J Comput Aided Mol Des (2006) 20:773–788
DOI 10.1007/s10822-006-9078-7

ORIGINAL PAPER

Efficient overlay of small organic molecules using 3D pharmacophores

Gerhard Wolber · Alois A. Dornhofer · Thierry Langer

Received: 12 June 2006 / Accepted: 4 September 2006 / Published online: 19 October 2006

Abstract Aligning and overlaying two or more bio-active molecules is one of the key tasks in computational drug discovery and bio-activity prediction. Especially chemical-functional molecule characteristics from the view point of a macromolecular target represented as a 3D pharmacophore are the most interesting similarity measure when describing and analyzing macromolecule-ligand interaction. In this study, a novel approach for aligning rigid three-dimensional molecules according to their chemical-functional pharmacophoric features is presented and compared to the overlay of experimentally determined poses in a comparable macromolecule coordinate frame. The presented approach identifies optimal chemical feature pairs using distance and density characteristics obtained by correlating pharmacophoric geometries and thus proves to be faster than existing combinatorial alignment methods and creates more reasonable alignments than pure atom-based methods. Examples will be provided to demonstrate the feasibility, speed and intuitiveness of this method.

Keywords Molecular alignment · Molecular superimposition · Pharmacophore · LigandScout

G. Wolber (✉) · A. A. Dornhofer
Inte:Ligand Softwareentwickungs- und Consulting GmbH, Mariahilferstrasse 74B/11, Vienna 1070, Austria
e-mail: wolber@inteligand.com

T. Langer
Computer-Aided Molecular Design Group, Center of Molecular Biology Innsbruck, Institute of Pharmacy, Innrain 52a, Innsbruck 6020, Austria

Introduction

Aligning small organic molecules in the context of a macromolecular target is the basic idea behind most model building approaches in computational drug discovery: Common understanding of molecular similarity induces that there exists a pair of similar conformations between two different chemical structures that share the same chemical features at similar positions in three-dimensional space. The ensemble of the shared chemical features responsible for binding, the *shared feature pharmacophore*, allows the two similar bio-active molecules to bind to the macromolecule in a comparable way and trigger similar biological responses.

3D pharmacophore models have proved to be a very comprehensive and thus powerful representation of small molecule binding, since they reflect the medicinal chemist's view of a molecule in a very intuitive way [1–5]. They capture the concept of bio-isosterism by not only comparing atoms or topological similarities but structural groups at similar locations with the same chemical functionality. It is essential to concentrate on the chemical features since topological molecule characteristics are often misleading for the superpositioning of molecules with respect to their biochemical interaction characteristics. Figure 1 shows an example with two dihydrofolate reductase inhibitors, where a topological overlay would result in a wrong superpositioning mode; however, if pharmacophore points (the hydrogen bonding pattern in this example) are taken as the only information source for the alignment, the correct overlay mode can be retrieved [6]. Furthermore, a pharmacophoric view onto a molecule allows comparing molecules with different scaffolds, but similar chemical functionality.

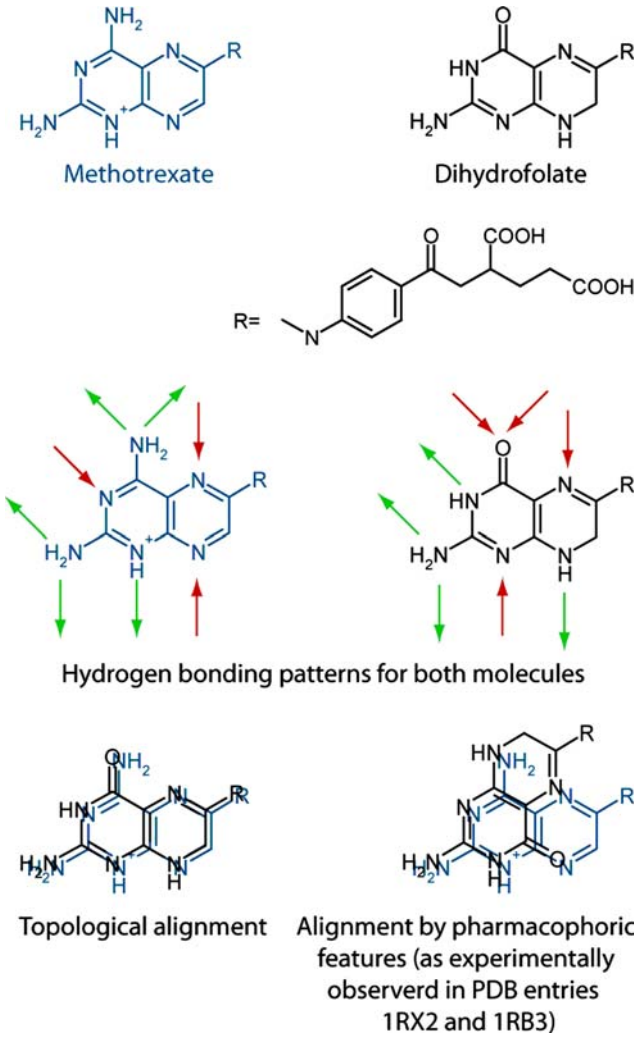

Fig. 1 The experimentally observed geometric alignment cannot be derived from atomic contributions algorithmically, but is correctly perceived as soon as the pharmacophore is considered [6]

The problem of finding the common pharmacophoric pattern from several molecules—although looking simple from a chemist's view—raises some computational problems that have not been solved in a satisfying manner so far: The number of conformations to compare in an infinite number of possible transformations is combinatorially growing with the degree of flexibility in the investigated ligands. A quantitative comparison is very difficult and impossible to address through exploring all possibilities in a brute-force approach using current computational infrastructure. An excellent review by Lemmen and Lengauer [7] describes more than 20 methods of addressing this problem in an optimized or heuristic way and the resulting limits of these approaches. Amongst those considering a pharmacophoric view on the molecules, there are two main trends for approaching the problem of extracting a shared pharmacophore pattern from several molecules with a binding mode that is assumed to be comparable: (i) Flexible and (ii) rigid molecule alignments allowing pharmacophore model elucidation in a subsequent step. The first successful and most prominent approach for rigid alignment and pharmacophore elucidation is the program DISCO developed by Martin and co-workers [8]. It addresses the combinatorial problem efficiently by applying the clique detection algorithm introduced by Bron and Kerbosch [9] to a complete distance graph of pharmacophore features represented as points. Two other algorithms for rigid pharmacophore pattern identification were published later [10] and have been intensely used for application studies in the past few years [11, 12], since the program was made commercially available as part of the program package CATALYST [13]. The well-known flexible pharmacophore elucidation program GASP [14] developed by Jones and co-workers in P. Willet's group uses flexible super-positioning and directed search within conformational space by a genetic algorithm while simultaneously optimizing the best overlap of common pharmacophoric features. Recent similarity approaches dealing with the alignment of molecules in 3D space consider a field-based measure of tanimoto to compare the electrostatic potential of two small molecules [15], such as OpenEye's tool EON [16]. Others use molecular shape fingerprints [17] for comparison, such as OpenEye's tool ROCS [18].

There are two different problems causing combinatorial explosion: First, there is conformational flexibility, which will not be subject to our algorithm, since this problem is well addressed by multi-conformational generators and has been discussed in recent publications [19, 20]. Conformational space and thus the ability to calculate a universal conformational model depends on the flexibility of the molecule, especially the number of freely rotatable bonds. It is possible to generate generally applicable conformational ensembles that sufficiently resemble bound conformations from the Protein Data Bank [21] for most small organic drug-like molecules [22]. These calculations are rather expensive in terms of computational cost and it therefore makes sense to store pre-computed conformational models [12, 23]. Regarding our studies on the quality of conformer generation programs, it can be summarized that existing programs cover conformational space for the purpose of rigid 3D fitting techniques like pharmacophore searches sufficiently well [20, 22]. Methods that tweak the molecule while fitting must dramatically reduce the search space while aligning in order to stay efficient and therefore bear the

danger of falling into a local minimum instead of investigating conformational space globally. Especially for the purpose of virtual screening, where a set of pre-computed conformers is used more than once, pre-computing conformers seems to be the method of choice for database searches and virtual screening.

The second combinatorial problem is the identification of the largest common sub-pharmacophore between several pharmacophores that were identified to be relevant for the respective molecules. An algorithm that is based on the comparison of all feature groups (i.e. forming all pairs, or all three-point pharmacophores and merging them back together) grows exponentially with the number of features involved, and cannot be solved in polynomial time thus being very slow with mid-sized drug molecules. Since acceptable solutions for the problem of conformer generation already exist, our work will focus on the second problem. In the following, a method will be introduced that is able to identify a maximum common pharmacophoric pattern between two molecules or pharmacophore models in a new and efficient way. These shared pharmacophore features can be used to transform the underlying molecules in 3D space thus producing meaningful alignments of topologically different molecules in three-dimensional space.

Aim and methods

The aim of this work is the creation of an algorithm that is able to align molecules on the basis of its chemical features (i.e. hydrogen bond donors, acceptors, positively and negatively ionizable features, and lipophilic aggregation spheres). It should be fast and propose a globally optimal solution that can be used for unsupervised ranking of different overlaid poses when virtually screening large amounts of molecules. Our algorithm is intended to (i) find important pharmacophoric features for a molecule and project them into three-dimensional space (*3D pharmacophore creation*), (ii) identify those features which are to be paired (*feature matching*), and (iii) calculate the necessary rotation and translation of the two 3D pharmacophores and apply this transformation to the underlying molecules (*translation and rotation*). These steps will be performed sequentially, and optionally refined by going back to feature matching (ii) and re-transformation (iii). Especially the assignment of correct feature pairs is a critical part, but given a good geometric classification for geometric relationships amongst the pharmacophoric feature points, an efficient implementation for optimal pairing of similar features should be possible. Distance-bin classification have proved to be successful for many chemistry-related problems, such as the implementation of atom-based 3D superimposition of molecules, presented in 2004 by Richmond et al. [24]. This method outperforms all previously known alignment methods by suggesting a single optimal solution, but still bears the disadvantage that it concentrates on atom contribution and therefore basically performs pattern recognition on atomic characteristics constrained by bond length distances. The mathematical base of the pairing algorithm is an algorithm that was already published in 1955 [25], and allows the identification of one optimal set of atom pairs in a very efficient way. Pharmacophore points show different distribution characteristics than atoms in 3D space, because the points are much sparser with less symmetry. 3D pharmacophore selectivity and thus its ability to discriminate between bio-active and non-bio-active molecules is solely defined by geometric relations between the chemical features, which also plays a crucial role in capturing important characteristics, but not being too specific to prevent the comparison of different scaffolds with similar chemical functions.

There are several algorithmic modules available from different disciplines that help building the algorithmic modules needed for this alignment technique: Fully automated 3D pharmacophore creation is possible through using our previously introduced code base from the LigandScout program [26], a single optimal chemical feature pairing solution can be built using maximum weighted maximum cardinality bipartite matching techniques, and once these pairs are built, the transformation to minimize distances between feature point pairs can be performed using an algorithm that was proposed by Kabsch in 1976 [27] and finally corrected in 1978 [28]. These algorithms and their adoptions to the given problems will be described in the following.

Pharmacophore creation

Pharmacophore creation is the initial step before starting the alignment. A new representation of the molecule is created, replacing topological information such as the connection table, atoms, and bonds by a more abstract 3D representation without connectivity information (Fig. 2). The pharmacophore is created using our LigandScout application framework [26] and concludes the detection of directed hydrogen bonding interactions, lipophilic regions, charge interactions and steric exclusions. Chemical features are detected in such a way that the algorithms checking for chemical feature complementarity between a ligand and a

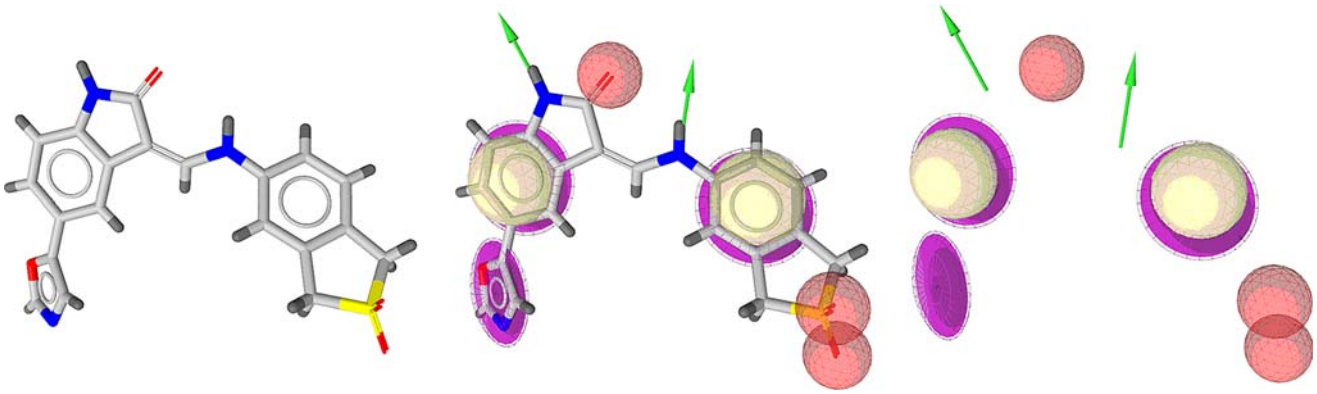

Fig. 2 3D structure of the Cyclin-Dependent Kinase 2 (CDK2) inhibitor 3-{[(2,2-Dioxido-1,3-dihydro-2-benzothien-5-yl)amino] methylene}-5-(1,3-oxazol-5-yl)-1,3-dihydro-2H-indol-2-one from PDB entry 1KE7 (left) with its abstract molecule pharmacophore (center). For the alignment algorithm only the pharmacophore representation (right) will be taken into account, all topological information (atoms and bonds as shown on the left) is omitted

macromolecule are omitted and all features that can be applied to the ligand side are added as three-dimensional objects. The level of abstraction for pharmacophore representation has already been presented with LigandScout and is shown in Table 1. Chemical features that require a projected point on chemical feature layer 3 are moved to the more abstract layer 4 at this point, because the projection geometry becomes incomplete without macromolecule on the other side. A hydrogen bond donor originating from a sp3 heavy atom on the ligand side, for example, is reduced to a point feature since the direction of the projected point would be undefined. Although plausible directions for this hydrogen bond will be checked in a post-processing step, the initial representation is a simple sphere for the initial alignment. For the case of a donor originating from a sp2 heavy atom on the ligand side, the projected point can be added, since it has a defined position without a macromolecule.

Pairing

Once pharmacophore features for both structures to be aligned are present, pairings between the two sets of pharmacophore features need to be found to perform a geometric transformation based on their minimum squared distance deviation later. Let $P_1 = \{F_{1,i}\}$, $i = 1,\ldots,n$, and $P_2 = \{F_{2,j}\}$, $j = 1,\ldots,m$,be two pharmacophore feature sets containing n and m chemical features, respectively. This paragraph sketches how to derive a matching defined as $\{(F_{1,i_k}, F_{2,j_l}) : i_k \in A_1, A_1 \subset \{1,\ldots,n\}, j_l \in A_2, A_2 \subset \{1,...,m\}, |A_1| = |A_2| \leq \min\{n,m\}\}$.

The basic concept behind the intended pair assignment is to match those chemical features, which show identical chemical and similar geometrical characteristics. In order to create a similarity measure, which allows for easy comparisons, the 3D position information of each pharmacophore is represented as one-dimensional distance bins with a specified resolution defined as $B_t \in \mathbb{N}^r, t \in \{\text{chemical feature types}\}$, r = maximum occurring distance category, for each feature type contained in the respective pharmacophore feature set. This means that the distance relation for the underlying pharmacophore feature set is to be encoded for each contained feature separately. Having derived all type-based distance bins for each feature in both feature sets, a cost matrix $C = (c_{i,j})$ strongly dependent on the similarity of both sets of distance bins is constructed with entries $c_{i,j} = \text{cost}\left(\bigcup_t B_{i,t}, \bigcup_t B_{j,t}\right)$, $B_{i,t}$= distance bin associated to feature F_i for feature type t, $B_{j,t}$ = distance bin associated to feature F_j for feature type t. The cost function will be explained in detail in the implementation paragraph below. This means that the more both sets of distance bins have in common the lower the costs for building this pair will be. Based on this cost matrix a maximum number of feature pairs shall be identified, i.e. not only the pairs with the minimum matching cost, but also a maximum number of pairs and therefore a global minimum matching cost is to be found. There is an established and efficient algorithm to create the best possible pairings: the Hungarian Matching algorithm [25], which performs maximum weighted maximum cardinality bipartite matching in polynomial time. After the optimal pairing based on the actual cost matrix has been identified, the transformation step as proposed by Kabsch has to be applied.

Table 1 Chemical feature abstraction layers in LigandScout

Layer	Classification	Universality	Specificity
4	Chemical functionality without geometric constraint, e.g. an H-Bond acceptor without a projected point or a lipophilic group	+++	–
3	Chemical functionality (H-bond acceptor, H-bond donor, positive ionizable, negative ionizable, hydrophobic) with geometric constraint, e.g. an H-bond acceptor vector including an acceptor point as well as a projected donor point; aromatic ring including a ring plane	++	+
2	Molecular graph descriptor (atom, bond) without geometric constraint, e.g. a geometrically unconstrained phenol group	–	++
1	Molecular graph descriptor (atom, bond) with geometric constraint, e.g. a phenol group facing a parallel benzenoid system within a distance of 2 to 4 angstroms		+++

3D Transformation: Kabsch alignment

Kabsch showed how to analytically find the best rotation $R \in \mathbb{R}^{3\times3}$ - represented as an orthogonal matrix - of one set of points x_n into another set of points y_n by minimizing the weighted sum of squared distance deviations, i.e. minimizing $\sum_n w_n(Rx_n - y_n)^2$, where $w_n$represents the weight factor for the n-th pair of positions [27, 28]. In order to find a global minimum, Lagrangian multiplicators are applied to the function to be minimized. A minimum of the resulting function is characterized by the first derivative being zero and the second derivative being positive definite. Using the symmetry and positive definite properties of matrices an intermediate matrix is derived which is decomposed into its Eigenvectors and Eigenvalues. These are used to construct an orthogonal matrix which represents the best transformation. Orthogonal matrices describe rotations as well as reflections, and the issue of reflections was not covered in the original publication, but corrected by Kabsch later [28]. Reflections can be distinguished from rotations by calculating the determinant of the transformation matrix. If $\det(R) = 1$, R defines a rotation; if $\det(R) = -1$,R defines a reflection. The case of reflections was resolved by negating one vector of an intermediate result yielding an orthogonal matrix with determinant 1 and thus the desired best rotation transformation. For this study, the Java implementation in the ilib framework [29], which is the base for the LigandScout application [26] was used.

Implementation

The paragraphs above already conceptually explained parts of the alignment algorithm. This section gives an overview of the complete algorithm and focuses on implementation details, modifications and necessary adoptions that were specific for pharmacophore point matching.

Figure 3 shows the iterative process of the alignment algorithm: Starting with two pharmacophores P_1 and P_2 to be aligned, each chemical feature $F_i \in P_1$ and $F_j \in P_2$ is assigned a set of feature-type based distance bins $B_{i,t} \in \mathbb{N}^{r_{i,t}}$ and $B_{j,t} \in \mathbb{N}^{r_{j,t}}$, respectively, encoding distance relations to all other features in the same pharmacophore. Depending on the similarity of the set of distance bins $\bigcup B_{i,t}$ and $\bigcup B_{j,t}$, a cost matrix $C = (c_{i,j})$ is constructed. The features $F_i \in P_1$ are represented as rows of C whereas features $F_j \in P_2$ are represented as columns of C. The potential feature pair (F_i, F_j) is assigned the cost value

$$c_{i,j} = \sum_{k=1}^{\min(\max(r_{i,t}),\max(r_{j,t}))} (\omega(k) \cdot \sum_t \min(B_{i,t}[k], B_{j,t}[k]))$$

where $v[k]$ represents the k-th component of the vector v, $v[k] := 0$ for $v \in \mathbb{N}^r, r < k$, the weight function $\omega(k) = \left(1 - \left(\frac{k}{n}\right)^3\right) \cdot 2\left(1 - \frac{k}{n}\right) \cdot p, k =$ current bin index,$n =$ maximum bin index, and $p =$ maximum element count for all bins and all types. The weight

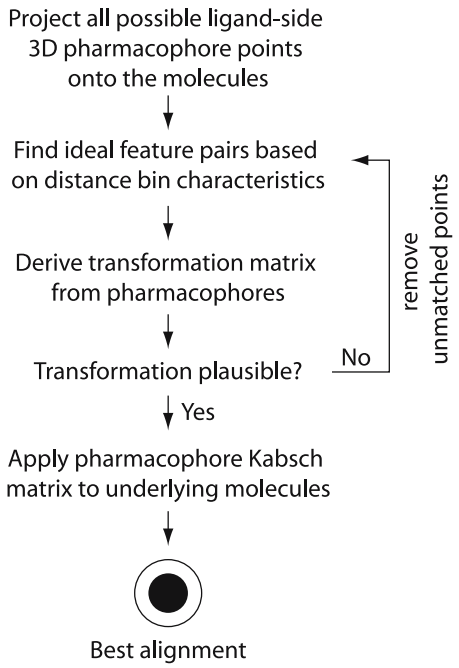

Fig. 3 Schematic overview of algorithmic steps for aligning two molecules

function is of order four concerning the variable k and is to be interpreted as compensation due to the information reduction from 3D to 1D (order three) as well as a preference of closer features over features further apart (order one). Figure 4 shows how distance bins are constructed: Considering the top-most sketch in Fig. 4, distance bins for feature A1 are to be constructed. The circles represent borders between the distance categories induced by a fixed resolution value. These circles are directly mapped to an array whose entries are the number of features found at the specific distance category. In this case features L1 and L2 are in the third circle which is equivalent to distance category 3, therefore the resulting array for the lipophilic feature type would be $A1(L) : 0|0|2$. However, due to reasons which will be explained below, a coarse binomially distributed filter $F = 1|2|1$ must be applied to all bins resulting in $A1(L) : 0|2|4$. No acceptor feature other than A1 occurs in this pharmacophore, therefore the resulting bin is $A1(A) = 0|0|0$. The same procedure is applied in the middle and bottom-most sketch in Fig. 4 yielding the respective distance bins. Figure 5 motivates the introduction of the binomially distributed filter. Both features to be inserted into the distance bins are very close to each other, which can be verified easily by observing that both features are close to the border between distance category 2 and distance category 3. (i) in Fig. 5 shows the distance bins without applying the filter. In this special case no overlap occurs, which complicates a match of both central features and does not reflect the real distance similarity. After applying the filter as in (ii) of Fig. 5 the overlap is $0|1|1|0$ which facilitates a potential match of both central features and resolves the problem.

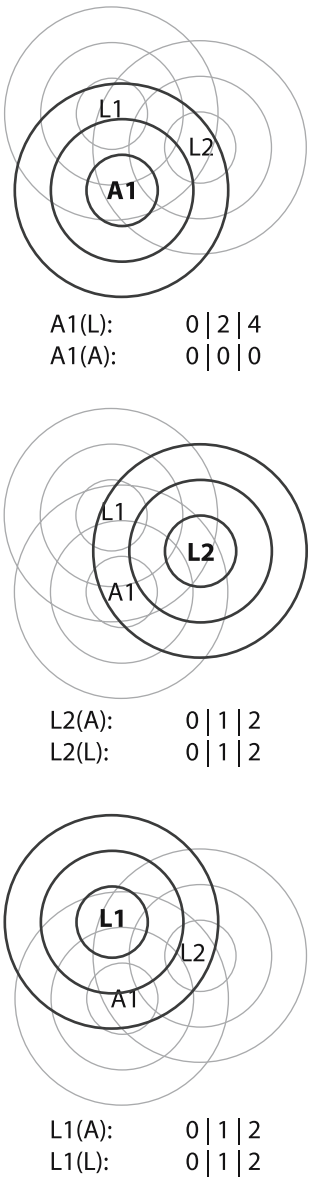

Fig. 4 Creation of distance bins. Acceptor feature A1, and lipophilic features L1 and L2 are shown with their respective distance categories depending on a fixed resolution. After the number of occurrences per distance category for the currently considered feature type is inserted into the distance bin, a binomially distributed filter F = 1|2|1 is applied to the distance bin yielding the final result

The cost function returns values that are higher if there is more similarity between both sets of distance bins, which would represent a maximization problem. The Hungarian Method, however, as implemented in the ilib framework [29], expects a minimization problem as input, which means that the cost matrix needs to be transformed to reflect this restriction. Let $m = \max(C)$ the maximum value in the cost matrix. Each entry in $C = (c_{i,j})$ will be transformed by $c_{i,j} = |m - c_{i,j}|$ resulting the desired minimization

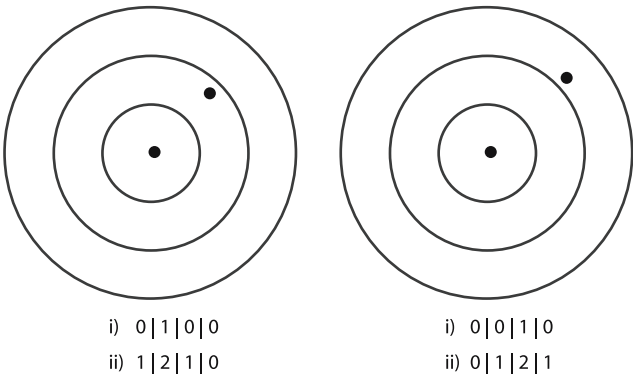

Fig. 5 For the two central features distance bins are to be constructed. The feature to be inserted on the left hand side would be in distance category 2, whereas the feature on the right hand side would be in distance category 3, although their geometric distance is quite small. (i) shows the distance bin without filtering, (ii) shows the filtered version

problem. This cost matrix is the basis for all further calculations. In the next step an optimal matching is returned from the Hungarian Matching algorithm, the Kabsch rotation step is applied to the resulting matching and is finally verified geometrically. Due to the reduction of information, which occurs when 3D position information is transformed to one-dimensional distance information, mismatches may emerge as a result of artificially introduced symmetry in the distance bins by virtue of missing orientation information. To identify the case of an inadequate matching, the mean value of the geometric distances between the features in each pair is calculated. Then, for each distance a check is performed to find out if the deviation relative to the mean value is abnormally high with respect to a specified threshold. In this case a mismatch was detected and the associated pairing must be excluded from the next matching. Therefore, the cost matrix C is adapted and the appropriate value for the inadequate feature pair $c_{i,j}$ is set to the maximum allowable value to omit this potential feature pair in subsequent iterations, i.e. $c_{i,j} = \mathrm{MAX}$. Based on the adapted cost matrix a new iteration step will be carried out. Each occurrence of a mismatch leads to another iteration step until either (i) an assignment of pairs is not possible because all potential pairings were forbidden, which reflects that no solution for the alignment could be found or (ii) no inadequate pair assignments exist, yielding the best solution for the alignment based on the given cost matrix. In order to consider additional geometric constraints in chemical features, such as a projected point of hydrogen bonding vectors, a final filtering was implemented as a post-processing step.

Post processing: remove points violating vector constraints

Vector features, such as hydrogen bonds, optionally are represented by a projected point in addition to the point positioned on the molecule. In a final step, the corresponding projected points are compared to each other by checking their positions against their tolerances. A vector feature is removed if the projected point position cannot be positioned within the other projected point's tolerance sphere. After this optional removal, a final geometric alignment step is performed representing the best alignment.

Examples and results

For the purpose of validation different ligands from the Protein Data Bank (PDB) were selected that bind to the same protein at the same location in a comparable way. Since the PDB file format does not encode atom types or valences for non-standard residues, and some PDB entries contain distorted geometries or wrong atom descriptions, all used entries were carefully checked manually and corrected using the graphical user interface of LigandScout [26]. The ligand-based alignment of two ligands as described in the first part of this study was compared to a transformation that is derived from an overlay of the alpha carbons of the two binding sites that surround the two ligands. All alpha carbons from the surrounding amino acids within a range of seven Angstroms from the ligand were subject to a Kabsch transformation as described in the 'Aim and methods' section of this article. The resulting transformation matrix was applied to the ligands to be compared. The difference between the pure ligand-based alignment and the alignment within the alpha carbon coordinate frame (i.e. the ''protein alignment'') was measured in terms of root mean square deviation (RMSD) of the distances between the atoms of the two instances of the same second molecule that are aligned to a molecule kept rigid within its coordinates as illustrated in Fig. 6.

Ligands for five pharmacologically relevant targets were chosen for comparison: Abelson Tyrosin Kinase, Phosphodiesterase 5 (PDE5), Cyclooxygenase 2 (COX2), Dihydrofolate Reductase (DHFR), and Cyclin-Dependent Kinase 2 (CDK2). The tables shown below denote the RMS deviation between the reference position within the protein and the ligand-based alignment as calculated by the described algorithm. The slight asymmetries between the RMSD values when comparing two distinct ligands to each other are caused by the fact that the ligand-based alignment algorithm starts removing points on the second alignment element, which may lead to slightly different results in the phase of the algorithm, where unmatched points are removed.

Abelson Tyrosin Kinase

The Novartis group developed the potent inhibitor STI-517 marketed under the trade-name Gleevec for the fusion protein BCR-ABL, which is produced due to a genetic defect causing chronic myelogenous leukemia (CML). During development a piperazine moiety was added to increase solubility. When aligning the two molecules and regarding the common binding motif, the piperazine should be ignored, which is perfectly reflected in the ligand-based alignment as well in the protein alignment as shown in Fig. 7. The atom-based RMSD value for this alignment is 0.5A.

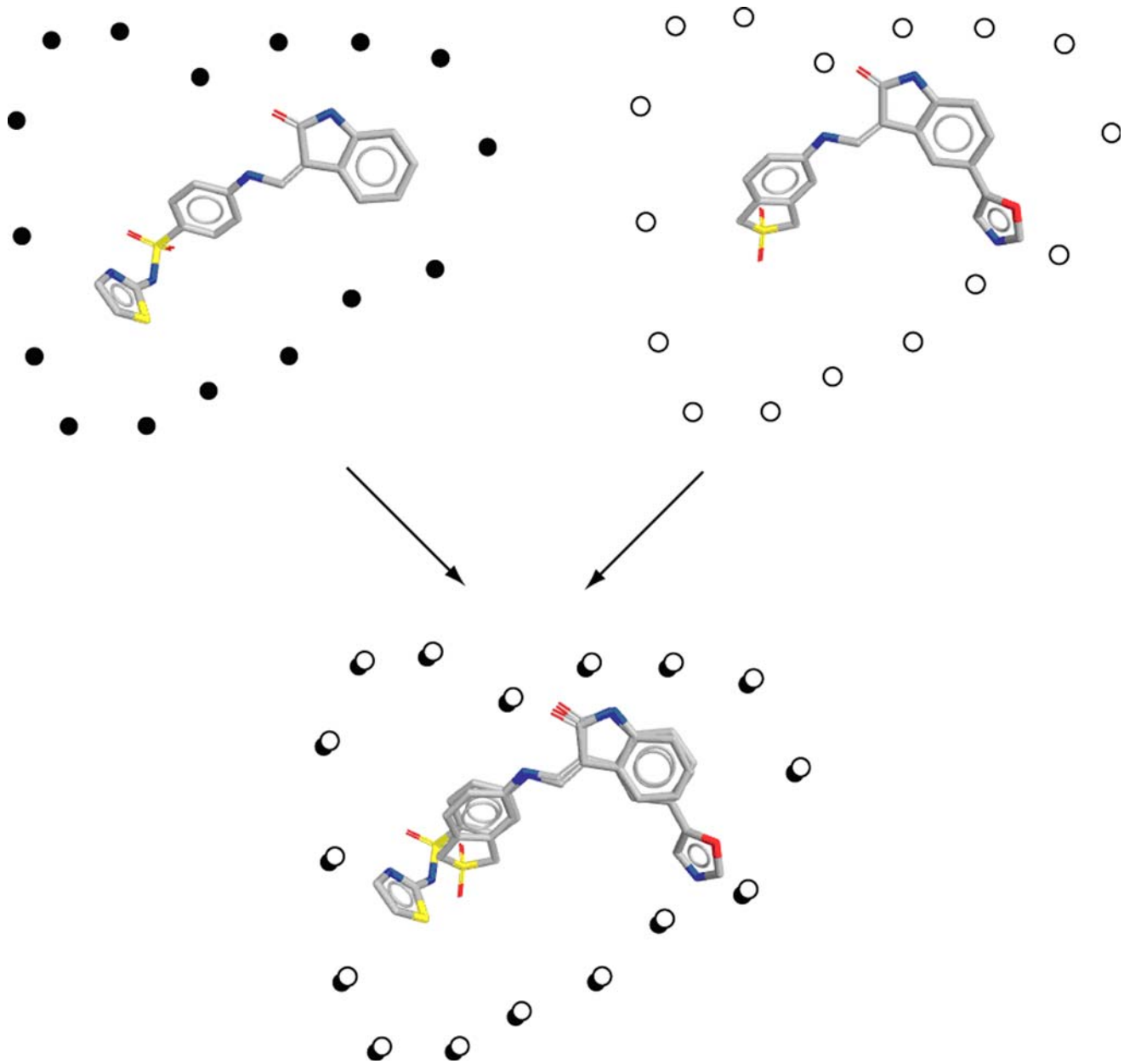

Fig. 6 Schematic representation for alpha carbon atoms surrounding two ligand molecules: The reference alignment within the protein by overlaying the alpha carbons is compared to the ligand-based alignment done by the proposed algorithm

Phosphodiesterase 5

Four PDB entries of Phosphodiesterase 5 complexed with its potent inhibitors Sildenafil and Tadalafil, respectively, were compared. Similar to the previous example (Gleevec) there is a piperazine moiety that changes position within the binding site of the protein. In the case of Sildenafil in PDB entry 1TBF the piperazine is differently oriented than in the conformation from PDB entry 1UDT. An automated, purely atom-based alignment that minimizes atom distances as shown in Fig. 8 would result in a bad alignment not capturing the pharmacophoric motif reflected in the observed binding mode and thus would not be suitable as a prerequisite to derive a common pharmacophoric model for virtual screening. The atom-based alignment shown in Fig. 8 was created in Accelrys' DS Visualizer 1.5 [13] by defining manual tethers for all atoms and minimizing the distances. The change of the piperazine position can also be observed in the two conformations of Tadalafil in PDB entries 1XP0 and 1UHO, and is also correctly perceived by the presented implementation of

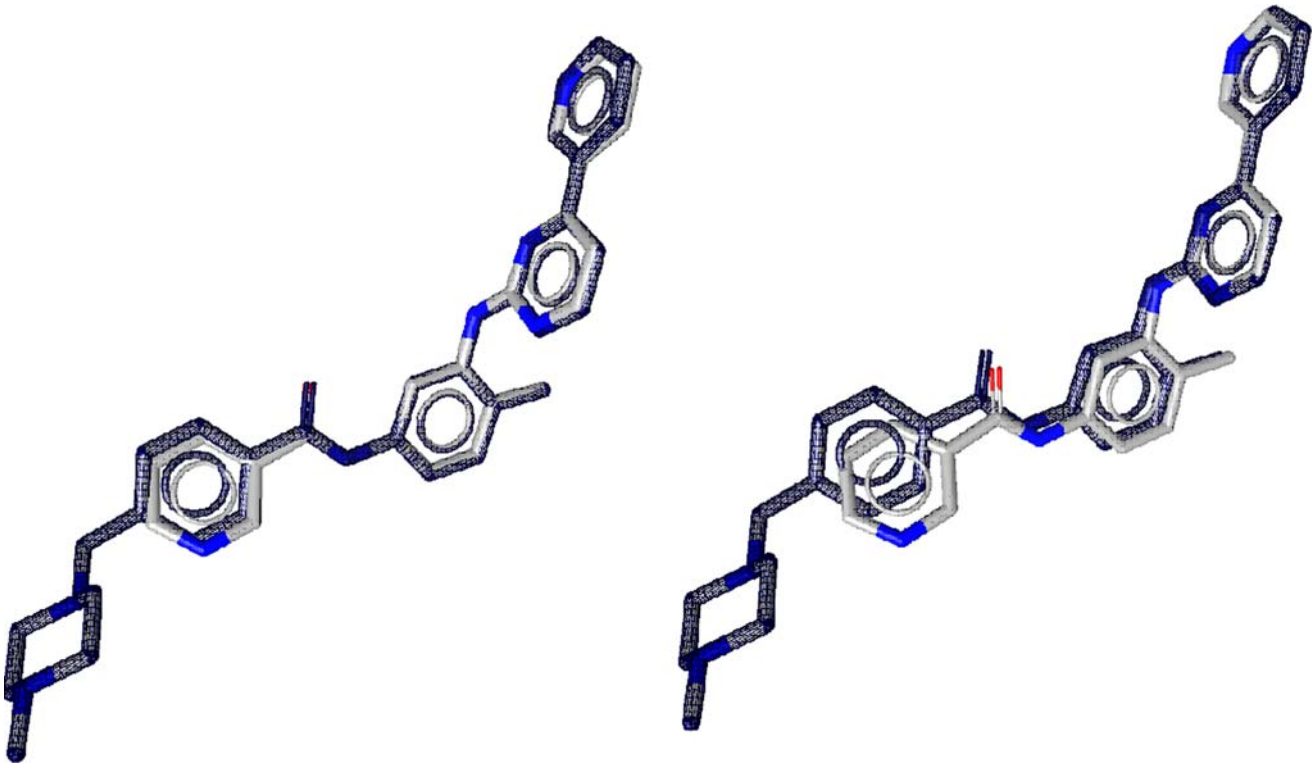

Fig. 7 The piperazine moiety is not taken into account for this alignment since the largest common pharmacophoric motif is correctly recognized by the algorithm

the pharmacophoric ligand-based alignment as shown in Fig. 9. The results in Table 2 show that all alignments of these four PDB entries resemble the experimentally determined alignments within a very low range of RMS deviation.

Cyclooxygenase 2 (COX2)

The selective COX-2 inhibitor SC558 from PDB entry 1CX2 is perfectly aligned onto the topological identical conformation from PDB entry 6COX. The alignment of Indomethacine, a non-selective COX-2 inhibitor to SC558 looks worse as illustrated in Fig. 7 and Table 3, but is still acceptable since most chemical features are mapped to similar locations. The inclusion of steric constraints, such as a shape around the molecule, would further improve the results in this case (Fig. 10).

Dihydrofolate Reductase

As already mentioned in the introduction and illustrated in Fig. 1, the common pharmacophoric pattern between Dihydrofolate and Methotrexate would be a problem for pure atom-based alignment methods. Using the pharmacophoric ligand-based alignment the pharmacophoric pattern between the two molecules is correctly perceived, and also the super-positioning of the different conformations of Dihydrofolate from the PDB entries 1RX2 and 7DFR are very similar to the protein alignment (Fig. 11 and Table 4). The common hydrogen bonding pattern as described by Böhm, Klebe, and Kubinyi even looks better in the ligand-based alignment, which can be explained by inaccuracies due to slightly different side chains in the two crystal structures of Dihydrofolate Reductase.

Cyclin-dependent Kinase 2

Finally, a series of rigid CDK2 inhibitors was tested. Using the pharmacophoric ligand-based alignment, all inhibitors perfectly align on top of each other in any pair-wise combination, as illustrated in Table 5. A pair-wise alignment comparing the inhibitors from PDB entries 1KE5 and 1KE8, respectively, and all five inhibitors aligned on top of each other are shown in Fig. 12.

Flexible alignment using multi-conformational models

In the beginning of this article we mentioned that the problem of flexible alignment can be solved by

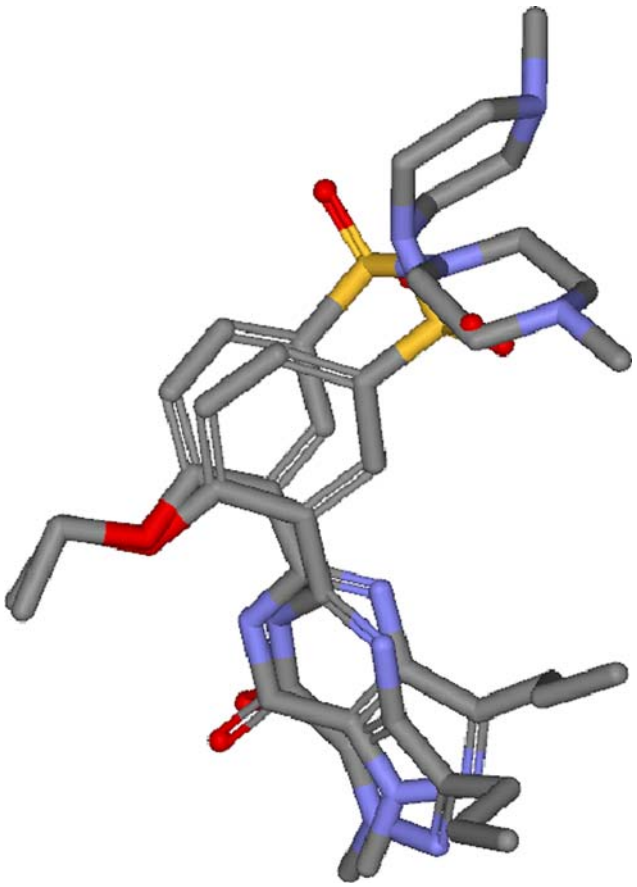

Fig. 8 The atom-based alignment of the two experimentally determined conformations of Sildenafil in PDB entries 1TBF and 1UDT shows minimal atom distances, but is unable to capture the common pharmacophoric motif between the two conformations and looks different than the observed or pharmacophoric ligand-based overlay as shown in Fig. 9

separating the problem into a rigid body alignment procedure and conformer generation. In an earlier study we showed that state-of-the-art conformer generation programs are able to cover conformational space sufficiently [22]. In order to investigate how this algorithm performs in a multi-conformational context, the program OMEGA was used to generate multi-conformational models from ligands that were selected from the set introduced in the first part of this section. The experimentally determined coordinates of the molecules were discarded, OMEGA 2.0 [19, 30] was used to create an unbiased conformational model and all conformations were then aligned to the experimentally determined conformation extracted from the PDB file. In a next step, each of the generated conformations was aligned to the experimentally determined conformation and ranked according to the scoring function described below. The best scored artificial conformation was then compared to the original PDB ligand in terms of atom RMSD.

To get a ranking among several solutions in an alignment, a scoring function has been developed, which favors solutions with a high number of geometrically matched feature pairs and which penalizes those with higher RMS deviations among those feature pairs. The scoring function is calculated as follows: $\text{score} = (10 \cdot n) + (9 - 3 \cdot \min(r, 3))$ (n = number of geometrically matched feature pairs and r = RMSD of these feature pairs). The advantage of this function is that the number of geometrically matched feature pairs can be perceived on first sight because the second additive term lies in the interval [0, 9].

The most important point to mention is that we use an abstract pharmacophoric representation of the molecule together with a scoring function that only operates on the pharmacophoric features. The aim of this experiment was to find out whether this abstract and simplified molecule representation performs well for picking the right conformation and for aligning it on top of the reference molecule.

The results shown in Table 6 demonstrate how the pharmacophore alignment worked for the selected examples. The atom based RMSD value shown in the second column is within the range that is considered to be a success for the reproduction of bio-active conformations. The pharmacophore RMSD as denoted in the next column shows the actual RMS deviation between the pharmacophore points identified by LigandScout [26]. For completeness we included the number of chemical features points that could be geometrically overlapped and the number of conformations that OMEGA generated with the settings ''maxConfs = 800, rmsThreshold = 0.2, searchForceField = mmff94sNoEstat, buildForceField = mmff94s_Trunc''. This gives an impression about the flexibility of the molecule: If the maximum number of conformations of 800 was not reached, this is an indication that the number of degrees of freedom of the molecule was lower and therefore conformational space is covered by the generated conformational model in a better way. Figure 13 graphically shows three examples, amongst which the alignment with the worst result from Table 6 (1rb3-mtx-161) is displayed: Despite its atom-based RMSD of 1.49 the most important chemical features overlap well.

Benchmarks

The developed ligand-based alignment proves to be an efficient algorithm for applications to small molecules since it does not explore all combinatorial possibilities, but still finds a single optimal solution due to the nature of the Hungarian Algorithm. All experiments have been carried out on a Pentium IV, 2.8 GHz processor

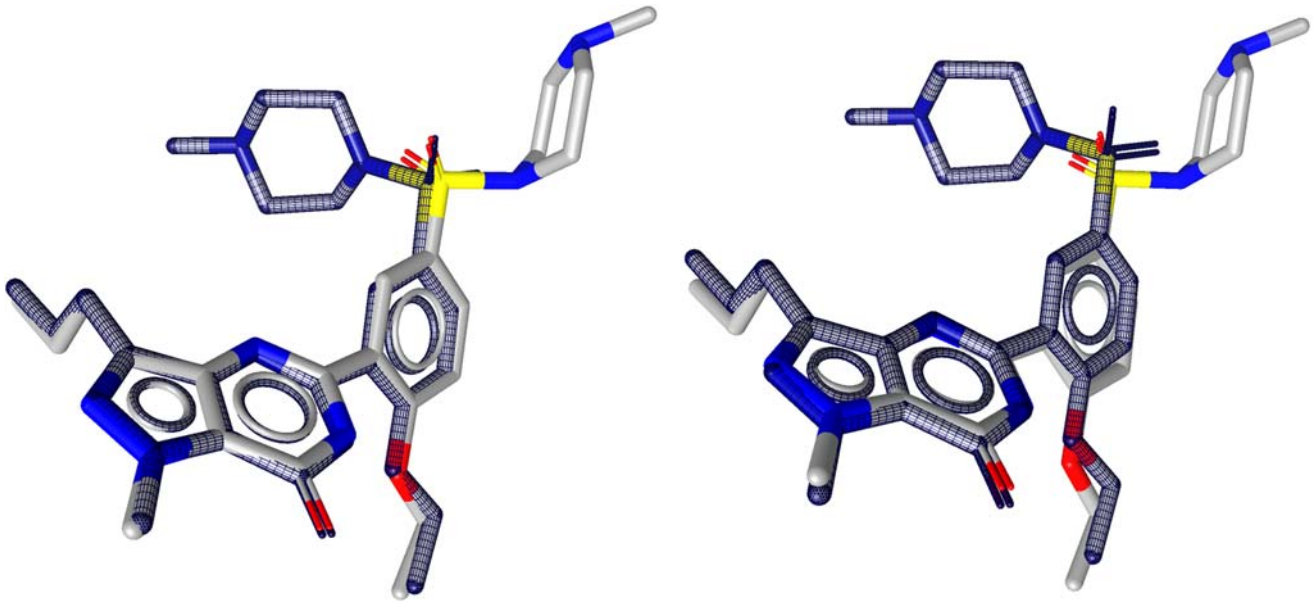

Fig. 9 The ligand-based alignment of Sildenafil is nearly identical to the observed super-positioning within the protein reflecting the common pharmacophoric motif and the flexible piperazine moiety

with 2 GB RAM running Gentoo Linux version 2006.0. The average time requirement for all of the presented application examples using our Java implementation is 27.7 ms, which qualifies this algorithm to be used in a large-scale environment. As an example, Table 7 shows the time in milliseconds required for each of the CDK2 inhibitor alignments. This high performance can be explained by the use of the maximum weighted maximum cardinality bipartite matching technique on the one hand, and the reduction of the molecule to few specific pharmacophore points on the other hand, as described in the 'Aim and methods' and 'Implementation' parts of this article.

Summary and discussion

In this study, a pharmacophore-based alignment algorithm was presented that efficiently superimposes ligands with respect to their pharmacophoric representation. Many algorithmic components used in this approach are known to solve particular problems in an optimal way, such as the solution to the pair assignment problem by Kuhn [25], and the 3D rotation of pair-wise points by Kabsch. The novel idea of this algorithm is that only the pharmacophore representation of a 3D conformation forms the basis for the alignment. This adds a new semantic layer on top of the molecule adding chemical knowledge. This semantic layer implicitly allows for a comparison of molecules capturing the concepts of bioisosterism and in a further stage intuitively permits chemists to compare different chemical compound classes with different scaffolds showing a similar pharmacophoric profile. This leads to the first application area of this alignment method: its use for field-based methods and 3D QSAR [31, 32] where the alignment of molecules has a considerable impact on the results.

Due to the efficiency of the implementation and the advanced geometric similarity measure for chemical features, the presented algorithm is very fast and can be used on standard hardware for a variety of tasks. By basically comparing pharmacophore feature positions

Table 2 Atom-based RMS deviations of the ligand-based alignments of Sildenafil and Tadalafil relative to their protein alignments

	1UDT	1TBF	1XP0	1UHO
1UDT	–	0.7	0.9	0.3
1TBF	0.7	–	0.4	0.7
1XP0	0.8	0.4	–	0.8
1UHO	0.3	0.7	0.7	–

Table 3 Cyclooxygenase-2 (COX2) inhibitors from three PDB entries and atom RMS deviation in Å denoting the difference between the ligand-based super-positioning and the protein reference frame alignment

	1CX2	4COX	6COX
1CX2	–	2.6	0.4
4COX	2.4	–	2.7
6COX	0.4	2.8	–

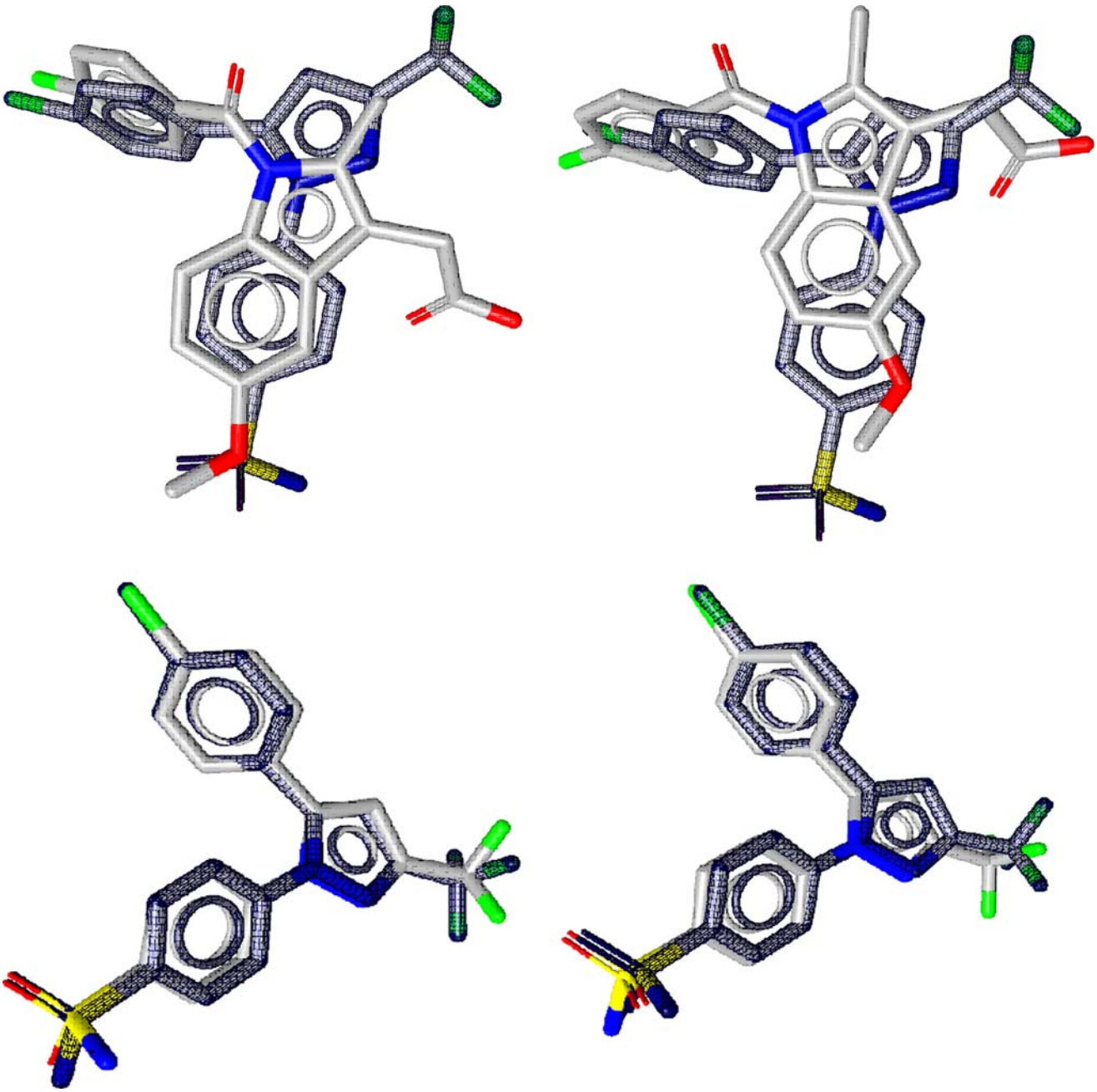

Fig. 10 Comparison of the ligand-based alignment (left) of Cyclooxygenase 2 (COX2) inhibitors to the experimentally determined protein-based alignment (right). Despite the relatively high RMSD value of 2.4 in the comparison of the ligands from PDB entries 4COX and 1CX2 (top), a high overlap of chemical features is perceived. In the case of the ligand from 1CX2 aligned to 6COX (bottom) a perfect alignment was calculated using the presented algorithm

instead of atom positions, a molecule can directly be compared to a 3D pharmacophore model and vice versa without changing the algorithm at all.

This leads to the second application area: the interactive creation of shared feature pharmacophores from several structure-based 3D pharmacophores by overlaying and interpolating the features the two pharmacophore have in common. A similar application is the creation of merged feature pharmacophores, which can be created by aligning two pharmacophores, adding features of both pharmacophores to a new one and interpolating overlapping features. Due to its speed this method has been included in a recent extension of the LigandScout software and can be used to interactively model pharmacophores and test their plausibility by interactively aligning them to their source molecules.

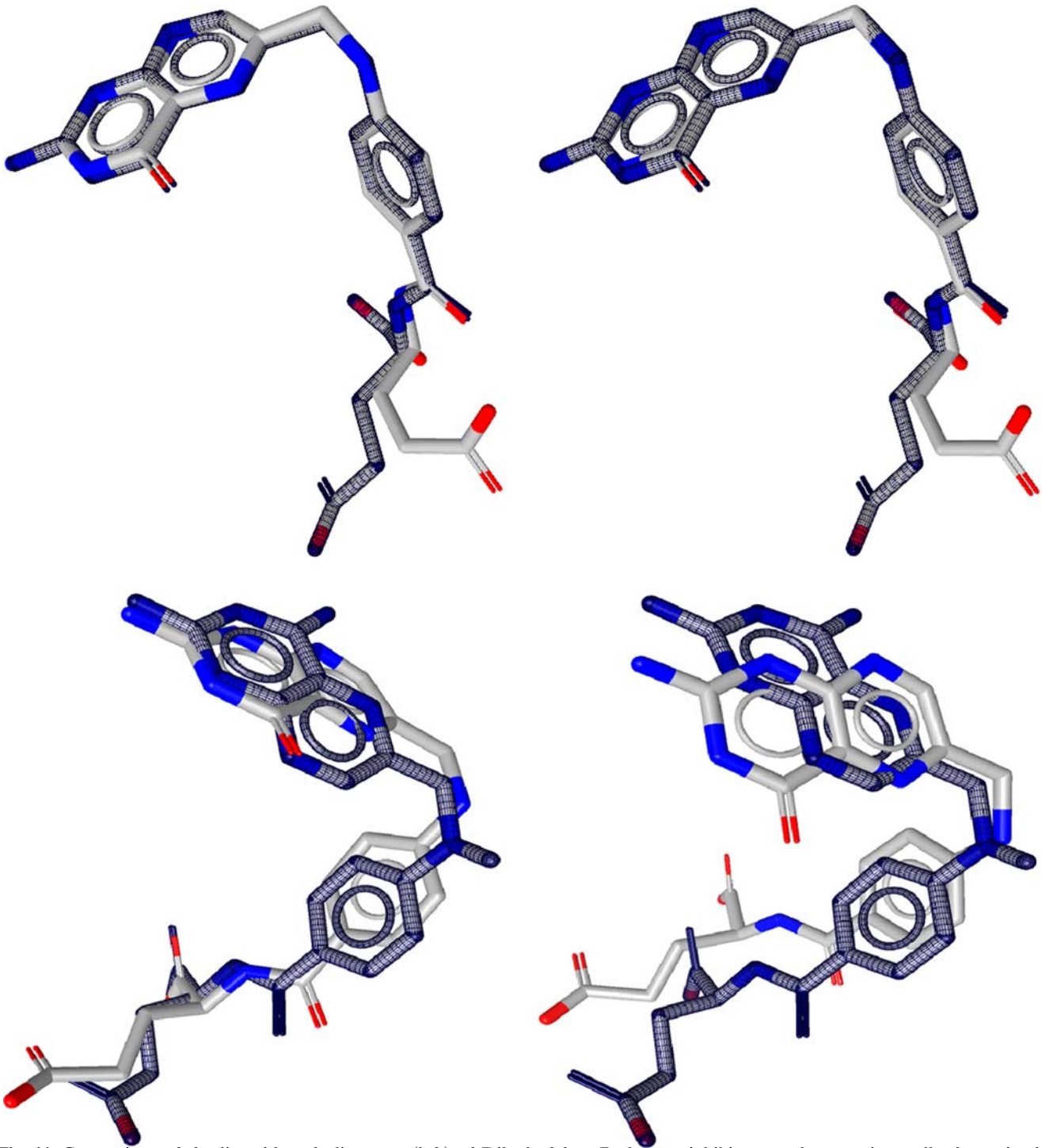

Fig. 11 Comparison of the ligand-based alignment (left) of Dihydrofolate Reductase inhibitors to the experimentally determined protein-based alignment (right). The two different conformations of Dihydrofolate align like the reference alignment, while the Methotrexate and Dihydrofolate alignment looks slightly different, but more plausible in terms of pharmacophore overlap

A controversial point about the presented approach is that the calculation results in only one alignment. However, this solution is guaranteed to be optimal within the defined representation of the problem. Due to inaccuracies in experimental data, the relevant and therefore correct solution may not be the algorithmic global optimum, but a solution that lies close to the computed solution. In the LigandScout implementation, this has been overcome by generating a set of solutions by combinatorially omitting up to two

Table 4 Atom RMS deviations from the reference alignment for Dihydrofolate Reductase inhibitors from the PDB entries 7DFR, 1RB3 and 1RX2

	7DFR	1RB3	1RX2
7DFR	–	2.3	0.2
1RB3	2.3	–	2.6
1RX2	0.3	2.6	–

Table 5 CDK2 inhibitors showing low deviations from the reference alignments

	1KE5	1KE6	1KE7	1KE8	1KE9
1KE5	–	0.4	0.6	0.2	0.3
1KE6	0.5	–	0.5	0.4	0.3
1KE7	0.6	0.6	–	0.7	0.3
1KE8	0.3	0.4	0.7	–	0.4
1KE9	0.3	0.3	0.3	0.3	–

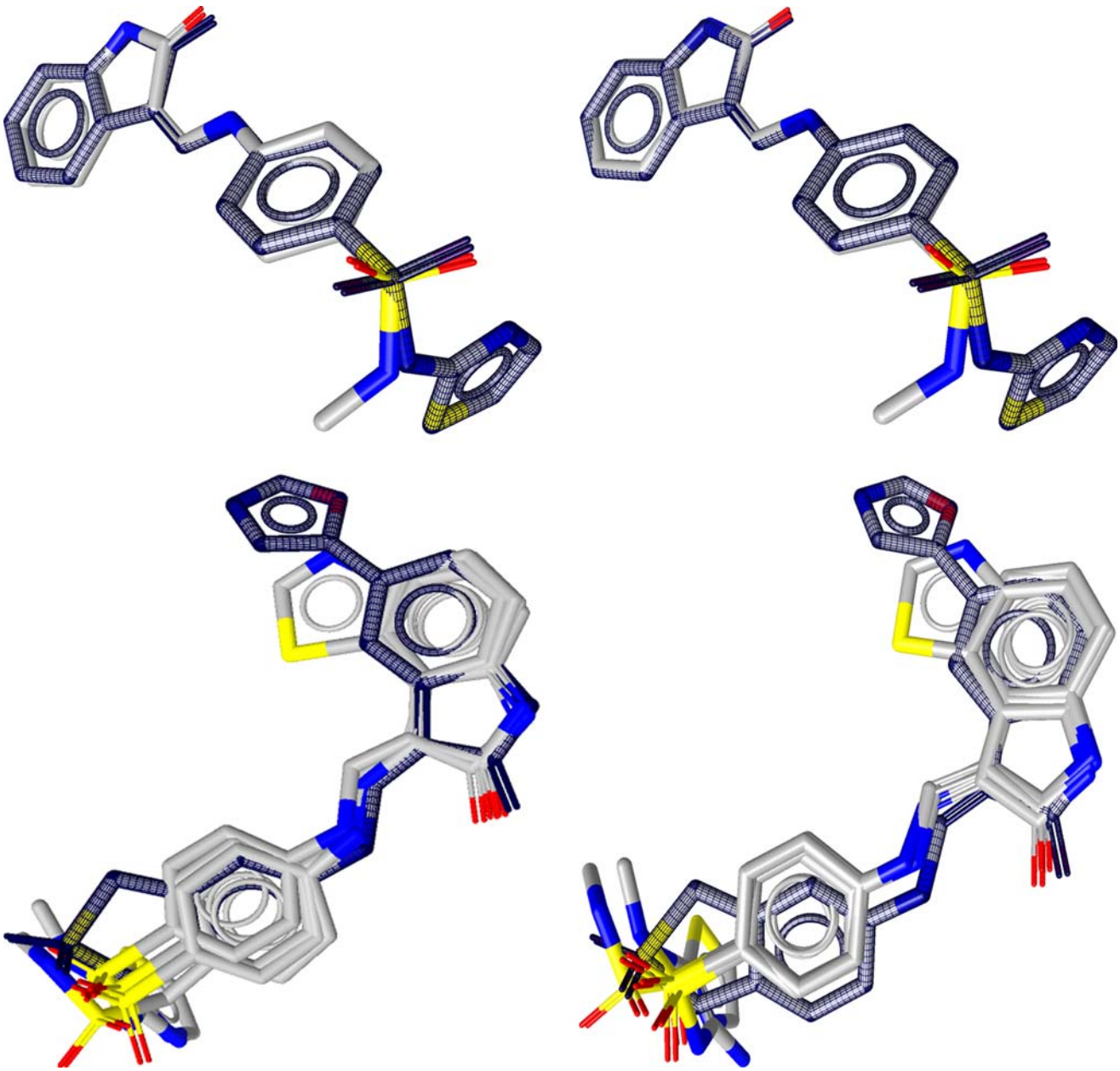

Fig. 12 CDK2 inhibitor alignments: The rigid molecules could be aligned on top of each other exactly like it was observed in the protein alignment

Table 6 PDB Ligands with experimentally determined coordinates were aligned to several artificial conformations generated by OMEGA

PDB identifier (4-letter code, 3-letter code, residue number)	Atom-based RMSD	Pharmacophore-based RMSD	Score	Number of matched features	Number of conformations
1cx2-s58-701	0.66	0.58	137.26	13	14
1fpu-prc-1	0.86	0.76	136.71	13	800
1iep-sti-201	1.04	0.67	146.73	14	800
1ke5-ls11	0.44	0.43	107.51	10	96
1ke6-ls2-201	0.44	0.35	137.95	13	474
1ke7-ls3-201	0.51	0.46	107.61	10	96
1ke8-ls4-2	0.37	0.27	138.19	13	800
1ke9-ls5-1	0.45	0.28	138.15	13	290
1rb3-mtx-161	1.49	0.71	185.49	18	800
1tbf-via-501	0.93	0.60	157.20	15	800
1xp0-vdn-201	0.83	0.56	167.33	16	800
4cox-imn-701	0.60	0.43	127.72	12	800
6cox-s58-701	0.48	0.48	137.56	13	14

A fully automated scoring function was used to select the best pharmacophore overlap, which forms the basis for the subsequent alignment. A low value for the measured atom-based RMSD in the second column denotes that the automatic alignment and scoring procedure was able to pick a conformation close to the reference structure.

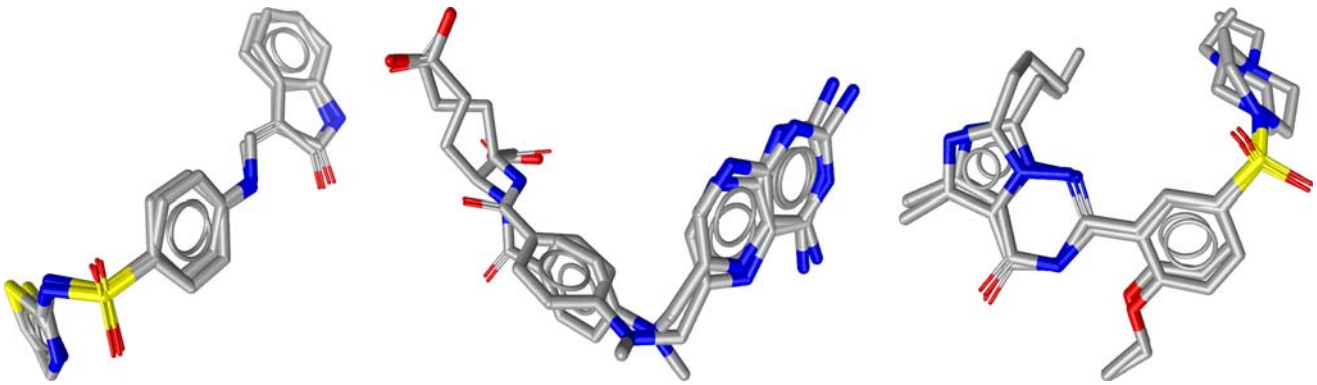

Fig. 13 Flexible alignment examples for 1ke8-ls42 (left), 1rb3-mtx (center), and 1xp0-vdn201 (right). Although the atom-based RMSD value is relatively high for 1rb3-mtx, the overlap of chemical features is still acceptable

Table 7 The time in standard CPU milliseconds required to perform alignments based on pharmacophore features of these CDK2 inhibitors

	1KE5	1KE6	1KE7	1KE8	1KE9
1KE5	–	15	17	11	18
1KE6	15	–	19	26	26
1KE7	19	23	–	22	13
1KE8	16	30	19	–	34
1KE9	21	28	18	28	–

chemical features, which still shows acceptable performance, but delivers more solutions, where the appropriate one can be selected manually in the graphical user interface. However, leaving it to the Hungarian Algorithm to find a single optimal solution, because it is not required to search the whole solution space, makes this algorithm that efficient. This has also been demonstrated by the pseudo-flexible alignment examples using multi-conformational models generated by OMEGA. This leads to the third application area, where manual decisions are difficult: due to its speed and easy representation of intermediate steps, this algorithm is an ideal candidate for implementing pharmacophore matching for virtual compound database screening in multi-conformational space.

Acknowledgments We thank Fabian Bendix and Robert Kosara (Inte:Ligand) for their excellent work on LigandScout as well as Christian Laggner, Johannes Kirchmair, Daniela Schuster, Theodora Steindl, and Eva Kleinrath (University of Innsbruck) for testing and helpful discussions.

References

1. Krovat EM, Fruhwirth KH, Langer T (2005) J Chem Inf Model 1:146
2. Laggner C, Schieferer C, Fiechtner B, Poles G, Hoffmann RD, Glossmann H, Langer T, Moebius FF (2005) J Med Chem 15:4754

3. Schuster D, Laggner C, Steindl TM, Palusczak A, Hartmann RW, Langer T (2006) J Chem Inf Model 3:1301
4. Schuster D, Laggner C, Steindl TM, Langer T (2006) Curr Drug Discov Technol 1:1
5. Steindl T, Laggner C, Langer T (2005) J Chem Inf Model 3:716
6. Böhm H-J, Klebe G, Kubinyi H (1996) Spektrum Akademischer Verlag
7. Lemmen C, Lengauer T (2000) J Comput Aided Mol Des 3:215
8. Martin YC, Bures MG, Danaher EA, DeLazzer J, Lico I, Pavlik PA (1993) J Comput Aided Mol Des 1:83
9. Bron C, Kerbosch J (1973) Commun ACM 9:575
10. Barnum D, Greene J, Smellie A, Sprague P (1996) J Chem Inf Comput Sci 3:563
11. Langer T, Krovat EM (2003) Curr Opin Drug Discov Devel 3:370
12. Langer T, Hoffmann RD (2001) Curr Pharm Des 7:509
13. DS Visualizer, version 1.5, available from Accelrys, Inc., 10188 Telesis Court, Suite 100, San Diego, CA 92121, USA
14. Jones G, Willett P, Glen RC (1995) J Comput Aided Mol Des 6:532
15. Prabhu NV, Zhu P, Sharp KA (2004) J Comp Chem 16:2049
16. EON, available from OpenEye Scientific Software (www.eyesopen.com), 3600 Cerrillos Rd., Suite 1107, Santa Fe, NM 87507, USA
17. Haigh JA, Pickup BT, Grant JA, Nicholls A (2005) J Chem Inf Model 3:673
18. ROCS, available from OpenEye Scientific Software (www.eyesopen.com), 3600 Cerrillos Rd., Suite 1107, Santa Fe, NM 87507, USA
19. Bostrom J (2001) J Comput Aided Mol Des 12:1137
20. Kirchmair J, Laggner C, Wolber G, Langer T (2005) J Chem Inf Model 2:422
21. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) Nucleic Acids Res 1:235
22. Kirchmair J, Wolber G, Laggner C, Langer T (2006) J Chem Inf Model 46:1848
23. Bostrom J, Greenwood JR, Gottfries J (2003) J Mol Graph Model 5:449
24. Richmond NJ, Willett P, Clark RD (2004) J Mol Graph Model 2:199
25. Kuhn HW (1955) Naval Res Logist Quart 2:83
26. Wolber G, Langer T (2005) J Chem Inf Model 1:160
27. Kabsch W (1976) Acta Crystal 922
28. Kabsch W (1978) Acta Crystal 827
29. Wolber G, Langer T (2001) In: Rational approaches to drug design, H.-D.H.W. Sippl, Editor. 2001, Prous Science: Barcelona, pp. 390–399
30. OMEGA, version 2.0, available from OpenEye Scientific Software (www.eyesopen.com), 3600 Cerrillos Rd., Suite 1107, Santa Fe, NM 87507, USA
31. Cramer RD 3rd, Patterson DE, Bunce JD (1989) Prog Clin Biol Res 161
32. Kubinyi HF, G, Martin YC (1998) Vol. 1–3, Kluwer/ESCOM, Dordrecht