Journal of Computer-Aided Molecular Design, **12:** 491–502, 1998.
KLUWER/ESCOM
© 1998 *Kluwer Academic Publishers. Printed in the Netherlands.*

RigFit: A new approach to superimposing ligand molecules

Christian Lemmen*, Claus Hiller & Thomas Lengauer
German National Research Center for Information Technology (GMD), Institute for Algorithms and Scientific Computing (SCAI), Schloß Birlinghoven, D-53754 Sankt Augustin, Germany

Received 28 October 1997; Accepted 11 February 1998

Key words: drug design, Fourier space alignment, molecular similarity, molecular superposition, rigid-body superposition, structural alignment

Summary

If structural knowledge of a receptor under consideration is lacking, drug design approaches focus on similarity or dissimilarity analysis of putative ligands. In this context the mutual ligand superposition is of utmost importance. Methods that are rapid enough to facilitate interactive usage, that allow to process sets of conformers and that enable database screening are of special interest here. The ability to superpose molecular fragments instead of entire molecules has proven to be helpful too. The RIGFIT approach meets these requirements and has several additional advantages. In three distinct test applications, we evaluated how closely we can approximate the observed relative orientation for a set of known crystal structures, we employed RIGFIT as a fragment placement procedure, and we performed a fragment-based database screening. The run time of RIGFIT can be traded off against its accuracy. To be competitive in accuracy with another state-of-the-art alignment tool, with which we compare our method explicitly, computing times of about 6 s per superposition on a common day workstation are required. If longer run times can be afforded the accuracy increases significantly. RIGFIT is part of the flexible superposition software FLEXS which can be accessed on the WWW [http://cartan.gmd.de/FlexS].

Introduction

Quite often a prerequisite to the exploitation of data from a set of ligand molecules binding to the same receptor is the knowledge of the relative orientation that the ligands adopt inside the receptor pocket. If the receptor structure is given, this task can be performed quite effectively by docking methods [1]. Frequently, the receptor structure is not available, however. Then ligand superposition is the method of choice to compute the desired relative orientation. Obviously this is a difficult task, since usually neither do we know in advance the correspondences between key functional groups of the ligands, nor do we have any prealignments that could serve as starting points for local optimizations. Rather, finding the key functional groups and intermolecular correspondences between them, are the kinds of problems that are to be addressed by the subsequent steps of ligand analysis (e.g., by pharmacophore elucidation). Therefore, methods are required that propose a small set of plausible relative orientations for the ligands. These are then further processed by the subsequent analysis. An aggravating fact is that ligands are usually quite flexible and a rigorous analysis has to incorporate molecular flexibility.

The rigid-body superposition method RIGFIT which is described in the present paper has been incorporated into the FLEXS system for flexible ligand superposition [2]. In this context RIGFIT is employed to perform superpositions of sets of conformers either of molecular fragments or of entire molecules. On the basis of the fragment placements generated by RIGFIT, the flexible construction procedure in FLEXS is able to produce reasonable alignments for flexible ligands. RIGFIT is a global optimization strategy which utilizes concepts known from X-ray crystallography, combined with an effective modeling of properties, and an efficient optimizer.

*To whom correspondence should be addressed.

Currently, various methods are available that deal with the ligand superposition problem. A breakthrough in the reduction of computational costs was the SEAL approach [3] which has been extended later [4]. Various combinatorial approaches aiming at enumerating efficiently possible correspondences of chemical features on the ligands [5–7] are computationally quite intensive. The GASP program [8] which employs a genetic algorithm for optimization allows to handle a set of flexible molecules with run times in the range of an hour per problem instance.

More recently, a variety of approaches, aiming at field- or shape-based comparison, have been presented. Quite an efficient van der Waals volume overlap optimization with the aid of Gaussian functions can be found in [9]. The method has been tested successfully on 21 pairs of 15 structurally dissimilar ligands binding to three distinct protein receptor sites. Unfortunately, explicit timings are given only for the initial self-overlap tests. MIMIC [10] is a method that matches steric and electrostatic fields of two molecules. It includes flexibility by considering multiple conformers that are clustered in a preprocessing step. Later, this approach has been extended to multi-molecule alignments [11] which enhances predicting the observed relative orientation for difficult cases. The results presented provide a very detailed analysis, however, they are limited to quite a restricted test set. An approach to 3D pattern recognition via steric and electrostatic alignment based on sphere overlap optimization is presented in [12]. This approach has been tested on proteins as well as on small molecules. Unfortunately neither timings nor rms deviations are tabulated for the different test applications, but only provided for single specific cases in the text. This makes the comparison of results difficult.

Another source of valuable superposition techniques is protein X-ray crystallography [13–15]. A brief survey of recent related approaches in this field is provided after the Methods section since the terminology required is introduced there. Recent reviews to molecular superposition can be found in [16, 17].

Our overall superposition strategy will be detailed in the next section, followed by a description of the physico-chemical modeling and the mathematical foundations of the different similarity indices used as objective functions. Afterwards, the algorithms used for sampling and optimization, are described. Then, we detail the results of three different experiments we carried out with RIGFIT. Finally, we give conclusions and an outlook to future work.

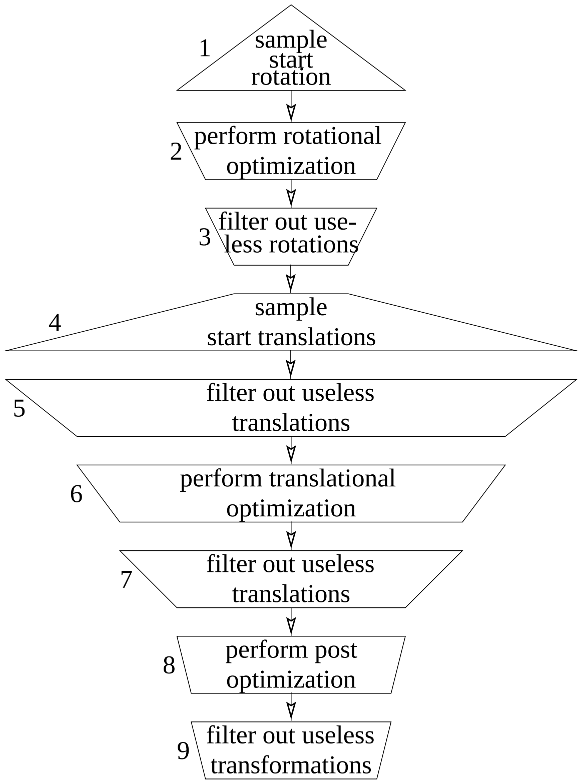
1
sample start rotation
2
perform rotational optimization
3
filter out useless rotations
4
sample start translations
5
filter out useless translations
6
perform translational optimization
7
filter out useless translations
8
perform post optimization
9
filter out useless transformations

Figure 1. Flow chart of the different optimization phases during a RIGFIT run. The lengths of the horizontal lines of the boxes illustrate the increase/decrease of the total number of placements to be processed in the respective phase.

Methods

The overall optimization strategy comprises three phases that are performed separately in the following sequence: rotational optimization, translational optimization and post-optimization. Between these phases several sampling and data reduction steps are performed. The entire procedure is outlined in Figure 1.

Boxes 2, 6, and 8 correspond to the three *phases of optimization*. Each phase comprises a set of *local optimizations* aiming at moving into a local maximum of the objective function in a sequence of *iterations* until a *termination criterion* is fulfilled. The three similarity indices used for the objective functions of the three optimization phases and, especially, the separation of rotational and translational optimization will be detailed after a description of the modeling of the physico-chemical properties in the next section.

Gaussian functions describing chemical features

We utilize Gaussian functions in order to model the various physico-chemical properties of the molecules.

$$\rho(x) = \sum_{\text{Atom } i} a_i e^{-b(x-x_i)^2} \tag{1}$$

The parameter b is identical for every Gaussian, whereas the a_i differ such as to represent the considered property value.

Carbó [18] proposed the following measure for the similarity of two molecules A and B.

$$Z_{AB} = \int \rho_A(x)\rho_B(x)\mathrm{d}x \tag{2}$$

Let us assume molecule B to be rigid but movable, whereas molecule A is kept fixed in space. Then $Z_{AB}(t, \Omega)$ depends on the translation t and rotation Ω of molecule B. With this notation, the similarity index proposed by Hodgkin [19], which we utilize for optimization, reads as

$$H_{AB}(t, \Omega) = \frac{2Z_{AB}(t, \Omega)}{Z_{AA} \cdot Z_{BB}}. \tag{3}$$

Since in rigid-body alignment Z_{AA} and Z_{BB} remain constant during optimization, in the following, we focus on $Z_{AB}(t, \Omega)$. If we substitute the Gaussian representation (Equation (1)) into the similarity measure (Equation (2)), we obtain the SEAL function [3].

$$\begin{aligned} Z_{AB}(t, \Omega) &= \int \sum_{\text{Atoms } i,j} a_i e^{-b(x-x_i)^2} a_j e^{-b(x-x_j)^2} \\ &= \sum_{\text{Atoms } i,j} w_{i,j} e^{-b/2(x_i-x_j)^2} && (4) \\ &= S_{AB}(t, \Omega) && (5) \end{aligned}$$

Modeling different physico-chemical properties extends the SEAL approach and was first introduced by Klebe et al. [4].

$$w_{i,j} = \sum_{\text{property } p} k_p a_{i_p} a_{j_p} \tag{6}$$

Here the k_p denote user-defined weights for the different properties p.

Gaussian functions have several useful properties which make them especially well-suited to represent chemical features of molecules [20]. First, the Fourier transform of a Gaussian function (short a *Gaussian*) is a Gaussian. Second, the absence of a boundary in Gaussians enables the optimization to intercorrelate Gaussians even if their maxima are distant from each other. This, in turn, helps to overcome the problem of non-existing prealignments and the lack of knowledge of the intermolecular correspondence of features displayed by the ligands. Third, derivatives can be computed easily. Finally, Gaussians have the character of a convex potential, i.e. the attraction of two Gaussians increases as their maxima approach each other.

We model the following chemical properties by Gaussians: steric occupancy, partial atomic charge, hydrophobicity, and hydrogen bonding potential.

Figure 2 depicts the observed relative orientation of two ligands binding to fructose bisphosphatase together with the different sets of Gaussians we utilize. It can be seen that, even if the chemical structures of the two molecules deviate significantly (center of the figure), most of the properties represented by Gaussians match convincingly. In fact, Gaussians of either molecule usually appear in pairs of close proximity. However, different molecular fields give rise to matches in different portions of the molecules. Thus, optimizing the alignment using different combinations of the different corresponding similarity measures will result in different alignments [10]. We determined the coefficients weighting in the different similarity measures (cf. below) empirically by a brute force search on a grid in parameter space.

Three phases of optimization in Fourier space and real space

The rotation is optimized independently from the translation by using the Patterson function ρ' instead of the original set of Gaussians ρ [21]. This can be done effectively by transforming the functions into Fourier space [15]. There, the translation of the Gaussians simply results in a change of the so-called phase factor which is multiplied to the Gaussian. In X-ray crystallography the lack of knowledge of the phase factors is responsible for the well-known *phase-problem*. In the present case the intended disregard of the phases separates rotational and translational optimization.

Figure 3 outlines the derivation of the similarity measure used for the rotational optimization. The key factor of this derivation is that using the squares of the absolute values of $\hat{p}$ in P_{AB} has the same effect as if Patterson densities $\hat{\rho}'$ are used instead of the real densities $\hat{\rho}$. Since the Patterson function does not depend

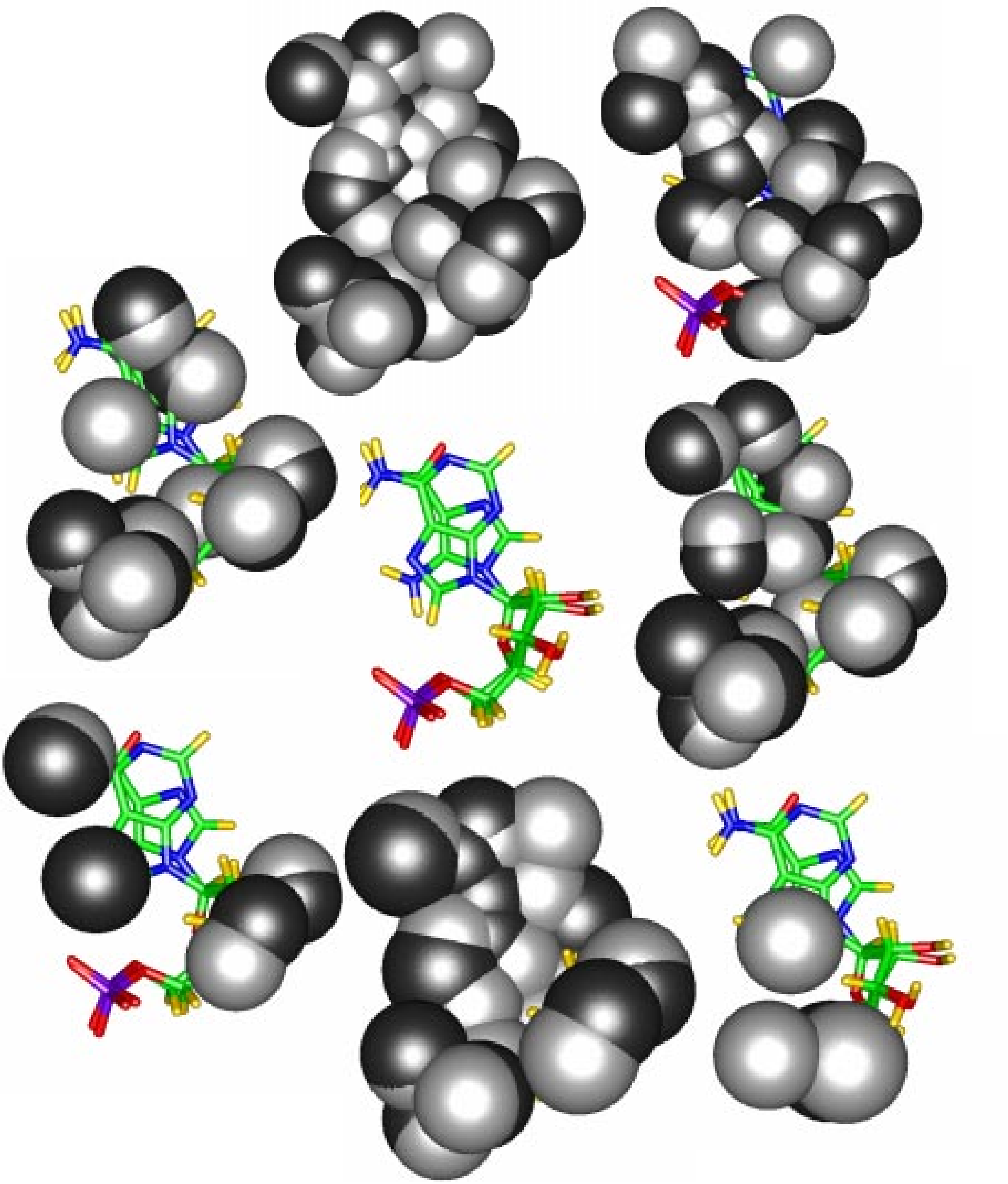

Figure 2. Different Gaussian functions describing the physico-chemical properties of the molecules are utilized in RIGFIT. To illustrate the match of these properties for two molecules binding to the same receptor, the crystallographically observed relative orientation of two fructose-bisphosphatase ligands (extracted from the structures with PDB-code: 4fbp and CASP2-label: t0039, respectively) is shown. The Gaussian functions are truncated to the respective iso-contour surfaces, in order to facilitate visual inspection. For the sake of clarity, positive and negative property values of the Gaussians are displayed separately. In order to distinguish the two molecules, the Gaussians representing 4fbp are given with a dark shading, while the light shaded Gaussians belong to t0039. In clockwise sequence starting at 12 o'clock the properties given are steric occupancy, positive partial charge, negative partial charge, hydrophobicity, lipophilicity, H-bonding acceptor and H-bonding donor potential. The chemical structures of the two molecules are provided in the center of the figure.

Starting with the similarity measure of Carbó:

$$Z_{AB}(t,\Omega) = \int_x \rho_A(x)\rho_{B,t,\Omega}(x)\,\mathrm{d}x$$

Parsevals theorem [32] gives:

$$Z_{AB}(t,\Omega) = \int_h \hat{\rho}_A(h)\hat{\rho}_{B,t,\Omega}(h)\,\mathrm{d}h$$

With periodic density functions, we obtain:

$$Z_{AB}(t,\Omega) = \sum_h \hat{\rho}_A(h)\hat{\rho}_{B,t,\Omega}(h)$$

Using Patterson functions $\hat{\rho}'_A$, $\hat{\rho}'_B$ instead [15]:

$$Z'_{AB}(t,\Omega) = \sum_h |\hat{\rho}_A(h)|^2|\hat{\rho}_{B,t,\Omega}(h)|^2 \qquad (13)$$

$$= \sum_h \hat{\rho}'_A(h)^2\hat{\rho}'_{B,\Omega}(h)^2 = P_{AB}(\Omega) \qquad (14)$$

Molecule representation by Gaussian functions:

$$\hat{\rho}(h) = \sum_{\text{Atom } i} \hat{g}_i(h) = \sum_{\text{Atom } i} A_i e^{-B_i(h-h_i)^2}$$

Extension to various properties:

$$R_{AB}(\Omega) = \sum_{\text{property } p} P^p_{AB}(\Omega)$$

Figure 3. The major steps of the derivation of the similarity measure $R_{AB}(\Omega)$ used for rotational optimization. Here $\hat{\rho}$ denotes the Fourier transformed of ρ. The consideration of the squared absolute values of the densities $\hat{\rho}$ in (13) is tantamount to using the respective Patterson densities $\hat{\rho}'$ in (14). The indices t and Ω denote the translational and rotational dependence of molecules B.

on the translation, P_{AB}, and accordingly R_{AB}, merely depend on the rotation Ω. The asymptotic run time of a single iteration in a local optimization using R_{AB} is given by

$$\mathcal{O}(N_{\text{Lv}} \cdot N_{\text{atoms}} \cdot N_{\text{properties}}), \qquad (7)$$

where N_{Lv} is the number of vectors h in Fourier space (cf. below), N_{atoms} is the number of atoms in the mobile molecule, and $N_{\text{properties}}$ is the number of different fields considered (4 in our case).

After rotational optimization we continue inside Fourier space and optimize the translation by application of the *convolution theorem* [15]. Figure 4 provides a rough description of the derivation of the similarity measure used for the translational optimization. The most important point here is that only the phase factors $e^{i2\pi ht}$ depend on the translation t. Therefore, during the first iteration of the first local optimization, the terms $\hat{\rho}_A(h)\hat{\rho}_B(h)$ are stored in a table indexed by h and reused in the subsequent local optimizations. The asymptotic run time of a single iteration in a local optimization using T_{AB} is given by

$$\mathcal{O}(N_{\text{Lv}} \cdot N_{\text{properties}}). \qquad (8)$$

Starting with the similarity measure of Carbó with fixed Ω:

$$Z_{AB_\Omega}(t) = \int_x \rho_A(x)\rho_{B_\Omega,t}(x)dx = \rho_A(t) * \rho_{B_\Omega}(t)$$

With the convolution theorem [32]:

$$\hat{Z}_{AB_\Omega}(h) = \hat{\rho}_A(h)\hat{\rho}_{B_\Omega}(h)$$

Transformed back to real space:

$$Z_{AB_\Omega}(t) = \int_h \hat{\rho}_A(h)\hat{\rho}_{B_\Omega}(h)e^{i2\pi ht}\,\mathrm{d}h$$

With precomputed $\hat{\rho}(h)$, periodic boundary conditions and extended to various properties:

$$T_{AB_\Omega}(t) = \sum_{\text{property } p}\sum_h \text{Table}(h)e^{i2\pi ht}$$

Figure 4. The major steps of the derivation of the similarity measure $T_{AB}(t)$ used for translational optimization. Here $\hat{Z}$ denotes the Fourier transformed of the similarity measure Z and B_Ω indicates that an arbitrary but fixed rotation Ω is applied to molecule B. The index t denotes the translational dependence of molecule B.

There are four features that make the optimization in Fourier space especially attractive. First, we split a six-dimensional search, such as for example performed in SEAL, into two successive three-dimensional searches, which inherently speeds the optimization. Second, if we artificially impose periodic boundary conditions, the integration of overlapping Gaussians simplifies to a summation over discrete periodic integral points (*Laue vectors*) in Fourier space, which again increases the efficiency of the algorithm. Third, approximations can be computed easily. If the summation (or integration) of overlapping Gaussians is not performed over the entire reciprocal space but restricted to a spherical region around the origin, we effectively remove the high frequency contributions

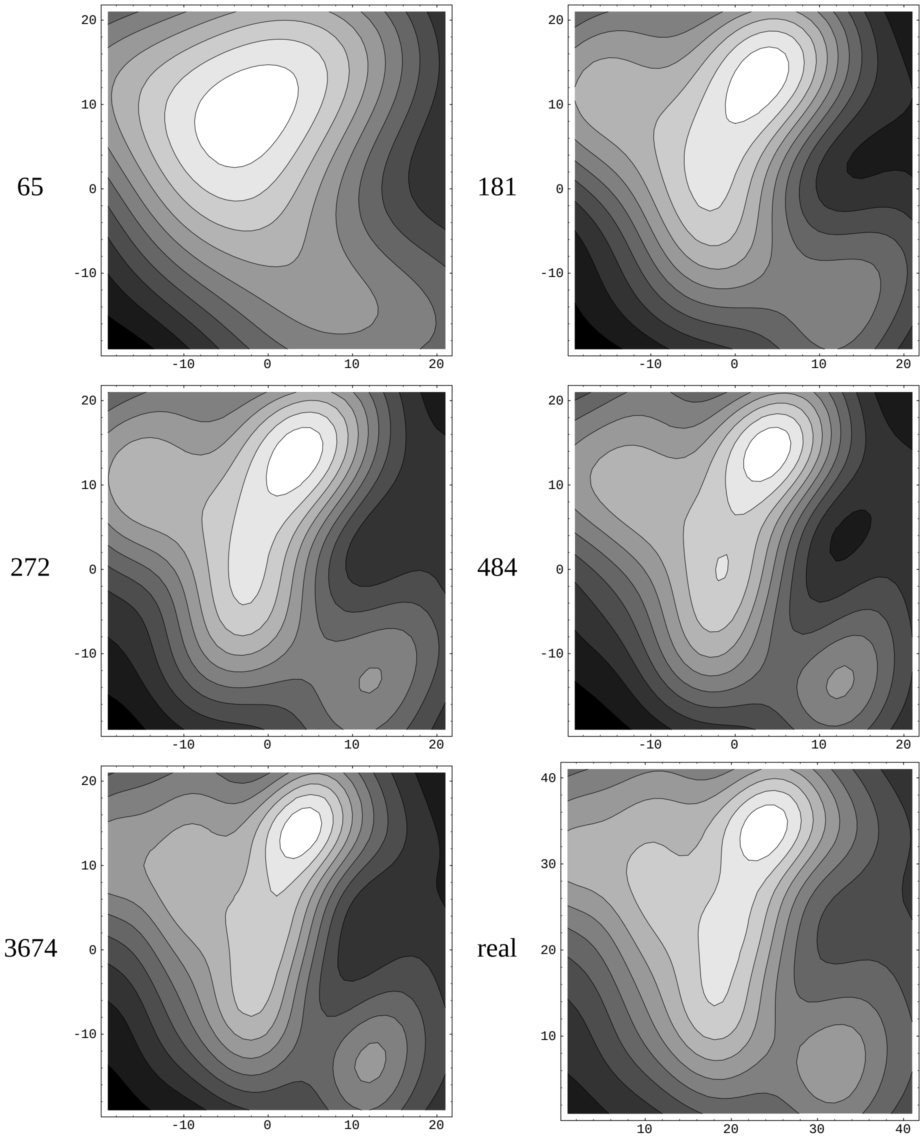

Figure 5. Two-dimensional cuts through the 3D-space of translations at various levels of resolution. The example considered originates from the observed relative orientation of two thermolysin ligands (extracted from the structures with PDB-codes 1tlp and 2tmn, respectively). The translation has been sampled in steps of 0.2 Å in positive and negative direction 4 Å around the origin. The axis labels indicate the number of steps between the origin and each sample point. The number of Laue vectors (N_{Lv}) used to calculate each diagram is indicated left to the figure. It can be seen that on the one hand the landscape becomes smoother with decreasing N_{Lv} and on the other hand with increasing N_{Lv} the approximation ($T_{AB_{\Omega}}$) of the real space similarity measure ($Z_{AB_{\Omega}}$, displayed bottom right) becomes increasingly accurate.

and smooth the function to be optimized [15]. Figure 5 illustrates this effect.

Finally, returning to Figure 1, it can be seen that most of the optimizations have to be performed during translational optimization, since here the number of starting points corresponds to the product of the number of rotational optima and the number of translational sample points. Fortunately, the similarity measure used for translational optimization T_{AB} can be computed especially efficient, as stated in Equation (8). In addition, we included an extra filtering step to avoid useless starting points prior to optimization.

Since the alignments produced by the optimization in Fourier space are only approximate solutions, we perform a real space post-optimization using a six-dimensional search with the extended SEAL function S_{AB} as a third and final optimization phase. The asymptotic run time of a single iteration in a local optimization using S_{AB} is given by

$$\mathcal{O}(N_{\text{atoms}}^2 \cdot N_{\text{properties}}). \tag{9}$$

Details of the mathematical description, the algorithms applied, and the implementation of the approach can be found in [22].

Algorithm

After introducing the global optimization strategy and the different similarity indices used during the different phases of optimization, we now focus on algorithms used for sampling and local optimization.

Rigorous sampling techniques

Preceding the rotational and translational optimization phases we perform a sampling step in order to find starting positions for the local optimizations.

Since using Eulerian angles is inefficient and results in an uneven distribution of angles we utilize the 'optimal sampling of the rotation function' described by Lattman [23]. In this approach Euler angles θ_1, θ_2, and θ_3, are modified to

$$\hat{\theta}_+ := \theta_1 + \theta_3, \quad \hat{\theta}_2 := \theta_2, \quad \hat{\theta}\_ := \theta_1 - \theta_3 \tag{10}$$

and the ranges of these pseudo-orthogonal angles are given by

$$0^\circ \leq \hat{\theta}_+ < 720^\circ, \ 0^\circ \leq \hat{\theta}_2 < 180^\circ, \ 0^\circ \leq \hat{\theta}\_ < 360^\circ. \tag{11}$$

We sample $\hat{\theta}_2$ in the range [0°, 180°] with a fixed stepwidth $\Delta\hat{\theta}_2$ and set $\Delta\hat{\theta}_+$ and $\Delta\hat{\theta}\_$ to

$$\Delta\hat{\theta}_+ = \frac{\Delta\hat{\theta}_2}{\cos\left(\frac{\hat{\theta}_2}{2}\right)}, \quad \Delta\hat{\theta}\_ = \frac{\Delta\hat{\theta}_2}{\sin\left(\frac{\hat{\theta}_2}{2}\right)}. \tag{12}$$

The smaller $\Delta\hat{\theta}_2$, the more efficient this sampling technique becomes. Reasonable results can be achieved with $\Delta\hat{\theta}_2 < 180^\circ$. This corresponds to a minimum of 15 starting orientations, which appears to be a reasonable lower limit. Finer grating in angular space is possible and facilitates more detailed analysis of single alignment pairs.

Using a grid of translation vectors as starting points quickly leads to a large number of local optimizations to be performed. Instead, we move the center of gravity of the movable ligand onto the center of gravity of each so-called *component* of the other ligand. Even if this ligand, as a whole, is treated as rigid, its chemical structure can be used to determine rotatable bonds and ring systems. A component contains at most a single acyclic rotatable bond and ring systems always belong entirely to a single component. This compromise of coverage versus computational effort for the translational sampling yielded the best results during empirical testing. Other alignment methods basically ignore the necessity of sampling translations during starting point generation. Often, simply the centers of gravity are superimposed [3, 9, 10, 15], or translational sampling is performed only in a limited region around the origin [4]. This may be adequate for ligands of roughly the same size. For the placement of molecular fragments onto a usually much larger reference this procedure is absolutely inappropriate.

Efficient optimization technique

The optimization technique we utilize is a quasi-Newton method. The gradient is computed by difference quotients for the similarity indices $R_{AB}(\Omega)$ and $T_{AB}(t)$ used during rotational and translational optimization respectively. The gradient of the objective function based on the similarity measure $S_{AB}(t, \Omega)$ is determined analytically. Trials showed that this combination is most efficient. The termination criterion is fulfilled when, for a sufficient number of iterations, the function value does not change significantly. Empirical testing showed that, if the function value changes by less than 10^3 for two successive iterations, the optimization may be terminated. Applying stricter termination criteria does not reveal any additional local optimum on our test set (cf. below). The corresponding thresholds have been adjusted empirically and can easily be modified by the user. The BFGS-update [24] of the Hessian matrix was found to be the most effective way to approximate the second derivative of the scoring function. For the parameterization of the rotation function we employed *quaternions* in order to avoid the singularities that occur if Euler angles are used instead [25].

Relationship with previous work

Recently, quite an efficient implementation of the molecular replacement method, used in X-ray crystallography [26], has been described in [13]. Our approach benefits from this technique, but rather than fitting a model into a measured electron density, our aim is to superimpose substantially different molecules. Other approaches in the literature, utilizing Fourier space concepts in small molecule superposition [15, 27], solely work on electron density comparison in the calculation and are usually tested on rather small sets of examples. Various approaches to molecular superposition are known which sample rotational space and perform an efficient optimization procedure for the translational part in the inner loop of the sampling [14, 28]. The distinct separation of rotational and translational optimization that we employ is substantially different from these concepts. The extensions to the SEAL approach by Klebe et al. [4], and the method developed by Nissink et al. [15] have been the basis of our work.

Results and discussion

We implemented RIGFIT as part of the FLEXS system for flexible ligand superposition and tested it in three ways. First, we performed rigid-body superpositions for a set of ligand pairs for which the crystal structures of the respective receptor-ligand complexes are given. This enables us to evaluate how closely RIGFIT approximates the observed relative orientation. Second, we employed RIGFIT as a fragment placement procedure in combination with our flexible superposition tool FLEXS. And third, we performed a fragment-based database screening using RIGFIT. All calculations have been performed on a SUN UltraSparcII processor with 250 MHz clock speed.

Rigid-body superposition

We superimposed pairs of rigid ligands for which the *observed relative orientation* can be extracted from experimental data. If both ligands bind to the same protein and the respective complexes have been structurally resolved, the observed relative orientation is derived from an rms-fit of corresponding backbone atoms in the protein structures. We measured the rms deviation between observed and computed relative orientation of the movable ligand's atomic coordinates and compared our results with one of the leading programs, the extended version of the SEAL approach [4]. It turns out that the results are comparable in accuracy while we reduced the run time by about a factor of two. Furthermore, we could improve in accuracy over the SEAL approach if we performed more detailed calculations in Fourier space, then requiring about the same amount of computing time as SEAL. On average, RIGFIT produces reasonable results within a run time of about 10 to 20 s per test case. Our test set contains 161 ligand pairs extracted from 50 protein-ligand complexes. The ligands span a whole range of drug size molecules from 18 to 158 atoms. The set of 9 proteins considered comprises trypsin, endothiapepsin, hiv-protease, dihydrofolate reductase, carboxypeptidase a, thrombin, thermolysin, glycogen phosphorylase, and human rhinovirus. Consult Table 1 for detailed result statistics. Several facts in this table clearly indicate that run time can be traded off against accuracy. The quantity N_{miss} indicates the number of cases where the observed relative orientation scores higher than any of the local optima detected by the algorithm. Apparently such deficiencies result from inaccuracies of the approximation. Indeed, the table shows decreasing N_{miss} with growing N_{Lv}. Another indication for this trade-off are the quantities N_{r_1} and $N_{r_{1-3}}$, giving the number of cases where the placements computed can be considered reasonable. These quantities rise with growing N_{Lv}.

Also, the table indicates that, even in the setting with minimum N_{miss} (Lr=6.0 Å), still about one third of the examples is missed ($\%r_{1-3} = 67.7\%$). There are several explanations for this phenomenon. First, some additional reasonable solutions may be present among the placements at rank 4 and below. Second, it must not be overlooked that the threshold of 1.2 Å is an arbitrary and relatively strict criterion to discriminate between good and bad superpositions. Due to experimental as well as conceptual uncertainties with the chosen type of evaluation, the error in the data considered as reference may well be estimated to be about 0.7 Å [4]. Our test cases span a wide variety of molecules, both in size as well as in chemical composition. Especially for the rather large examples (e.g. the peptidic ligands of hiv-protease and endothiapepsin) rms deviations of about 1.5 to 2.0 Å may also be considered reasonable. Third, in addition to the parameter Lr, there are other user-definable parameters that largely influence the results obtained. The setting we use is clearly a compromise between runtime and accuracy. The number of starting points for rotation optimization, e.g., is currently set to 15. A denser sampling obviously would help to reduce the number of cases where the local optimum found, starting from the observed relative orientation, is missed by the approach (cf. figure N_{miss} in Table 1). Finally, obviously there remain some examples that cannot be tackled by our

Table 1. Rigid-body superposition

Lr	N_{Lv}	ϕ time (s)	N_{r_1}	ϕ r_1	% r_1	$N_{r_{1-3}}$	ϕ r_{1-3}	% r_{1-3}	N_{miss}
2.0	14	6.2	67	0.61	41.6	81	0.64	50.3	37
2.4	31	9.1	73	0.62	45.3	90	0.63	55.9	33
3.0	65	16.3	83	0.63	51.6	102	0.65	63.4	24
3.5	95	23.6	84	0.60	52.2	109	0.64	67.7	19
4.3	181	45.8	85	0.60	52.8	103	0.63	64.0	17
6.0	484	127.2	86	0.61	53.4	109	0.63	67.7	13
8.0	1095	300.3	84	0.60	52.2	105	0.63	65.2	14
SEAL:		<14	67	0.67	41.6	93	0.68	57.8	–

Symbol description: Lr Laue radius; N_{Lv} number of Laue vectors; ϕ time average computation time per instance; N_{r_1} number of cases with an rms deviation <1.2 Å in the first place; ϕ r_1 average rms deviation for the r_1-examples; % r_1 percentage of r_1-examples compared to all; $N_{r_{1-3}}$ number of cases with an rms deviation <1.2 Å among the first three solutions; ϕ r_{1-3} average rms deviation for the r_{1-3}-examples; % r_{1-3} percentage of r_{1-3}-examples compared to all; N_{miss} number of cases where the value of the scoring function is higher for the observed relative orientation than for any optimum detected by RIGFIT.

method. In most cases this will be due to the lack of knowledge about the receptor structure. The presence of multiple binding modes [29] makes this situation even worse. Methods that rely purely on a pairwise comparison of ligand data cannot be expected to reveal results that are as accurate as those recently reported for docking [1, 30]. Despite the above stated difficulties and compared to a hit rate of about 70% recently achieved in docking [30] our result of 67.7% hits is encouraging. Even the hit rate of 63.4%, achieved with a much lower resolution (Lr=3.0 Å) and correspondingly much faster (ϕ time=16.3 s), appears reasonable. Multiple ligand superposition may be an alley towards further decreasing the number of erroneous solutions, as suggested recently in [11]. E.g., the three thermolysin ligands 1tmn, 5tln, and cbz reveal the following pairwise rms deviations if the optimization is started from the observed relative orientation: 1tmn/cbz, 0.90 Å; 1tmn/5tln, 0.49 Å; cbz/5tln, 3.32 Å. Obviously, the observed relative orientation of the pair cbz/5tln is not detectable from the ligands alone. In fact, visual inspection shows that the overlapping portion of the two ligands in this orientation is rather small. The inclusion of another ligand used as a reference (1tmn in this case) allows for a substantial enhancement of the superpositions revealed.

Fragment placement

The second test application was to use RIGFIT as an alternative to the combinatorial fragment placement procedure in FLEXS [2]. FLEXS heavily relies on directional intermolecular interactions (e.g., H-bonds). Accordingly, some hydrophobic ligands in our FLEXS test set could not be handled so far (e.g., steroids). Exactly these have been used as a test suite for RIGFIT as an alternative anchor placement procedure in FLEXS. In this application different conformers of a selected anchor fragment are processed sequentially utilizing the rigid-body placement method presented here. RIGFIT works perfectly on such cases, and places the anchor fragment with an accuracy of about 1 Å rms deviation. Taking this as a starting point for the subsequent flexible ligand superposition, FLEXS computes solutions with an rms deviation of less than 1.5 Å for most of the test cases. Table 2 shows the detailed results.

The increased run times here are due to the consideration of multiple conformers. The examples considered contain five steroid-type ligands binding to an immunoglobuline (extracted from the structures with PDB-codes 1dbb, 1dbj, 1dbk, 1dbm, and 2dbl) and six trypsin ligands of limited size (extracted from the structures with PDB-codes 1tnh, 1tni, 1tnj, 1tnk, 1tnl, and 3ptb). The anchor fragments selected comprise in each case the complete ring system for the steroids and the phenyl-ring extended by the next one or two heavy atoms attached to the ring (in some of the cases the entire ligand) for the trypsin ligands. The number of conformers of the anchor varies from 1 to 9 with an average of 4.5, the average run time of the anchor placement is below 1 min, and the average accuracy is 1.36 Å. Except for the ligand in 1tni as reference, which occupies a substantially different position in the

Table 2. Alternative anchor placement

Reference	Test	N_{conf}	Anchor placement		Complex constr.	
ligand	ligand		time (s)	rms (Å)	time (s)	rms (Å)
1dbb	1dbj	6	139	1.35	16	1.35
1dbb	1dbk	1	23	1.57	6	1.57
1dbb	1dbm	9	205	0.67	1	1.68
1dbb	2dbl	2	42	1.17	1	1.43
1dbj	1dbk	1	23	0.50	1	0.50
1dbm	1dbj	6	143	1.57	22	1.57
2dbl	1dbj	6	111	1.66	26	1.66
1dbm	1dbk	1	24	1.27	6	1.27
2dbl	1dbk	1	19	1.54	5	1.54
1dbm	2dbl	2	43	1.19	2	1.61
3ptb	1tnh	6	61	1.38	9	1.33
1tni	1tnh	6	32	0.84	5	1.31
1tnh	1tnj	6	50	1.05	10	0.92
1tnk	1tnh	6	43	0.90	10	0.86
1tnl	1tnh	6	20	1.25	15	1.25
1tni	3ptb	2	26	2.93	3	2.02
1tni	1tnj	6	26	3.18	1	1.95
1tni	1tnk	6	57	1.39	4	1.07
1tni	1tnl	6	31	2.41	20	1.87
3ptb	1tnj	6	28	1.23	4	0.93
1tnk	1tnj	6	55	1.45	14	0.46
1tnj	1tnl	6	20	0.78	9	0.78
1tnk	3ptb	2	55	1.33	11	0.49
1tnk	1tnl	6	17	0.73	8	0.73
1tnl	3ptb	2	5	0.56	17	0.56
Average		4.5	51.9	1.36	9.0	1.23

The table shows results on those examples of our current FlexS test set which contain less than three specific directional interaction partners. So far, FlexS could not handle these ligands. However, RigFit has been applied successfully as an alternative anchor placement routine in these cases. Reference ligand denotes the structure which is taken as a rigid reference onto which the test ligand is flexibly fitted. N_{conf} gives the number of conformations of the respective anchor fragment of the test ligand. The anchor placement results by RigFit as well as the subsequent complex construction results by FlexS are provided with run times and accuracies.

binding pocket of trypsin, all the anchor placements could be performed with below 2.0 Å rms deviation. The subsequent flexible completion of the ligand to be fitted, starting from these anchor placements of RigFit, is performed by FlexS in 9 s on average, revealing placements with an average rms deviation of 1.23 Å.

Database screening

The rigid-body superposition results suggest that RigFit is able to provide approximate superpositions very fast, if only a few Laue vectors are considered during optimization in Fourier space. The ability to determine a solution close to the observed relative orientation obviously depends on the degree of similarity of the aligned molecules in their observed relative orientation. This can be seen by the fact that RigFit determines the self-fit of each of the molecules perfectly, even at the lowest resolution level. The fragment placement results suggest that RigFit is able to align even molecular fragments onto a reference molecule reasonably. Therefore, in the final application we tuned RigFit to computing rough approximations quickly in order to perform database searches of molecular fragments (*target fragments*) against a large set of ligands. In order to do so, we restricted N_{Lv} to 17. Each of the alignments required 2 s with this setting.

The idea behind the DB-screening application is to find ligands in a database which carry functional groups with similar physico-chemical properties as the target fragment. As a first test, we used the pteridine ring moiety in dihydrofolate (ligand of dihydrofolate reductase, PDB-code: 1dhf) as a target fragment and screened against a set of 357 ligands. These were selected from the Cambridge Structural Database (CSD, [31], release 5.09) by the keywords *drug* and *active* in the annotation [G. Klebe, personal communication]. RigFit reliably placed the pteridine ring onto its native location inside dihydrofolate at highest rank. More interestingly, the placement of the target fragment onto the structure of methotrexate ranked as high as 20, thus clearly being among the highest-ranking solutions.

Figure 6 illustrates these two placements of the target fragment in the upper two pictures. Additionally, in its lower part, the figure provides a diagram showing the similarity score (y-axis) against an index of the ligand in the database (x-axis). For clarity reasons, the indices have been sorted in decreasing order by score. The distribution illustrates the ability of RigFit to screen out a high percentage of ligands, maintaining the close analogues to the target fragment. If only the top ranking solutions (e.g., scores above 2000) are considered, about 40 ligands are maintained (indicated by the dashed line in the diagram), while about 90% of the ligands can be screened out. Methotrexate (at position 20), which by no means contains an equivalent chemical substructure but shows similar field

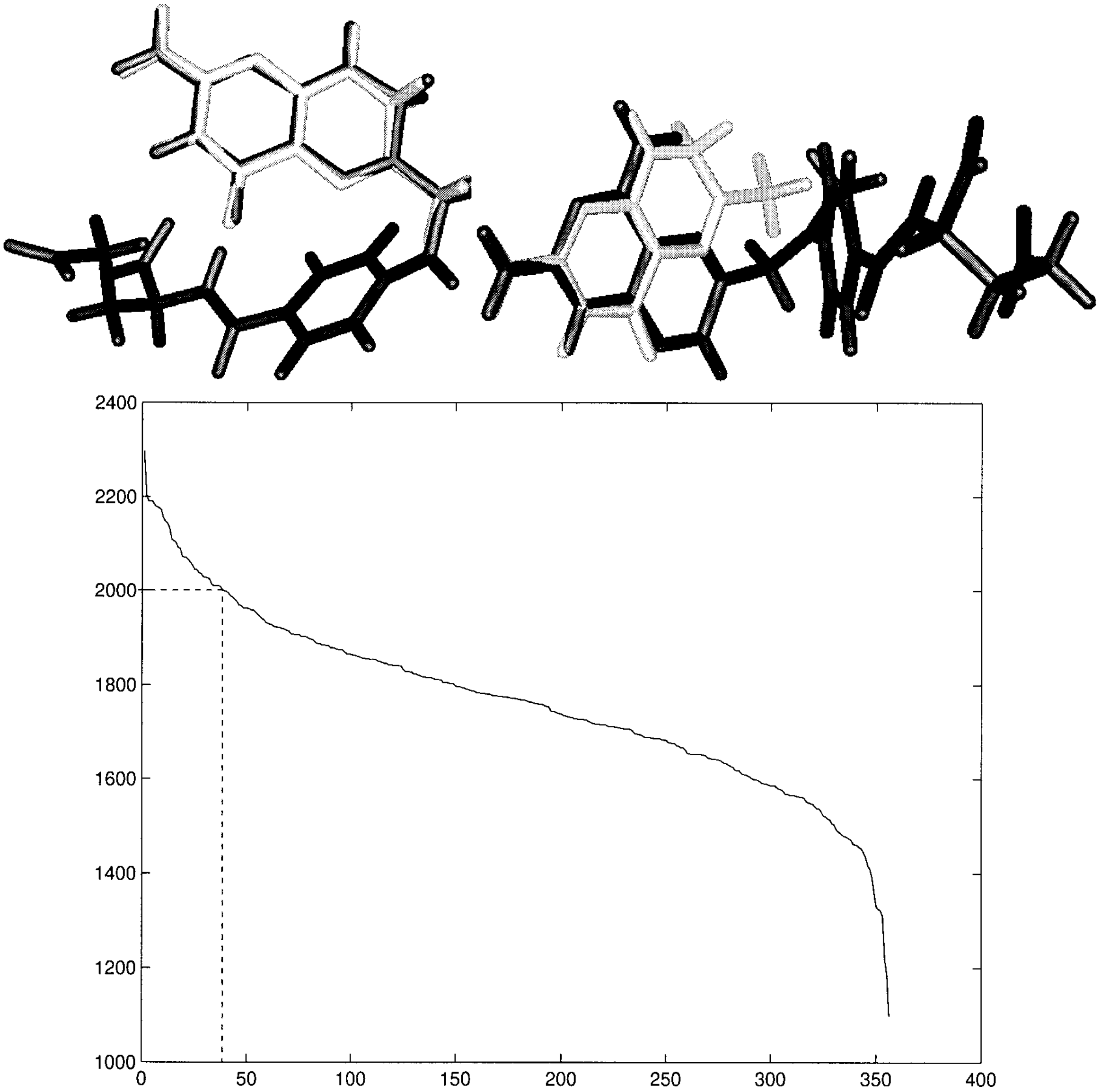

Figure 6. The results of a database scan aligning the pteridine moiety of dihydrofolate to each of 357 ligands extracted from the CSD. The upper part of the figure provides the first and the 20th superposition according to the ranking by RigFit-scores. These are the placements for the head group of dihydrofolate onto its native position inside dihydrofolate (top left) and onto the observed relative orientation inside methotrexate (top right) respectively. The diagram in the lower part of the figure provides the similarity score (y-axis) of each of the 357 superpositions (x-axis) sorted by score. It can be argued that about the top 40 placements (accordingly about 90%) can be discriminated from the rest (indicated by the dashed line).

properties in a certain region, clearly belongs to the maintained set of ligands. Methotrexate is known to bind in another binding mode but into the same pocket as dihydrofolate. The RigFit-placement corresponds to the observed relative orientation for this pair of ligands. Therefore, RigFit in a DB-screening application deserves for more than just a confidentiality index (i.e., the similarity score). Additionally, it supplies a structural alignment, worthwhile to be analyzed separately, which leads to the respective score.

Conclusions and outlook

We present the rigid-body superposition method RigFit, which combines and extends computational concepts from three different domains to assemble an efficient superpositioning tool. These concepts are the independent rotational optimization performed in Patterson space [15], the modeling of physico-chemical properties by sets of Gaussian functions [4] and a fast quasi-Newton optimizer.

In addition to its speed (compare, e.g., to [3, 4, 15]), the resulting superposition method has several advantages over conventional methods. Two of them should be emphasized separately: first, the wide applicability as demonstrated by three test scenarios (in [3, 10, 15] e.g., tests have been carried out only on a few examples of limited size), second, the usefulness of RIGFIT as a fragment placement routine, which already led to the incorporation of the tool into the flexible ligand superposition software FLEXS. To the best of our knowledge, fragment-based alignment has not been performed with any of the existing rigid-body superposition methods so far.

We will further investigate the applicability of RIGFIT for database screening. Also, we expect the run time to be further reduced by the use of interpolation methods in Fourier space.

Acknowledgements

We thank W. Nissink from Utrecht University for his suggestions and for making available unpublished material on his Fourier space approach to electron density fitting. We are grateful to our collaborating partners in the RELIWE-project, especially G. Klebe and T. Mietzner from BASF Main Laboratory, for many stimulating discussions on the different concepts of molecular similarity. The RELIWE-project is partially funded by the German Federal Ministry for Education, Science, Research and Technology (BMBF) under Grant no. 01 IB 302 A.

References

1. Rarey, M., Kramer, B., Lengauer, T. and Klebe, G., J. Mol. Biol., 261 (1996) 470.
2. Lemmen, C. and Lengauer, T., J. Comput.-Aided Mol. Design, 11 (1997) 357.
3. Kearsley, S.K. and Smith, G.M., Tetrahedron Comput. Methodol., 3 (1990) 615.
4. Klebe, G., Mietzner, T. and Weber, F., J. Comput.-Aided Mol. Design, 8 (1994) 751.
5. Kato, Y., Inoue, A., Yamada, M., Tomioka, N. and Itai, A., J. Comput.-Aided Mol. Design, 6 (1992) 475.
6. Marshall, G.R., Barry, C.D., Bosshard, H.D., Dammkoehler, R.D. and Dunn, D.A., In Olson, E.C. and Christoffersen, R.E. (Eds.) Computer-Assisted Drug Design, Vol. 112, American Chemical Society, Washington, DC, 1979, pp. 205–222.
7. Martin, Y.C., Bures, M.G., Danaher, E.A., DeLazzer, J., Lico, I. and Pavlik, P.A., J. Comput.-Aided Mol. Design, 7 (1992) 83.
8. Jones, G., Willett, P. and Glen, R.C., J. Comput.-Aided Mol. Design, 9 (1995) 532.
9. Grant, J.A., Gallardo, M.A. and Pickup, B.T., J. Comput. Chem., 17 (1996) 1653.
10. Mestres, J., Rohrer, D.C. and Maggiora, G.M., J. Comput. Chem., 18 (1997) 934.
11. Mestres, J., Rohrer, D.C. and Maggiora, G.M., J. Mol. Graphics Mod., 15 (1997) 114.
12. Petitjean, M., J. Chem. Inf. Comput. Sci., 36 (1996) 1038.
13. Navaza, J., Acta Crystallogr., A50 (1994) 157.
14. Diederichs, K., Proteins, 23 (1995) 187.
15. Nissink, J.W.M., Verdonk, M.L., Kroon, J., Mietzner, T. and Klebe, G., J. Comput. Chem., 18 (1997) 638.
16. Klebe, G., In Kubinyi, H. (Ed.) 3D QSAR in Drug Design. Theory, Methods and Applications, ESCOM, Leiden, 1993, pp. 173–199.
17. Bures, M.G., In Charifson, P.S. (Ed.) Practical Application of Computer-Aided Drug Design, Marcel Dekker, New York, NY, pp. 39–72.
18. Carbó, R., Leyda, L. and Arnau, M., Int. J. Quant. Chem., 17 (1980) 1185.
19. Hodgkin, E.E. and Richards, G., Int. J. Quant. Chem.: Quant. Biol. Symp., 14 (1987) 105.
20. Good, A.C., Hodgkin, E.E. and Richards, W.G., J. Chem. Inf. Comput. Sci., 32 (1992) 188.
21. Rossmann, M.G. and Blow, D.M., Acta Crystallogr., 15 (1962) 24.
22. Hiller, C., Optimierungsmethoden zum strukturellen Alignment von Ligandmolekülen. Master's thesis, University of Bonn, 1997.
23. Lattman, E.E., Acta Crystallogr., B28 (1972) 1065.
24. Dennis, J.E. and Schnabel, R.B., Numerical Methods for Unconstrained Optimization and Nonlinear Equations, Prentice-Hall, NJ, 1983.
25. Griewank, A.O., Markey, B.R. and Evans, D.J., J. Chem. Phys., 71 (1979) 3449.
26. Rossmann, M.G., The Molecular Replacement Method, Gordon & Breach, New York, NY, 1972.
27. Cooper, D.L. and Allan, N.L., J. Comput.-Aided Mol. Design, 3 (1989) 253.
28. Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A.A., Aflalo, C. and Vakser, I.A., Proc. Natl. Acad. Sci. USA, 89 (1992) 2195.
29. Mattos, C., Rasmussen, B., Ding, X., Petsko, G.A. and Ringe, D., Nat. Struct. Biol., 1 (1994) 55.
30. Jones, G., Willett, P., Glen, R.C. and Taylor, R., J. Mol. Biol., 267 (1997) 727.
31. Allen, F.H., Bellard, S., Brice, M.D., Cartwright, B.A., Doubleday, A., Higgs, H., Hummelink-Peters, T., Kennard, O., Motherwell, W.D.S., Rodgers, J.R. and Watson, D.G., Acta Crystallogr., B35 (1979) 2331.
32. Bracewell, R.N., The Fourier Transform and its Applications, 2nd ed., Electrical and Electronic Engineering Series, McGraw-Hill, New York, NY, 1978.