.5% 0.44 (1)	1.37 (7) 1.26 (43) 100.0% 1.33 (2)
1dwd	1.88 (2) 1.44 (64) 84.6% 1.31 (2)	0.45 (1) 0.45 (1) 100.0% 0.0 (1)	0.81 (2) 0.81 (2) 100.0% 1.35 (1)	
3tapap	1.42 (6) 1.10 (56) 100.0% 0.42 (2)	1.41 (4) 1.12 (18) 100.0% 1.39 (1)	0.82 (3) 0.82 (3) 100.0% 0.01 (1)	1.76 (3) 1.52 (192) 87.5% 1.32 (1)
4tapap	1.87 (2) 1.33 (115) 100.0% 1.02 (1)	1.53 (9) 1.42 (58) 84.6% 1.09 (1)	1.61 (4) 1.61 (4) 50.0% 1.04 (1)	0.65 (2) 0.65 (2) 100.0% 0.78 (1)

[a] Sequence of values within each entry: BEST-10, BEST-ALL, PARTIAL-GOOD, RIGID.

Table 16. Dihydrofolate Reductase[a]

TL \ RL	1dhf	4dfr
1dhf	0.48 (1) 0.48 (1) 100.0% 0.0 (1)	1.39 (4) 1.36 (38) 100.0% 3.11 (1)
4dfr	1.68 (4) 1.47 (91) 100.0% 1.39 (1)	0.46 (2) 0.46 (2) 100.0% 2.71 (1)

[a] Sequence of values within each entry: BEST-10, BEST-ALL, PARTIAL-GOOD, RIGID.

axis.[19] The table reflects this behavior. Molecules adopting similar binding modes are aligned well by FLEXS. Among the two subsets rms deviations of 12−14 Å are found corresponding to the reversed orientation of the ligands. FLEXS does not detect the alternative orientation as a second possible solution of the alignment problem. In contrast, both solutions are observed if the RIGFIT method is applied to the crystallographically given conformations (Table 17).

4. Fructose Bisphosphatase. Two ligands, adenosine monophosphate and a ligand with a substituted

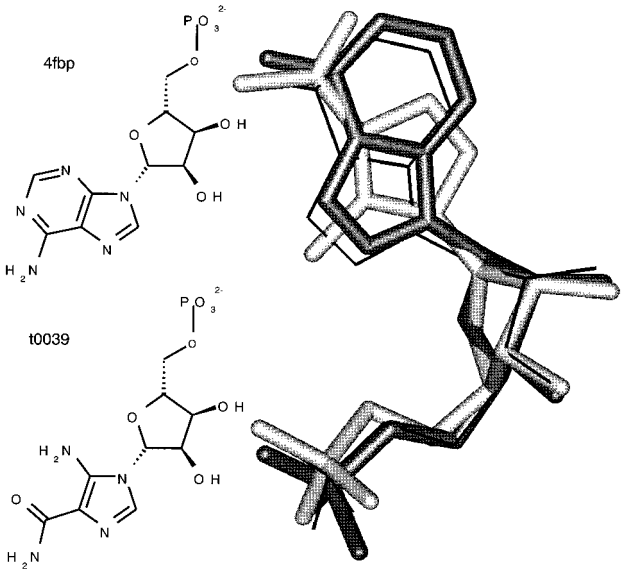

Figure 10. Flexible fit (dark gray) of two fructose bisphosphatase ligands (t0039, light gray; 4fbp, black, lines) reveals an rmsd of 0.61 Å at rank 2.

Table 17. Human Rhinovirus[a]

TL \ RL	2r04	2r06	2r07	2rs5	2rs1	2rr1	2rs3	2rm2
2r04	0.89 (2) 0.67 (59) 100.0% 0.0 (1)	1.51 (8) 1.39 (184) 100.0% 0.32 (1)	1.79 (9) 1.72 (222) 87.5% 0.70 (1)	1.58 (1) 1.28 (121) 100.0% 0.38 (1)				
2r06	0.89 (9) 0.78 (104) 100.0% 0.41 (1)	0.44 (3) 0.44 (3) 100.0% 0.0 (1)	0.75 (8) 0.53 (206) 100.0% 0.35 (1)	0.67 (6) 0.60 (190) 100.0% 0.51 (1)				
2r07	0.87 (4) 0.78 (13) 100.0% 0.70 (1)	0.72 (1) 0.72 (1) 100.0% 0.35 (1)	0.43 (1) 0.43 (1) 100.0% 0.0 (1)	0.87 (10) 0.87 (10) 100.0% 0.73 (1)				
2rs5	0.99 (4) 0.88 (235) 100.0% 0.34 (1)	0.70 (2) 0.52 (69) 100.0% 0.51 (1)	1.06 (7) 0.75 (119) 100.0% 0.73 (1)	0.64 (4) 0.64 (4) 100.0% 0.0 (1)				
2rs1					0.48 (5) 0.48 (5) 100.0% 0.0 (1)	0.49 (6) 0.49 (6) 100.0% 0.16 (1)	0.63 (9) 0.51 (13) 100.0% 0.16 (1)	0.58 (1) 0.58 (1) 100.0% 0.28 (1)
2rr1					0.58 (6) 0.58 (6) 100.0% 0.16 (1)	0.73 (8) 0.73 (8) 100.0% 0.0 (1)	0.69 (3) 0.68 (14) 100.0% 0.29 (1)	0.66 (2) 0.66 (2) 100.0% 0.32 (1)
2rs3					0.61 (8) 0.61 (8) 100.0% 0.16 (1)	0.99 (1) 0.90 (19) 100.0% 0.29 (1)	0.69 (3) 0.69 (3) 100.0% 0.0 (1)	0.65 (3) 0.65 (3) 100.0% 0.33 (1)
2rm2					0.50 (3) 0.50 (3) 100.0% 0.28 (1)	0.70 (4) 0.69 (46) 100.0% 0.32 (1)	0.65 (8) 0.64 (54) 100.0% 0.35 (1)	0.50 (1) 0.50 (1) 100.0% 0.0 (1)

[a] Sequence of values within each entry: BEST-10, BEST-ALL, PARTIAL-GOOD, RIGID.

imidazole instead of the purine moiety, are compared. Using the flexible alignment, an approximate superposition is produced in one case and a convincing fit in the other case. On the basis of the crystallographically determined conformations, RIGFIT performs excellently. Figure 10 shows the alignment of 4fbp onto t0039 with an rmsd of 0.61 Å at rank 2 (Table 18).

Conclusion and Outlook

FLEXS is an approach to flexible ligand fitting based on combinatorial principles. Discrete optimization techniques have been combined with numerical optimization methods. The approach presented here is fast and applicable to various problem settings ranging from interactive usage up to database screening. Considering the prerequisite of a given reference ligand geometry and the inherent limitation of any superposition method,

Table 18. Fructose Bisphosphatase[a]

RL \ TL	4fbp	t0039
4fbp	0.51 (1) 0.51 (1) 100.0% 0.0 (1)	1.75 (1) 1.75 (1) 100.0% 0.20 (1)
t0039	0.61 (2) 0.61 (2) 100.0% 0.17 (2)	1.51 (2) 1.51 (2) 71.4% 0.0 (1)

[a] Sequence of values within each entry: BEST-10, BEST-ALL, PARTIAL-GOOD, RIGID.

FLEXS has proven to produce reasonable superpositions and can successsfully be applied to the search for novel ligands in terms of virtual screening. With respect to a test set of 284 experimentally given alignments, 60% of the examples can be reproduced with an rmsd below 1.5 Å. This figure should be compared to similar validation studies performed for docking methods that achieved a reproduction rate of about 70%.[3]

The general scope of the method has been extended significantly by the introduction of various new methodologies and a systematic parameter adjustment. Especially the calibration of coefficients for the different contributions used in the scoring function has improved the results and in particular the ranking of good candidate solutions.

There is still a discrepancy between the accuracy of the geometries computed and the associated ranking of the obtained solutions, as obvious by comparing rows 1 and 2 in the various results tables. Often enough convincing geometries are detected; however, in many cases an inappropriate scoring is attributed. The same phenomenon is well-known from docking.[45] Further studies have to show how an empirical similarity scoring can be expressed to better discriminate the experimentally observed solution from computer-generated superpositions.

Additional aspects to be implemented in further investigations are the simultaneous consideration of multiple ligands to be superimposed and the incorporation of at least moderate conformational flexibility for the reference ligand structure. The goal at this stage is to be able to exploit additional information inherently present in sets of reference ligands, showing conformational flexibility in complementary molecular portions.

Acknowledgment. We thank our colleagues Bernd Kramer and Matthias Rarey for stimulating discussions, for preparing parts of the input data, and for contributions to the hydrophobicity classification scheme. Alexander Zien provided the calibration methods and Claus Hiller implemented the RIGFIT optimizer. This work has been performed as part of the RELIWE-Project, which was partially funded by the German Federal Ministry for Education, Science, Research, and Technology (BMBF) under Grant No. 01 IB 302 A. The cooperation partners were BASF AG (Ludwigshafen), E. Merck (Darmstadt), EMBL (Heidelberg), and the GMD institutes IPSI (Institute for Integrated Publication and Information Systems, Darmstadt) and SCAI (Institute for Algorithms and Scientific Computing, Sankt Augustin). We are grateful to all our partners in the project for excellent cooperation, especially to Jens Sadowski, Thomas Mietzner, and Hugo Kubinyi from BASF AG.

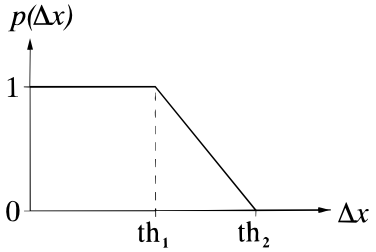

Figure 11. Penalty function p is used to scale down a contribution depending on a certain deviation Δx.

Table 19. Penalty Function Thresholds[a]

	HB	II	HI phenyl	HI amide
ΔS_m	4.7	8.3	0.7	0.7
$th_1(\Delta r)$	0.3 Å	0.3 Å		
$th_2(\Delta r)$	0.7 Å	0.7 Å		
$th_1(\Delta\alpha)$	30°	30°	70°	35°
$th_2(\Delta\alpha)$	80°	80°	90°	90°

[a] S_m gives the contribution of a single match distinguished by type. The thresholds $th_i(\Delta x)$ correspond to the thresholds for the penalty function as illustrated in Figure 11.

Appendix: Scoring and Algorithms

In the following, we describe in more detail the applied scoring scheme and the algorithms used. Additional data and an interface to FLEXS can be found on our web page (http://cartan.gmd.de/FlexS).

Scoring Scheme. The scoring function S is a sum of various terms.

$$S = S_{\text{match}} + \Delta S_{\text{vdW}} \cdot S_{\text{vdW}} + \Delta S_{\text{stm}} \cdot S_{\text{stm}} + \sum_{\text{property } p} \Delta S_p \cdot S_p \quad (1)$$

Here, each term $\Delta S_x \cdot S_x$ is composed of a *coefficient* ΔS_x and a *contribution* S_x, where x is one of the properties considered.

S_{match} is a type of energy term that accounts for paired intermolecular interactions (called *matches*, m).

$$S_{\text{match}} = \sum_{\text{match } m} S_m(\text{type}(m)) \cdot p(\Delta r, \Delta\alpha) \quad (2)$$

Three types of matches m are considered: uncharged hydrogen bonds (HB), ionic interactions (II), and hydrophobic interactions (HI). The geometry of these interactions and the definition of deviations have been described elsewhere.[25] The function p is used to penalize deviations from an ideal geometry.[33] Deviations in length (Δr) and angle ($\Delta\alpha$) are considered. The form of the penalty function is illustrated in Figure 11 for a single argument. Multiple arguments are evaluated separately, and the results are composed multiplicatively. Table 19 summarizes the corresponding thresholds th_1 and th_2.

The contributions S_p in eq 1 correspond to overlap volumes of the different properties expressed by Gaussians.

$$p \in \left\{ \begin{array}{l} \text{electron density } ED, \text{ partial charge} \\ CHG, \text{ hydrophobicity } HYD, \text{ H-bond} \\ \text{acceptor } HA, \text{ H-bond donor } HD \end{array} \right\}$$

S_{vdW} denotes the overlap volume of the atomic van der Waals spheres (*vdW-overlap volume*). S_{stm} is a contribution that quantifies to what extent subtrees match in the test and reference ligand, respectively (*subtree matching*). It is calculated as follows.

During incremental construction of the test ligand, each partial placement involves bonds toward fragments that have not been placed yet (*open bonds*, *OB*). First, the fragment added in the last iteration is searched for such OB. Then, the

reference ligand is inspected for bonds that are closest in space to these OB (*matched bonds*, *MB*). For each such bond b the size of the remaining subtree ST_b in the topological structure of the respective ligand rooted at b, size(ST_b), is evaluated. Then

$$S_{stm} = \sum_{OB\ b\ MB\ b'} 1 - \frac{size(ST_b) - size(ST_{b'})}{max(size(ST_b), size(ST_{b'}))}$$

determines the subtree matching score.

The rationale behind the stm term is to bias the selection of placements toward solutions where the size of the remaining substructure of the test ligand corresponds to the size of the respective part in the reference structure. The hypothesis is that larger portions of the test ligand superpose better with larger portions of the reference ligand and vice versa.

The coefficients ΔS_x, weighting the different contributions S_x to the total score S, have been rigorously calibrated using novel methods.[27] The basic idea is to set the coefficients such as to favor the native solution s^*, thus to fulfill

$$S(s^*) > S(s_i) \quad (3)$$

for as many as possible solutions s_i. Formally this leads to the so-called *pattern recognition problem* which is well-known in machine learning.[34] Following this approach we avoid reproducing superpositions in the inner loop of an optimizer. Instead, we calculate a reasonable sample of superpositions s_i once and store the various contributions $S_x(s_i)$ to the score $S(s_i)$ in a file. Subsequently, we apply the calibration algorithms to this fixed data set, aiming at minimizing the number of violated inequalities (eq 3). Table 20 shows the coefficients that we derived by this procedure.[35] It turned out to be necessary to vary the weighting scheme applied during the different phases of the algorithm. Since the test ligand size increases as the algorithm proceeds, comparisons are performed on different scales. Thus, intermediate changes of the parametrization appear appropriate. Therefore, we utilize different coefficients during base placement, complex construction, and final scoring.

The parameter sets are relatively consistent throughout the different algorithmic phases. As a point of reference the paired intermolecular interactions (accounted for in the S_{match} term) are given a fixed weight of 1. It can be observed that partial charges are the dominating parameter, while van der Waals overlap (S_{vdW}) and electron density (S_{ED}) are almost neglected. It must be assumed that these contributions are unable to discriminate between 'good' and 'bad' solutions, suggested by the algorithm. The change in weights for the Gaussians accounting for hydrogen-bonding potential (S_{HA} and S_{HD}) appear counterproductive, especially since intermolecular interactions are the basic concept in FLEXS to generate placements (which implies that all solutions contain a high number of paired interactions). Nevertheless, it has to be mentioned that our data set could contain a bias with respect to this particular scoring contribution. Clearly, variations of the data set and the calibration methods have to follow, to obtain more robust and further improved scoring functions.

Since the RIGFIT optimizer described below operates in Fourier space as well as in real space, two additional sets of coefficients are required. These coefficients have been adjusted by a brute force grid search in parameter space.

Two Base Placement Methods. We implemented two alternatives for the initial phase of our superposition procedure (base placement). Since different conformers of the base fragment are processed sequentially, both methods perform the placement of a rigid molecular fragment onto the reference ligand.

1. Combinatorial Placement Procedure. The first alternative originates from computer vision where it is called *pose clustering*.[36] This method has been applied successfully in different computational approaches to handle receptor−ligand interactions.[37−39] In our approach, triangles of interaction points (IPs), assigned to the rigid reference structure, are stored in a hash table, and triangles of IPs, corresponding to the base fragment, are enumerated in order to find compatible triangles in both (i.e., triplets of paired intermolecular interactions). Each such pair of triangles determines a unique transformation of the base fragment onto the reference ligand. For each pair of successfully matched intermolecular interactions, we control whether a potential receptor atom would be able to satisfy both interactions (to the reference and to the placed fragment). We further check whether the commonly occupied volume exceeds a user-specified threshold in order to discard irrelevant placements. The placements passing these criteria are clustered in order to reduce their number to a tractable size for the subsequent steps of the algorithm. Furthermore, during clustering, the triplets of paired intermolecular interactions are merged to obtain larger *interaction lists* (see Figure 12).

This kind of approach is efficient and accurate for placing the base fragment. However, it requires at least three putative interaction points belonging to different functional groups that can be matched on both structures. Sometimes ligands are largely hydrophobic and lack a sufficient number of directional interactions (e.g., steroids). As a consequence, we integrated into FLEXS a second alternative placement approach, called RIGFIT.

2. Numerical Placement Procedure. RIGFIT optimizes the common volume of two molecules expressed by various Gaussian functions associated to different physicochemical properties. In this respect, RIGFIT is similar to SEAL[4] which has been extended to consider different chemical properties.[7] However, the optimization strategy in our approach is completely different and has several advantages.[26]

The basic algorithmic idea in RIGFIT originates from X-ray crystallography and uses the concept of the *Patterson function*. Its application to molecular superposition has been described first by Nissink.[40]

One way to approach the well-known *phase problem* in crystal structure determination is to consider Patterson densities instead of real space electron densities. Since the Patterson function contains only information about interatomic distances (the actual spatial location of the atoms is unknown), this description is independent of the translation of the molecule. By transforming the Gaussians to Fourier space and neglecting the phases artificially, we mimic the *molecular replacement* approach of X-ray structure determination and reveal a translation-independent description of the molecules. To compare two molecules, we evaluate the similarity measure proposed by Hodgkin.[41] Since the measure derived is invariant under translation, rotation can be optimized separately.[42] After determining the local optima of the rotation function, we optimize the translation in a second independent step. This

Table 20. Coefficient Sets for the Scoring Function[a]

algorithmic phase	ΔS_{vdW}	ΔS_{ED}	ΔS_{CHG}	ΔS_{HYD}	ΔS_{HA}	ΔS_{HD}	ΔS_{STM}	ΔS_{match}
base placement	33.6	0.2	985.0	21.5	167.7	0.0	0.8	1.0
complex construction	0.0	0.0	236.4	8.7	1.7	0.0	0.1	1.0
final scoring	0.0	0.0	996.8	29.1	0.0	74.3	0.4	1.0
Fourier space optimization		50	200	50	200	200		
real space optimization		30	10000	280	500	950		

[a] The displayed coefficient sets are used during the different algorithmic phases of the flexible superposition method (first three) and of the RIGFIT optimizer (last two). The indices x for the ΔS_x in the headline coincide with the different properties given with eq 1 above.

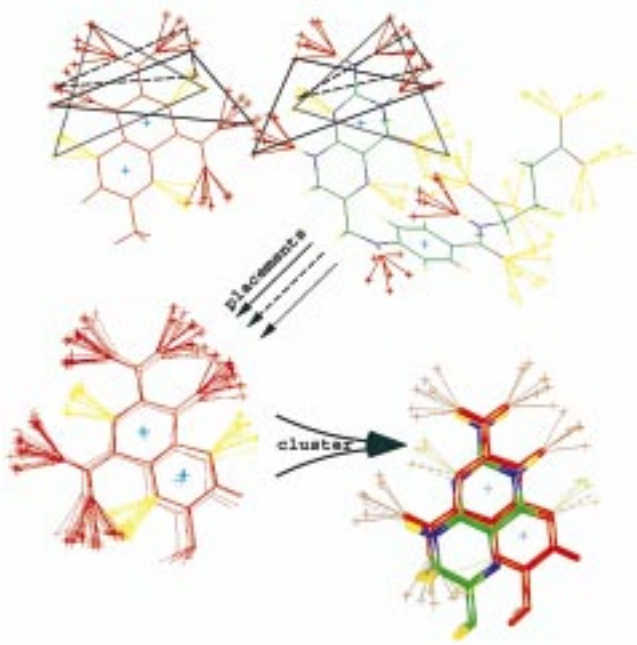

Figure 12. (Top) Base fragment of the test ligand (methotrexate, red) and the reference ligand (folate, atom colored) together with interaction points (used for the discrete approximation of the interaction geometries) and some compatible triangles of interaction points. Each pairing of compatible triangles determines a unique transformation of the base fragment (some examples are shown bottom left). During clustering close alternatives are merged into a single representative placement with potentially larger lists of paired intermolecular interactions (indicated by dashed lines, bottom right).

optimization is carried out in Fourier space efficiently, utilizing the *convolution theorem*.

For the purpose of efficiency, we employ another concept originating from X-ray crystallography. If, as in a crystal, periodic boundary conditions are assumed, the Fourier transform of a real space function becomes a discrete function; i.e., it is nonzero only for integral points in Fourier space (*Laue vectors*). Thus, the computation of an integral simplifies to the summation of function values for the Laue vectors. We benefit from this by treating our molecules as being located in a virtually infinite lattice of replications. Furthermore, if not the entire set of Laue vectors but rather a spherical region around the origin is considered, the high-frequency contributions are removed from the Fourier series. As a consequence, the computational costs for the summation decrease and the scoring function becomes smoother. From technical applications this effect is known as *low-pass filtering*. Of practical importance for us is the possibility to trade off accuracy in the density description against computing speed. This has been demonstrated for three application scenarios.[26]

One major advantage of the optimization in Fourier space is the possibility to divide a 6D search (as performed, e.g., in SEAL) into two successive 3D searches. This inherently speeds up the optimization process. Note that the transformation to Fourier space can easily be performed analytically (since Gaussians remain Gaussians during transformation); thus the computational overhead is small.

Since the superpositions computed in Fourier space are only approximate, we added a 6D postoptimization step in real space. A flowchart for the whole RIGFIT procedure is outlined in Figure 13.

The numerical optimization technique which we employ is a standard quasi-Newton optimizer with BFGS update of the Hessian matrix.[43] We use *quaternions* for the parametrization of the rotation and employ analytically determined first derivatives for the 6D postoptimization in real space. Theo-

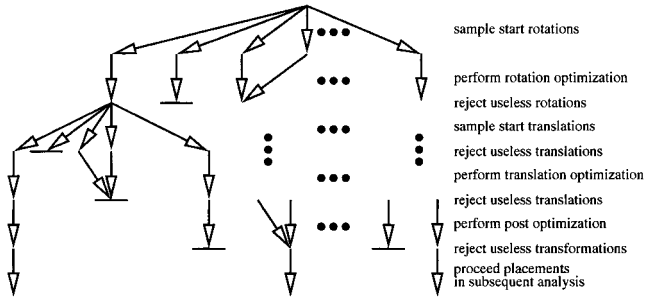

Figure 13. Flowchart of the different optimization steps during a RIGFIT run. The arrows depict changes in the superposition performed by either sampling or local optimization. Dead ends (⏚) show transformations discarded by different filters.

retical evaluation and empirical testing have shown that this strategy is fast and effective.[44]

Variable Sequence Construction. After the base placement has been executed, either by the combinatorial or by the numerical placement procedure, the entire test ligand is constructed by incrementally attaching the remaining fragments. During this algorithmic phase conformational flexibility is considered by connecting the respective fragment in all possible conformations. Through this attachment procedure, in each step, the original set of placements is *expanded* to a larger set of placements of an increased molecular entity. The treelike structure implied by this strategy is pruned at every iteration by selecting the k best placements according to the scoring scheme (cf. above).

Originally, we derived the sequence in which we add the remaining fragments directly from the input structure (*original sequence*). This strategy was adapted from the related docking method FLEXX.[1] There this selection scheme appears appropriate since every fragment has to be accommodated into the binding pocket. In contrast, in structural alignment the test ligand might extend (at least in part) beyond the space occupied by the reference ligand. The hard constraints imposed by the binding site in docking obviously correspond to much softer constraints originating from the occupied volume of the reference ligand in structural superposition.

If we expand an originally reasonable placement of a part of the test ligand into a direction beyond the reference ligand, its score will decrease. This may result in the rejection of the respective placement. In contrast, expanding into a direction that overlaps with the reference more convincingly frequently enables us to retain the respective placement (see Figure 14 for an illustration). For two reasons it is desirable to delay the placement of fragments extending beyond the reference until every other fragment has been connected. First, it is more likely to reveal a reasonable overall placement, and second, the size of the substructures that form reasonable partial placements will increase.

Unfortunately, it is unknown in advance to what extent fragments might expand beyond the reference. We developed a novel complex construction strategy that decides dynamically which fragment to add next. This decision depends on the actual partial placement. Accordingly, each partial placement is associated with a list of *candidate fragments* to be added in the next iteration. Upon placement expansion, FLEXS selects the most appropriate candidate fragment from the list. This evaluation considers (a) the amount of expected overlap with the reference, (b) the number of potential interactions in the candidate fragment, and (c) the size of the substructure tree rooted at the actual candidate fragment (Figure 15 illustrates this situation).

Contribution (a) is computed dynamically by counting the number of reference ligand atoms intersecting the torus that describes the possible positions of the first atom in the candidate fragment. The rationale of this contribution is to delay the processing of fragments that extend beyond the reference ligand. Contributions (b) and (c) are static properties

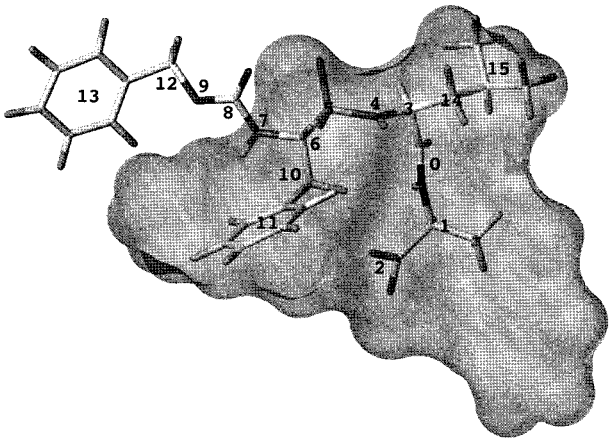

Figure 14. Experimentally observed superposition of two thermolysin ligands (PDB files: 1tlp and 4tmn). The reference ligand 1tlp is represented by its Connolly surface. The test ligand 4tmn is given with the original sequence of fragment numbers. It can be seen that the fragments 8, 9, 12, and 13 do not overlap with the volume of the reference. During variable sequence complex construction, FLEXS adds these fragments last in most of the placements.

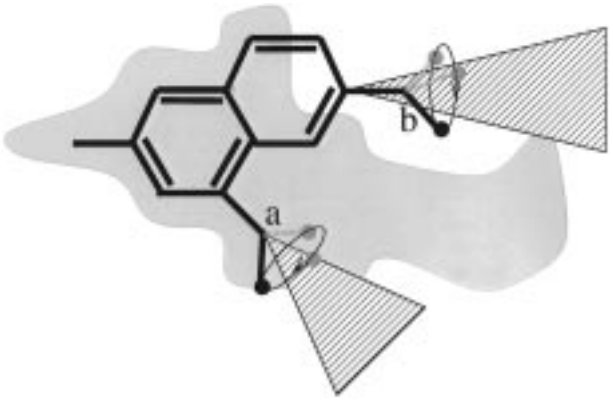

Figure 15. Situation in which we have to decide between two candidate fragments (a and b) to be added. The reference ligand is given as gray-shaded volume, and the test ligand is indicated by its chemical formula. For both candidates, the circular path describing the possible positions of the first atoms of these fragments and the size of the corresponding subtree is sketched. In this situation, the choice would be in favor of fragment (a) since it corresponds to the root of the smaller subtree and its torus overlaps better with the reference ligand.

and can be computed directly from the topology of a structure. These contributions are supposed to prefer small fragments that are capable to form many paired intermolecular interactions.

In every iteration of the incremental construction procedure several choices are possible to select a candidate fragment. Accordingly, the number of structurally distinct placements of the partial ligand (*number of buildup states*, N_{BUS}) will grow. Especially for ligands, composed of many fragments, N_{BUS} can explode. As a consequence, we apply two strategies to prune the construction. Initially, we define a sequence of adding the fragments (*original sequence*). We only deviate from this sequence if the test mentioned above indicates a more appropriate sequence. The threshold, beyond which a placement is accepted to be more appropriate, can be defined by the user. As a second criterion the user can restrict N_{BUS} to an upper limit. If FLEXS exceeds this threshold it returns to the original sequence.

The following severe problem arises during variable sequence construction. Different partial placements have to be scored and compared in order to select the k best solutions for the subsequent iteration step. To resolve this problem we

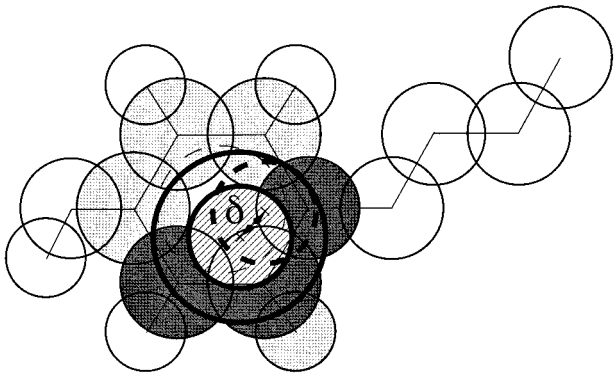

Figure 16. Computation of the overlap for a probe atom (bold circle, hatched) results in a caching list (all gray atoms) and the overlap volume with the intersected atoms (dark gray). To perform this calculation initially, the radius of the probe is enlarged by a locality parameter δ (large bold circle). The caching list then comprises all atoms intersecting with the larger probe. In a second step the overlap volume of the atoms in the caching list with the probe is determined. If in subsequent placements the probe atom is displaced by any distance smaller than δ, the caching list will be reused to determine the actual overlap volume instead of considering the entire set of spheres (thin circles).

enhanced our scoring and selection strategy in two ways. First, we compute an estimate for the total score of the entire ligand even if the placement is only partially performed. This is done by assigning estimated scores to the placements of fragments that are not yet considered in the respective BUS. In the beginning these estimates are simply the scores obtained by superimposing the test ligand onto itself. During complex construction we refine this first guess using the best score obtained so far among all placements for the respective fragment. Second, besides the k best solutions considered further in the subsequent iteration step, for each BUS a certain limited number k_{BUS} of placements is considered additionally.

Computational Shortcuts. During the computation of a superposition the different terms of the scoring function have to be determined quite frequently, typically up to 50 000 times, and the vdW-overlap volume must be computed even more frequently. Therefore special care has to be taken to implement these calculations efficiently. The major improvements in reducing computational costs have been achieved by the following strategies.

1. Caching. Whenever possible, scoring contributions calculated in previous steps are stored and reconsidered. For example, if a fragment is added to the partially placed test ligand during complex construction, we store the results of the overlap tests in special caching lists that depend on a locality parameter δ (see Figure 16). Subsequently, we distinguish between three possible cases. First, if the orientation is not altered (which is the case if no additional paired intermolecular interaction is found), only the overlap terms for the added fragment have to be computed. Second, if the position of the test ligand is altered, but the displacement of the atom under consideration is below δ, we use the data stored in the caching lists rather than entirely recomputing the overlap. The latter has to be done only if none of the above alternatives applies (third case).

2. Cutoff Values. During base placement, the algorithm rejects placements with a vdW-overlap volume below a particular user-defined threshold (currently set to 60%) from hundred thousands of alternatives. Therefore, during the determination of the overlap volume, an upper bound and a lower bound for the overlap volume are continuously tested against this threshold. The upper bound is computed by adding to the overlap volume of the atoms already processed an estimate for the overlap volume that could maximally be achieved by evaluating the overlap of the remaining atoms. The estimate simply comprises the total volume of these atoms,

thus anticipating that they would exactly coincide with the reference structure in the ideal case. The lower bound is given by the overlap volume already computed. If the upper bound falls below the postulated threshold, the computation is terminated and the attempted placement is rejected. If the lower bound exceeds the threshold, the computation is terminated as well and the placement is accepted.

3. Convergence Detection. During numerical optimization in RIGFIT, frequently, starting from different positions, the same local optimum is detected. In addition, an optimization is computationally expensive, especially in the final phase, when strict termination criteria are used in order to achieve accurate solutions. To save computing time, intermediate positions (transformation matrices) are stored in a hash table. During optimization we permanently check whether the actual position is close to one of the stored positions. In this case, the optimization is terminated since it can be assumed that it will converge to the same local optimum. Note that the hash table must not memorize such unfinished optimizations.

References

(1) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. Predicting receptor−ligand interactions by an incremental construction algorithm. *J. Mol. Biol.* **1996**, *261*, 470−489.

(2) Welch, W.; Ruppert, J.; Jain, A. N. Hammerhead: fast, fully automated docking of flexible ligands to protein binding sites. *Chem. Biol.* **1996**, *3*, 449−462.

(3) Jones, G.; Willett, P.; Glen, R. C.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267*, 727−748.

(4) Kearsley, S. K.; Smith, G. M. An alternative method for the alignment of molecular structures: Maximizing electrostatic and steric overlap. *Tetrahedron Comput. Methodol.* **1990**, *3*, 615−633.

(5) Grant, J. A.; Gallardo, M. A.; Pickup, B. T. A fast method of molecular shape comparison: A simple application of a Gaussian description of molecular shape. *J. Comput. Chem.* **1996**, *17*, 1653−1666.

(6) Parretti, M. F.; Kroemer, R. T.; Rothman, J. H.; Richards, W. G. Alignment of molecules by the Monte Carlo optimization of molecular similarity indices. *J. Comput. Chem.* **1997**, *18*, 1344−1353.

(7) Klebe, G.; Mietzner, T.; Weber, F. Different approaches toward an automatic structural alignment of drug molecules: Application to sterol mimics, thrombin and thermolysin inhibitors. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 751−778.

(8) Perkins, T. D. J.; Mills, J. E. J.; Dean, P. M. Molecular surface-volume and property matching to superpose flexible dissimilar molecules. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 479−490.

(9) Mestres, J.; Rohrer, D. C.; Maggiora, G. M. MIMIC: A molecular-field matching program. Exploiting applicability of molecular similarity approaches. *J. Comput. Chem.* **1997**, *18*, 934−954.

(10) Marshall, G. R.; Barry, C. D.; Bosshard, H. D.; Dammkoehler, R. D.; Dunn, D. A. The conformational parameter in drug design: The active analogue approach. In *Computer-Assisted Drug Design*; Olson, E. C., Christoffersen, R. E., Eds.; American Chemical Society: Washington, DC, 1979; Vol. 112, pp 205−222.

(11) Gerber, P. R.; Müller, K. Superimposing several sets of atomic coordinates. *Acta Crystallogr.* **1987**, *A43*, 426−428.

(12) McMartin, C.; Bohacek, R. S. Flexible matching of test ligands to a 3D pharmacophore using a molecular superposition force field: Comparison of predicted and experimental conformations of inhibitors of three enzymes. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 237−250.

(13) Martin, Y. C.; Bures, M. G.; Danaher, E. A.; DeLazzer, J.; Lico, I.; Pavlik, P. A. A fast new approach to pharmacophore mapping and its application to dopaminergic and benzodiazepine agonists. *J. Comput.-Aided Mol. Des.* **1992**, *7*, 83−102.

(14) Kato, Y.; Inoue, A.; Yamada, M.; Tomioka, N.; Itai, A. Automatic superposition of drug molecules based on their common receptor site. *J. Comput.-Aided Mol. Des.* **1992**, *6*, 475−486.

(15) Hurst, T. Flexible 3D searching: The directed tweak technique. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 190−196.

(16) Thorner, D. A.; Wild, D. J.; Willett, P.; Wright, P. M. Similarity searching in files of three-dimensional chemical structures: Flexible field-based searching of molecular electrostatic potentials. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 900−908.

(17) Hahn, M. Three-dimensional shape-based searching of conformationally flexible compounds. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 80−86.

(18) Jones, G.; Willett, P.; Glen, R. C. A genetic algorithm for flexible molecular overlay and pharmacophore elucidation. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 532−549.

(19) Klebe, G. Structural alignment of molecules. In *3D QSAR in Drug Design. Theory, Methods and Applications*; Kubinyi, H., Ed.; ESCOM Science Publishers: Leiden, The Netherlands, 1993; pp 173−199.

(20) Bures, M. G. Recent techniques and applications in pharmacophore mapping. In *Practical application of computer-aided drug design*; Charifson, P. S., Ed.; Marcel Dekker: New York, 1997; pp 39−72.

(21) Leach, A. R.; Kuntz, I. D. Conformational analysis of flexible ligands in macromolecular receptor sites. *J. Comput. Chem.* **1992**, *13*, 730−748.

(22) Moon, J. B.; Howe, W. J. Computer design of bioactive molecules: A method for receptor-based de novo ligand design. *Proteins* **1991**, *11*, 314−328.

(23) Böhm, H.-J. The computer program LUDI: A new method for the de novo design of enzyme inhibitors. *J. Comput.-Aided Mol. Des.* **1992**, *6*, 61−78.

(24) Rotstein, S. H.; Murcko, M. A. GenStar: A method for de novo drug design. *J. Comput.-Aided Mol. Des.* **1993**, *7*, 23−43.

(25) Lemmen, C.; Lengauer, T. Time-efficient flexible superposition of medium-sized molecules. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 357−368.

(26) Lemmen, C.; Hiller, C.; Lengauer, T. RigFit: A new approach to superimpose ligand molecules. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 491−502.

(27) Zien, A.; Lemmen, C.; Lengauer, T. Optimization methods for the calibration of empirical scoring functions. *SIAM J. Optimiz.* **1998**, submitted.

(28) Klebe, G.; Mietzner, T. A fast and efficient method to generate biologically relevant conformations. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 583−606.

(29) Hoflack, J.; De Clercq, P. J. The SCA program: An easy way for the conformational evaluation of polycyclic molecules. *Tetrahedron* **1988**, *44*, 6667.

(30) Gasteiger, J.; Rudolph, C.; Sadowski, J. Automatic generation of 3d-atomic coordinates for organic molecules. *Tetrahedron Comput. Methodol.* **1990**, *3*, 537−547.

(31) Klebe, G. The use of composite crystal-field enviroments in molecular recognition and the de-novo design of protein ligands. *J. Mol. Biol.* **1994**, *237*, 221−235.

(32) Viswanadhan, V. N.; Ghose, A. K.; Revankar, G. R.; Robins, R. K. Atomic physicochemical parameters for three-dimensional structure directed quantitative structure−activity relationships. 4. additional parameters for hydrophobic and dispersive interactions and their application for an automated superposition of certain naturally occurring nucleoside antibiotics. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 163−172.

(33) Böhm, H.-J. The development of a simple empirical scoring function to estimate the binding constant for a protein−ligand complex of known three-dimensional structure. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 243−256.

(34) Bennett, K. P.; Mangasarian, O. L. Robust Linear Programming Discrimination of Two Linearly Inseparable Sets. *Optimiz. Methods Software* **1992**, *1*, 23−34.

(35) Zien, A. Optimierungsmethoden zur Kalibrierung empirischer Bewertungsfunktionen. Master's thesis, University of Bonn, 1997.

(36) Linnainmaa, S.; Harwood, D.; Davis, L. S. Pose determination of a three-dimensional object using triangle pairs. *IEEE Trans. Pattern Anal. Machine Intell.* **1988**, *10*, 634−646.

(37) Rarey, M.; Wefing, S.; Lengauer, T. Placement of medium-sized molecular fragments into active sites of proteins. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 41−54.

(38) Fischer, D.; Lin, S. L.; Wolfson, H. L.; Nussinov, R. A geometry-based suite of molecular docking processes. *J. Mol. Biol.* **1995**, *248*, 459−477.

(39) Lenhof, H.-P. An algorithm for the protein docking problem. In *Bioinformatics: From Nucliec Acids and Proteins to Cell Metabolism*; Schomburg, D., Lessel, U., Eds.; VCH: Weinheim, Germany, 1995; Vol. 18, pp 125−139.

(40) Nissink, J. W. M.; Verdonk, M. L.; Kroon, J.; Mietzner, T.; Klebe, G. Superposition of molecules: Electron density fitting using Fourier transforms. *J. Comput. Chem.* **1997**, *18*, 638−645.

(41) Hodgkin, E. E.; Richards, G. Quantum biologiy symposium 14. *Int. J. Quantum Chem., Quantum Biol. Symp.* **1987**, *14*, 105−110.

(42) Rossmann, M. G.; Blow, D. M. The detection of subunits within the crystallographic asymmetric unit. *Acta Crystallogr.* **1962**, *15*, 24−31.

(43) Dennis, J. E.; Schnabel, R. B. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*; Prentice-Hall: New Jersey, 1983.

(44) Hiller, C. Optimierungsmethoden zum strukturellen Alignment von Ligandmolekülen. Master's Thesis, University of Bonn, 1997.

(45) Dixon, J. S. Evaluation of the CASP2 docking section. *Proteins* **1997**, Suppl. 1, 198−204.

(46) Mattos, C.; Rasmussen, B.; Ding, X.; Petsko, G. A.; Ringe, D. Analogous inhibitors of elastase do not always bind analogously. *Nature Struct. Biol.* **1994**, *1*, 55−58.
(47) Böhm, H.-J.; Klebe, G. What can we learn from molecular recognition in protein−ligand complexes for the design of new drugs? *Angew. Chem., Int. Ed. Engl.* **1996**, *35*, 2588−2614.
(48) Mattos, C.; Ringe, D. Multiple binding modes. In *3D QSAR in Drug Design. Theory, Methods and Applications*; Kubinyi, H., Ed.; ESCOM Science Publishers: Leiden, The Netherlands, 1993; pp 226−254.
(49) Kopple, K. D.; Baures, P. W.; Bean, J. W.; D'Ambrosio, C. A.; Huges, J. L.; Peishoff, C. E.; Eggleston, D. S. Conformation of Arg-Gly-Asp containing heterodetic cyclic peptides: Solution and crystal studies. *J. Am. Chem. Soc.* **1992**, *114*, 9615−9623.
(50) Briem, H.; Kuntz, I. D. Molecular similarity based on DOCK-generated fingerprints. *J. Med. Chem.* **1996**, *39*, 3401−3408.
(51) *MACCS Drug Data Report (MDDR)*; MDL Information Systems Inc., San Leandro, CA.
(52) *SYBYL Molecular Modeling Software Version 6.x*; TRIPOS Associates, Inc., St. Louis, MO; 1994.
(53) Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity − a rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219−3228.
(54) *DAYLIGHT Software Manual*; DAYLIGHT Inc., Mission Viejo, CA; 1994.

JM981037L