Graph-Based Molecular Alignment (GMA)

J. Marialke, R. Körner, S. Tietze, and Joannis Apostolakis*

Institute for Informatics, Research and Educational Unit for Bioinformatics and Practical Informatics, Ludwig-Maximilians-University, Amalienstr. 17 D-80333, Munich, Germany

Received September 5, 2006

We describe a combined 2D/3D approach for the superposition of flexible chemical structures, which is based on recent progress in the efficient identification of common subgraphs and a gradient-based torsion space optimization algorithm. The simplicity of the approach is reflected in its generality and computational efficiency: the suggested approach neither requires precalculated statistics on the conformations of the molecules nor does it make simplifying assumptions on the topology of the molecules being compared. Furthermore, graph-based molecular alignment produces alignments that are consistent with the chemistry of the molecules as well as their general structure, as it depends on both the local connectivities between atoms and the overall topology of the molecules. We validate this approach on benchmark sets taken from the literature and show that it leads to good results compared to computationally and algorithmically more involved methods. The results suggest that, for most practical purposes, graph-based molecular alignment is a viable alternative to molecular field alignment with respect to structural superposition and leads to structures of comparable quality in a fraction of the time.

INTRODUCTION

In many drug design projects, the structure or even identity of the target is unknown, and the medicinal chemist has to rely on the information provided by a number of compounds known to have the desired activity. One of the available options for identifying structural and chemical features that are important for the activity, also called pharmacophores,[1] is to align sets of chemically related molecules that show different levels of activity and to correlate structural similarities and differences with activity. The underlying assumption is that the different molecules exert their function on the basis of a common structural/physicochemical principle, for example, a common binding mode to the same receptor. On the basis of this assumption, it is even possible to derive quantitative models for activity prediction: by aligning all molecules with respect to each other, a common reference frame is obtained which can be analyzed to provide quantitative structure−activity relationships (QSARs). This approach builds the basis of 3D QSAR and, most famously, of comparative molecular field analysis.[2]

Recently, a number of methods have been suggested for graph-based virtual screening. The suggested methods are very fast and lead to very good results as measured by enrichment, or the ability to place compounds of similar biological activity into the same cluster.

Stahl et al.[3] have suggested a new graph-based molecular similarity metric that takes molecular topology into account. Their metric is an extension of a maximum common subgraph (MCS)-based metric previously also used by Raymond and Willett.[4] The extension consists in penalizing topologically inconsistent MCS in a postprocessing step. Here, we take topology into account already during the identification of the MCS by defining and optimally solving a generalization of the MCS problem. A comparable algorithm has been recently presented by Barker et al., who have applied it on reduced graph-based similarity.[5] This algorithm was tested for enrichment in a number of different biological activities against the MDL Drug Data Report and was shown to outperform other measures of similarity such as Daylight fingerprints.[6]

As these and other papers show, graph-alignment-based molecular similarity methods are highly effective for clustering, pharmacophore identification, and virtual screening. In order to combine the efficiency of graph-based alignment with the additional information provided by the three-dimensional structure and physical properties of the molecules, we investigated a graph-based molecular alignment (GMA) approach and evaluated it with respect to its performance in predicting the biologically active conformation of molecules with similar biological activity. The method is based on a graph-matching procedure which produces a mapping between the query molecule and a template molecule (of known conformation). The mapping is used to superpose the molecules: we obtain the 3D coordinates of the query molecule by optimizing its conformation with respect to the root-mean-square deviation (RMSD) of the atoms that are mapped to the template. The optimization itself is obtained with a simple Levenberg−Marquardt algorithm for solving the RMSD minimization problem in torsion space.[7,8] Besides the higher efficiency of torsion space optimization, the second advantage of this approach is that it does not require additional constraints or energy terms to keep the conformation of the ligand physically reasonable, that is, with correct bond lengths and angles.

By applying the approach on benchmark sets where the conformations of both the templates as well as the query ligands are known, it is possible to compare the predicted conformation of the query ligand against its biologically

* Corresponding author e-mail: Joannis.Apostolakis@bio.ifi.lmu.de.

10.1021/ci600387r CCC: $37.00
Published on Web 01/10/2007

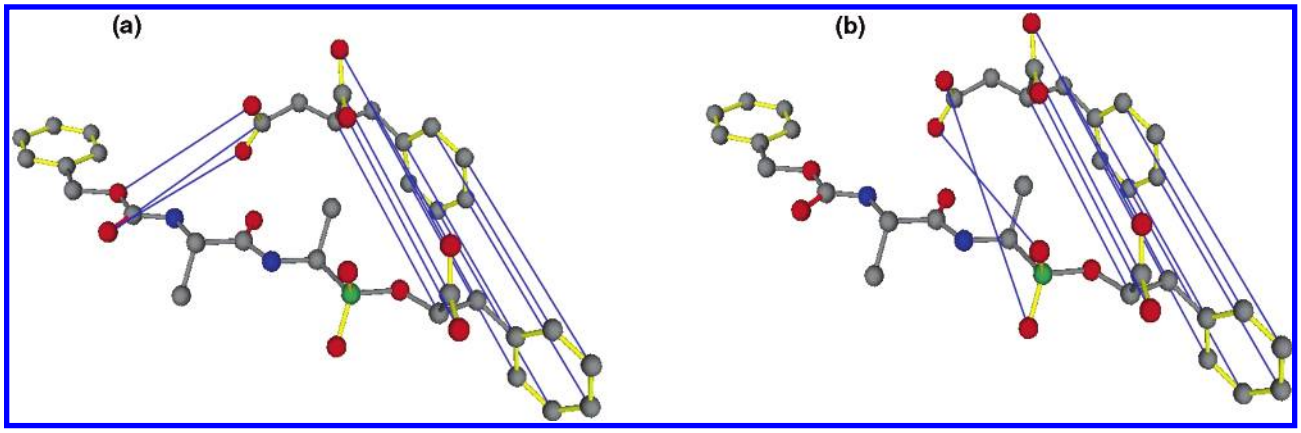

Figure 1. (a) Mapping between two different carboxypeptidase ligands with the original MCS consistency criterion. The blue lines connecting mapped atom pairs show how independent components of the common subgraph correspond to topologically unreasonable mappings. (b) Mapping between the same two molecules with the embedded MCS consistency criterion. The obtained mapping is smaller (since the C atom of the carboxylic moiety is not mapped) but still allows independent components in the common subgraph. Bond coloring convention: single bonds are gray; double bonds are red; bonds with partial π-bond character are yellow.

active conformation and thus to assess the accuracy of the conformation prediction. The benchmark data sets were taken from the literature.[9,10] The results reported in the literature for some other methods are rather anecdotal, in general showing three to five different alignments, and we will discuss a few of those other examples additionally.[11,12]

METHODS

In the current version of GMA, one of the molecules (the template) is kept rigid. Both sided flexibility is easily implemented; however, for the time being, we have avoided this, as we need the template to retain its biologically active structure, for the comparison to the X-ray structure overlay. GMA consists of three steps: a preprocessing step, which produces mappings between the query and template, the actual mapping of the dihedral angles, and the torsion space optimization of the flexible degrees of freedom of the query molecule.

In the preprocessing step, the molecules (query and template) are treated as graphs of connected atoms and the MCS of the two ligands is computed using a simplified variant of the RASCAL algorithm.[13] The MCS defines a one-to-one mapping between atoms in the query molecule graph and atoms in the template. We have found that the standard MCS is a poor choice for structurally mapping molecules. The reason for this is that, when the MCS has more than one independent component, topology between the corresponding components in the two molecules is not always preserved. For linear molecules, that would mean a mapping of A−B−C to A−C−B (where the uppercase letters denote independent components of the MCS, and a dash stands for nearest neighbor in terms of graph distance). A simpler kind of problem is shown in Figure 1a: the MCS is a subgraph consisting of two independent components, which are close to each other in one of the molecules and significantly further away in the other. This type of problem has recently also been discussed in the work by Stahl et al.[3]

This problem arises from the fact that the definition of MCS is based on the identification of maximal isomorphic subgraphs between the two molecules, and that the definition of graph isomorphism (and hence also subgraph isomorphism) takes only direct connectivity into account.

A graph G is an ordered pair of two sets (V, E), where V is the set of nodes and E is a set of edges with $e \in E$, and $e = (u, v)$, where $u \in V$ and $v \in V$. Two graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ are isomorphic if and only if there exists at least one bijective mapping $f: V_1 \rightarrow V_2$, from the node set of the first graph to the node set of the second graph such that the existence of an edge between two vertices in the first graph is equivalent to the existence of an edge between their mappings in the second graph, that is, $(v_1, v_2) \in E_1 \leftrightarrow (f(v_1), f(v_2)) \in E_2$. We call the pair $(v_1, f(v_1))$ a vertex mapping from G_1 to G_2. Thus, the isomorphism condition can be reformulated as follows: for any pair of mappings $(v_1, f(v_1))$ and $(v_2, f(v_2))$, we have either $(v_1, v_2) \in E_1 \wedge (f(v_1), f(v_2)) \in E_2$ or $(v_1, v_2) \notin E_1 \wedge (f(v_1), f(v_2)) \notin E_2$. For labeled graphs, the function f is allowed to map only vertices (and edges) of an identical type (label). In the context of molecular similarity, the above definition simply implies that any pair of mapped atoms is either connected by a bond in both molecules or not directly connected in either of the two molecules. The relative topology of independent components of the common subgraph is thus not necessarily maintained, leading to the mappings seen in Figure 1a.

To solve this problem, we use a variant of graph isomorphism, which we will call embedded subgraph isomorphism and which leads to topologically correct mappings. In embedded subgraph isomorphism, the main idea is the use of a stricter consistency condition, which takes into account not only the direct connectivity but also the equivalence of graph distances within the original graphs to assess the consistency of mapping. As was described above, in MCS, a vertex mapping is consistent with a second mapping if the mapped pairs of vertices are either both connected by an edge or both not connected by an edge in their respective graphs. In the definition of embedded subgraph isomorphism, the consistency between two mappings is given by the condition that the two pairs of mapped vertices have a similar distance in the original graphs (G_1 and G_2), where distance and distance similarity still need to be defined. For example, by defining two similarity classes, one for distances equal to one (vertices connected by a single edge) and one for all others, we recover the original formulation of subgraph isomorphism. Thus, embedded

subgraph isomorphism is a generalization of subgraph isomorphism (and therefore also nondeterministic polynomial-time hard).

For the distance similarity definition, there exist at least three obvious variants: shortest path length identity, shortest path length similarity, where path lengths are allowed to deviate within relative or absolute limits, and the identity or similarity of any of all possible nonrecurring path lengths.

In principle, the first definition is quite strict, and the two other variants have been introduced to weaken it and allow more structural diversity within the common subgraph. The second variant corresponds in a sense to gapped alignment, while the third variant explicitly makes use of multiple paths for the definition of a relatively tolerant yet still chemically reasonable consistency. It has to be mentioned that every weakening of the similarity definition leads to an increase in time complexity. We will note here that it is further possible to use distances that do not derive from the graph itself, but from the actual conformation of the molecule.

With first tests we have found that the first variant is too strict while the third rarely leads to significantly better results. The focus of the present work lies, therefore, on the second variant, which corresponds to a compromise between mapping quality and efficiency. The gapped length constraint is set as follows: a pair of nodes from the first graph is consistent (can be mapped) with a pair of nodes from the second graph if their labels are mutually identical and both pairs are either direct neighbors in the respective graphs or their graph distance differs by a threshold of 5 at most. A threshold of a certain size allows local topological inconsistencies that correspond to inversions of groups of that size. The choice of the threshold is a compromise between allowing local misfits and retaining global topology.

The identification of maximum embedded common subgraphs (EMCSs) is chemically intuitive, since it takes the comparison from finding identical independent subgraphs (chemical groups) to comparing also the relative connectivity (topology) between the groups. EMCSs corresponding to the first and second distance similarity variant are by definition not larger than MCS (since they fulfill more constraints). At the same time, they are in general larger than the largest connected common subgraph, since connected common subgraphs are in most cases also consistent with the shortest path distance criterion. The exception arises when the shortest paths between vertices in the connected common subgraph lead over vertices that are not contained in it. For example, the largest connected common subgraph of pentane and cyclopentane is a butane, while according to the shortest path distance constraint, the MCS is only a propane.

The example in Figure 1b shows the effect of the length constraint on the MCS. The mapping obtained, though smaller by one vertex, provides a significantly better basis for alignment between the two molecules. Stahl et al.[3] have recognized the topology problem and addressed it by postprocessing all obtained MCSs and penalizing poor topology. The problem with that approach is that, as shown in the example in Figure 1, the constraint on correct topology leads in general to smaller MCSs which are not part of the unconstrained solution set. Thus, in many cases, all obtained MCSs are topologically inconsistent and are penalized accordingly, even though slightly smaller topologically consistent solutions exist, which are not found although they would score higher than the larger yet inconsistent MCSs.

The identification of the maximum embedded subgraph is obtained with DIZZEE, our slimmed down version of RASCAL,[13] a branch and bound algorithm for finding cliques in the product graph of the two graphs that are being compared. The exact algorithm used for the identification of the maximum common subgraph is not particularly important as long as the solution is computationally efficient. We will, therefore, simply mention here the main ideas behind the identification of the EMCS in DIZZEE and RASCAL.

First, a so-called product or association graph is built on the basis of the two graphs that are being compared. Assuming we are looking for the MCS between two labeled graphs G_1 and G_2, the product graph M contains one node m for every pair of vertices with equivalent labels. Thus, each node m in M defines a mapping of a node in the first graph G_1 to a node in the second graph G_2. We define two projection functions, $g_1(m)$ and $g_2(m)$, that recover the nodes mapped by m from G_1 and G_2, respectively. Two nodes m and m' in M are connected by an edge if either $(g_1(m), g_1(m')) \in E_1 \wedge (g_2(m), g_2(m')) \in E_2$ or $(g_1(m), g_1(m')) \notin E_1 \wedge (g_2(m), g_2(m')) \notin E_2$; that is, the mappings defined by nodes m and m' define a common subgraph of size 2. In the case of EMCS, the condition is substituted by $D(g_1(m), g_1(m')) \approx D(g_2(m), g_2(m'))$, where $D(...)$ is the distance function and is the similarity relation. We say that m and m' are consistent. A clique of size k in M consists of k single-node mappings that are all consistent with each other and, therefore, define a common subgraph of size k. Conversely, any common subgraph of size k in G_1 and G_2 consists of k consistent nodes and therefore corresponds to a clique of size k in M. The maximum clique in M yields the MCS for graphs G_1 and G_2.

The main efficiency-enhancing trick in RASCAL is the partitioning of the product graph into sets consisting of independent nodes. The cliques are then built up by choosing nodes from the smallest independent set. It is important to note that RASCAL was originally developed and implemented for line graphs, derived from the molecules. Line graphs are derived in the following way: Edges (bonds) in the original graph correspond to vertices in the line graph. Two vertices in the line graph are connected by an edge, if the corresponding edges in the original graph have a common vertex. Line graph vertices are labeled by the atom types of the two atoms forming the bond and the bond type and are thus much more specific than vertices in the original graph. This leads to fewer nodes in the product graph, hence increasing the efficiency of the MCS solution. We have implemented DIZZEE for both the original as well as the line graphs. Again, initial tests have shown that the line graph approach is more efficient. In contrast to the original work, we do not consider atomic hybridization in the mapping; that is, the label of line graphs consists simply of the lexicographically ordered (to avoid order dependence) element types of the elements aligning the corresponding bond. Bond multiplicities are also not annotated in the label. This leads to somewhat slower comparisons (because more line graph nodes are compatible with each other) but also to higher sensitivity. DIZZEE has further been extended to use weighted matching for the underlying graph alignment. This

introduces the possibility of taking into account the physical properties that describe interaction similarity as well as the possibility of training the weights in order to obtain given alignments. However, the overall result was found to be relatively robust over parameter variation (such as weighting, type of the length constraint, and so on), and therefore, the results reported here will be limited to the unweighted MCS.

DIZZEE returns the maximal mappings between the two molecular graphs. For each mapping, the molecules are then flexibly superposed: The atomic mapping also defines the corresponding torsional degrees of freedom in the molecules and thus allows a very direct treatment of flexibility. The mapped torsion angles are "stamped"; that is, each mapped torsion angle in the query molecule is set to the angle of the corresponding torsion angle in the template molecule. A pair of torsion angles from the query and the template molecule is mapped when there exists a set of four mapped and linearly connected atoms in the test molecule with the two middle atoms being connected by the central bond of the torsion angle. This procedure is not strictly necessary. The results change only insignificantly by removing this step. However, it often leads to faster convergence of the final minimization, or even completely avoids the minimization altogether, if the obtained alignment between template and query is good enough at this stage. In the application of the method, we have noticed that the use of all mappings does not improve results, while it can lead to a significant slow down, when many symmetry-related mappings exist (which is quite common). The reason for the low effect on the accuracy lies with the fact that most maximal mappings are very similar. For the benchmarks in the Results section, we have used only the first mapping returned by the method.

Once the conformation of the query ligand has been determined, it still needs to be superposed to the template. Since we have an explicit atomic mapping between the template and query molecule, we can use the RMSD of the mapped atoms to calculate the quality of a superposition. This is more efficient than the calculation of overlap integrals of Gaussian functions.[14] The RMSD is a simpler function, and furthermore, it depends linearly on the number of atoms in the molecules as opposed to SEAL type functions,[14] which scale with the product of the number of atoms in the molecules. The most important advantage of the RMSD is, however, that the optimal rigid structural superposition between two molecules according to RMSD can be calculated in a single step, for example, with the Kabsch algorithm,[15] while solving the corresponding problem for overlap integrals is significantly more complex.[16] The query molecule is thus rigidly superposed on the template molecule and scored to obtain a reference RMSD.

Finally, if the RMSD of the mapped atoms lies higher than a certain threshold (0.5 Å), we additionally perform a torsion space optimization of the query molecule with the RMSD of the mapped atoms as the objective function. The optimization takes place in two parts: in the first, only the RMSD is optimized; in the second, a simple clash term is added to the RMSD score, to produce structures with reasonable geometries.

RESULTS

As a main benchmark set, we have used the fFlash data set, as used by Krämer et al.[10] This data set is a subset of the benchmark used for validating FlexS.[9] It consists of 7 groups of ligands. The ligands in each group are more or less similar to each other and bind to the same binding site. X-ray structures of the bound conformation exist for each ligand. Thus, the benchmark allows the evaluation of the alignment with respect to a known binding mode. In each group, every ligand is used in turn as a template for all of the other ligands. By aligning the query ligands to the template in its bound conformation in the binding site, we obtain a pose for the query ligands in the binding pocket of the protein, which can be compared to the known X-ray structure of the query ligand in complex with the protein.[9] The obtained RMSDs offer a direct measure of the predictive quality of the structural superposition. We compare our results on this benchmark with the results of fFlash[10] and FlexS.[9] In the discussion of the results, we will omit explicit comparison of the self-alignment results, since these are in general of only technical relevance, that is, to show that the alignment can solve the trivial problem correctly. As Table 1 shows, this is the case for all algorithms in almost all cases and never a problem for GMA. The fact that the self-alignment is never exactly 0 is due to the use of slightly different conformations for superposition and comparison. In the former case, the templates have been locally minimized with a gradient steepest descent method with the Sybyl force field,[9,17] while in the latter, the ligand conformations are taken directly from the Protein Data Bank structure.

Thrombin Ligands. In the example of 1dwd and 1dwc, the ability of the approach to superpose molecules that differ significantly at the graph level is demonstrated. A little bit over 70% of the atoms can be matched, which is less than in most other successful alignments (see Table 1). The main interacting moieties differ between the two ligands, and the "backbone" is shorter in 1dwc. This leads to the shortcut taken by the aliphatic linker of the guanidine group in 1dwc (Figure 2a), which as seen in Figure 2b is indeed a correct prediction. Allowing gaps is essential for obtaining a correct mapping in this case. With an exact length constraint that allows no gaps, the obtained alignment is quite poor (RMSD > 4 Å).

Rhinovirus Ligands. This group of long, linear ligands contains two heterocyclic and one aromatic ring that are connected via a long flexible carbon chain. In Figure 3a and b, the effect of gapped alignment can be seen as the formation of a "loop" in the middle of the aliphatic chain of the 2r04 ligand. The loop consists of unmatched atoms and allows the perfect superposition of the ring moieties at the ends of the ligands. While the comparison to the crystal structure of 2r04 shows this alignment to be suboptimal in this case, the obtained accuracy is reasonably good, with a RMSD of 1.96 Å.

The ligands show a certain degree of pseudo symmetry which leads to one pair of ligands (2r04 and 2r06) showing an inverse binding mode compared to the other pair (2rr1 and 2rs3). This has been previously noted in the original literature[18] and leads to poor results in terms of RMSD for all methods in the comparison (see Table 1). As can be seen in the comparison of the alignment between 2rs3 and 2rr1 (Figure 3c and d) and the alignment between 2r04 and 2rr1 (Figure 3e and f), the internal group similarity appears to be comparable to the cross group similarity, so that ligand

Table 1. Comparison of GMA vs fFlash[10] And FlexS[9] with Respect to the RMSD of the Solution to the Known Crystal Structure[a]

Query	Template	Flash (best 10)	Flash (best rank)	FlexS (best 10)	GMA	%matched	Time Flash	Time GMA
Thrombin								
1dwc	1dwc	1.28	1.32	0.60	0.30	1.00	2.70	0.27
1dwd	1dwc	1.93	1.93	1.44	1.68	0.71	2.20	0.89
1dwc	1dwd	2.28	2.42	1.64	1.69	0.74	0.50	1.11
1dwd	1dwd	1.86	2.26	0.45	0.48	1.00	0.60	0.21
							3.75	**2.49**
Rhinovirus								
2r04	2r04	1.37	1.37	0.67	0.32	1.00	1.50	0.07
2r06	2r04	1.80	1.90	1.30	1.33	0.96	1.50	0.06
2rr1	2r04	R	R	10.00	13.32	0.96	1.10	0.07
2rs3	2r04	R	R	10.00	13.91	0.93	1.30	0.11
2r04	2r06	1.49	2.56	0.78	1.96	0.88	0.30	0.09
2r06	2r06	1.28	1.96	0.44	0.11	1.00	0.40	0.06
2rr1	2r06	R	R	10.00	12.82	0.85	0.30	0.10
2rs3	2r06	R	R	10.00	13.08	0.81	0.30	0.13
2r04	2rr1	R	R	10.00	12.82	1.00	1.60	0.06
2r06	2rr1	R	R	10.00	11.55	0.96	2.10	0.07
2rr1	2rr1	1.21	1.64	0.73	0.14	1.00	1.50	0.07
2rs3	2rr1	1.66	2.16	0.68	0.58	0.96	1.30	0.11
2r04	2rs3	R	R	10.00	12.96	1.00	2.20	0.06
2r06	2rs3	R	R	10.00	11.59	0.96	2.30	0.06
2rr1	2rs3	1.29	1.29	0.90	0.43	1.00	1.80	0.07
2rs3	2rs3	1.64	1.92	0.69	0.14	1.00	1.70	0.07
							13.25	**1.33**
Fructose								
4fbp	4fbp	0.52	0.52	0.51	0.47	1.00	0.20	0.07
T0039	4fbp	0.61	0.61	1.75	1.42	0.90	0.20	0.08
4fbp	T0039	0.57	0.57	0.61	1.91	0.93	0.10	0.17
T0039	T0039	0.61	0.61	1.51	0.55	1.00	0.10	0.08
							0.37	**0.41**
Dhfr								
1dhf	1dhf	1.54	1.62	0.48	0.45	1.00	0.10	0.18
4dfr	1dhf	2.30	2.69	1.36	2.45	0.89	0.10	0.20
1dhf	4dfr	1.73	1.73	1.47	2.35	0.92	0.20	0.19
4dfr	4dfr	1.84	1.84	0.46	0.68	1.00	0.20	0.18
							0.37	**0.76**
Query	**Template**	**Flash (best 10)**	**Flash (best rank)**	**FlexS (best 10)**	**GMA**	**%matched**	**Time Flash**	**Time GMA**
Thermolysin								
1tlp	1tlp	1.48	1.76	0.86	0.87	1.00	92.00	0.18
1tmn	1tlp	3.43	3.54	1.24	1.68	0.82	9.40	0.26
2tmn	1tlp	1.10	1.88	0.79	1.08	0.88	1.10	0.03
3tmn	1tlp	1.25	1.36	0.95	1.05	0.89	5.30	0.08
1tlp	1tmn	2.94	3.11	1.34	2.30	0.74	11.80	0.50
1tmn	1tmn	1.85	2.20	0.72	0.88	1.00	20.70	0.18
2tmn	1tmn	1.24	2.00	0.63	1.49	0.65	0.80	0.06
3tmn	1tmn	1.24	1.24	0.78	1.09	0.89	3.80	0.07
1tlp	2tmn	X	X	12.00	2.75	0.35	0.10	0.3
1tmn	2tmn	X	X	12.00	6.84	0.28	0.10	0.21
2tmn	2tmn	0.97	1.42	0.37	0.37	1.00	0.10	0.06
3tmn	2tmn	X		12.000	5.76	0.37	0.10	0.10
1tlp	3tmn	3.86	3.86	12.00	4.02	0.56	0.10	0.26
1tmn	3tmn	4.10	4.10	12.00	4.18	0.64	0.10	0.22
2tmn	3tmn	1.22	1.34	3.01	3.19	0.59	0.10	0.06
3tmn	3tmn	1.00	1.25	1.15	0.37	1.00	0.10	0.07
							91.06	**2.77**
Carboxyptd-a								
1cbx	1cbx	0.88	0.88	0.44	0.10	1.00	0.10	0.03
6cpa	1cbx	X	X	12.00	5.30	0.31	0.10	0.29
7cpa	1cbx	X	X	12.00	8.49	0.30	0.10	0.30
1cbx	6cpa	1.03	1.03	0.78	0.66	0.73	0.30	0.0
6cpa	6cpa	1.37	1.61	1.31	0.32	1.00	11.60	0.15
7cpa	6cpa	X	X	0.93	2.13	0.81	10.60	0.30
1cbx	7cpa	1.13	1.13	0.88	8.50	0.87	0.80	0.05
6cpa	7cpa	1.39	1.60	0.69	0.68	1.00	24.30	0.13
7cpa	7cpa	2.32	2.40	1.25	0.17	1.00	46.60	0.24
							59.06	**1.56**
Elastase								
1ela	1ela	0.89	1.46	0.62	0.35	1.00	1.00	0.26
1ele	1ela	1.22	1.24	0.65	0.79	0.84	0.87	0.12
1ela	1ele	1.08	1.98	0.96	0.86	0.73	0.87	0.19
1ele	1ele	1.200	2.15	0.37	0.35	1.00	1.00	0.11
							2.33	**0.70**

[a] R marks reverse alignment; green = RMSD < 1.5; yellow = RMSD < 3.0; red = RMSD >3.0. %matched value corresponds to the percentage of atoms in the query molecule that are matched to an atom in the template. Below each group of ligands, the accumulated times for each group are given. For comparability purposes, the accumulated times for fFlash have been divided by 1.67 as discussed in the text.

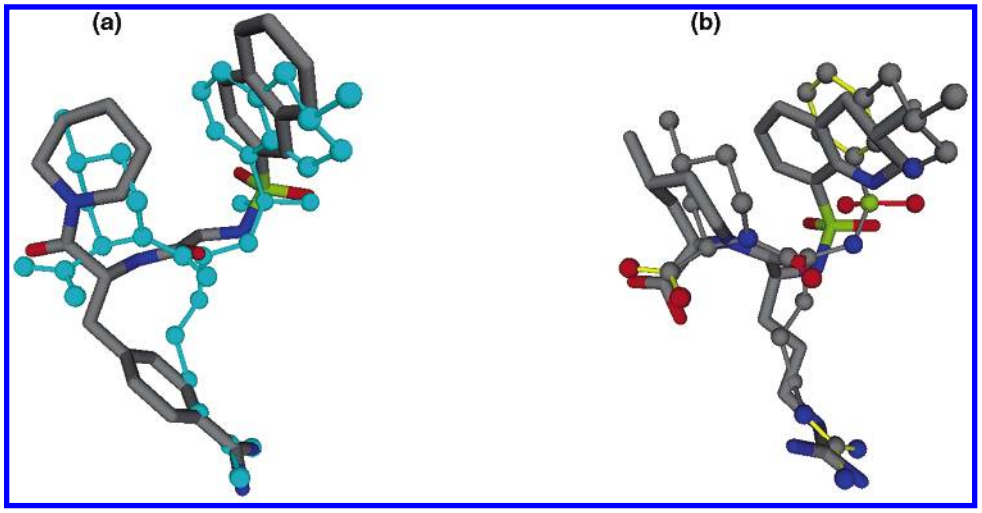

Figure 2. (a) 1dwc ligand (light-blue ball-and-stick) aligned on 1dwd ligand (thick sticks). (b) Comparison between aligned pose (ball-and-stick) and crystal structure for ligand 1dwc (thick sticks), RMSD = 1.69 (see also Table 1).

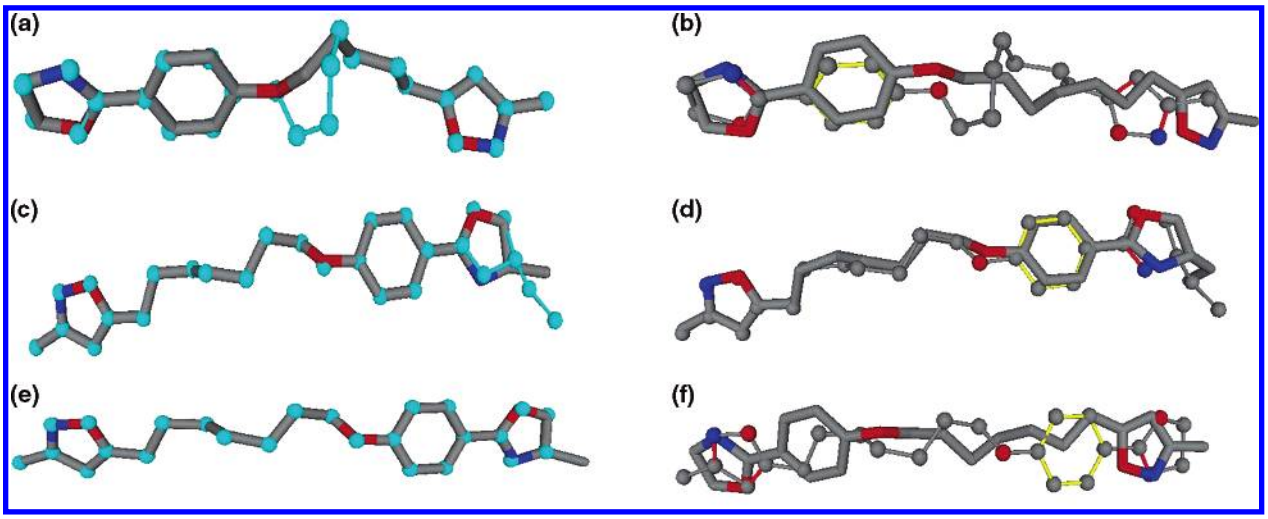

Figure 3. (a) 2r04 ligand (light-blue ball-and-stick) aligned on 2r06 (thick sticks). (b) Comparison between aligned pose (ball-and-stick) and crystal structure for ligand 2r04 (thick sticks), RMSD = 1.96 Å. (c) 2rs3 ligand (light-blue ball-and-stick) aligned on 2rr1 ligand (thick sticks). (d) Comparison between aligned pose (ball and stick) and crystal structure for ligand 2rs3 (thick sticks), RMSD = 0.59 Å. (e) 2r04 ligand (light-blue ball-and-stick) aligned on 2rr1 ligand (thick sticks). (f) Comparison between aligned pose (ball-and-stick) and crystal structure for ligand 2r04 (thick sticks), RMSD = 12.82 Å.

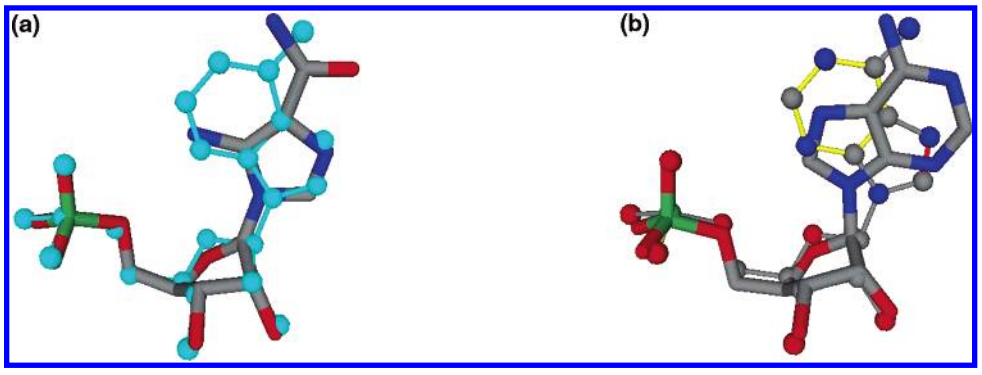

Figure 4. (a) 4fbp (light-blue ball-and-stick) aligned on t0039 ligand (AMP, thick sticks). (b) Comparison between aligned pose (ball-and-stick) and crystal structure for 4fbp (thick sticks), RMSD = 1.91 Å.

structure does not explain differences in binding orientation (Figure 3d and f).

Fructose Bisphosphatase. Adenosin monophosphate (AMP) is aligned to t0039, a ligand in which the purine base is substituted by an imidazole substructure. The actual alignment (Figure 4a) appears very good, matching either identical or homologous atom types in all cases except one: the aromatic nitrogen atom (N_3 of the purine ring) of AMP is placed over an amino group (H-bond donor) of t0039, thus placing an H-bond acceptor over an H-bond donor. Rotating the purine ring by 180° around the bond to the ribose avoids this problem but gives a poorer steric fit. The comparison to the crystal structure appears to suggest that the H-bonding pattern is more important in this case. Clearly, the relative

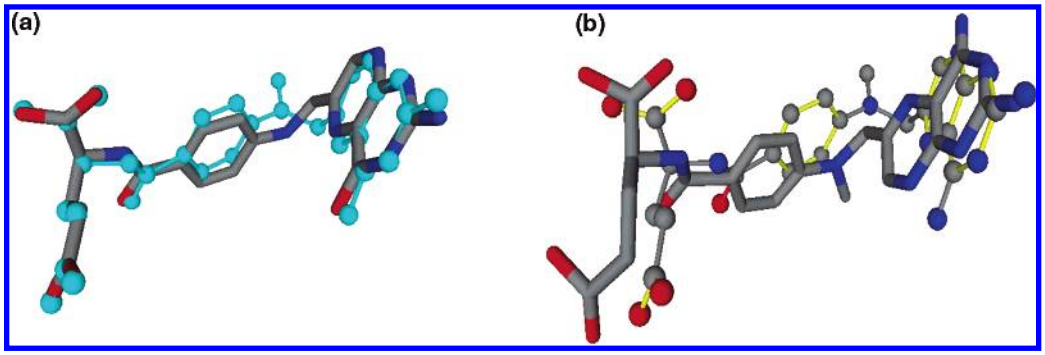

Figure 5. (a) 4dfr ligand (light-blue ball-and-stick) aligned on 1dhf ligand (thick sticks). (b) Comparison between aligned pose (ball-and-stick) and crystal structure for ligand 4dfr (thick sticks), RMSD = 2.45 Å.

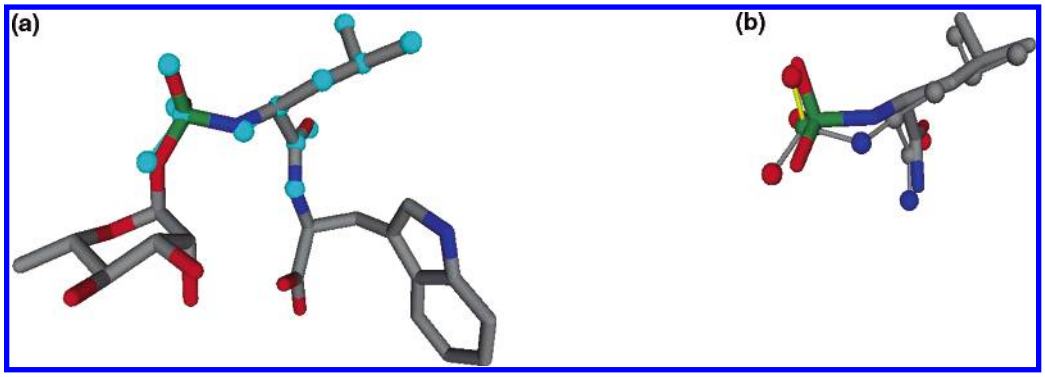

Figure 6. (a) 2tmn ligand (light-blue ball-and-stick) aligned on 1tlp ligand (thick sticks). (b) Comparison between aligned pose (ball-and-stick) and crystal structure for ligand 2tmn (thick sticks), RMSD = 1.08 Å.

weighting of steric over H-bond properties for the comparison depends also on the target. It is easily possible to increase the relevance of H-bond patterns by including polar H atoms or using weighted MCS. However, the reliance on H-bond patterns postulates exact knowledge of the protonation pattern and the tautomeric state of the ligands. These are not trivially calculable and may further depend on the microenvironment in the binding site of the target (which in general is not known).

Dihydrofolate Reductase (Dhfr). The two different ligands in dihydrofolate reductase share a similar structure (Figure 5 left part) but show different binding modes. The pteridine ring is rotated by 180° around the bond linking it to the rest of the structure, leading to significant differences in the prediction of the binding conformation and an overall shift of the molecule. This fact explains the significant RMSD in the comparison between predicted and crystal structure for the two ligands. This is another case where the hydrogen-bonding pattern appears more important than the steric overlap.

Thermolysin Ligands. Of the four thermolysin ligands, two (2tmn and 3tmn) are substantially smaller than 1tlp and 1tmn. This is reflected by the results in the benchmark run. The large molecules align well on each other. A reasonable alignment is found for the superposition of the small compounds on larger template structures (see Figure 6). Because of the different sizes of the ligands, the alignment of the large molecule on the smaller ligand fails (see Figure 7), since not enough information is given by the alignment to place the complete ligand. This is clearly a general problem, which is method-independent as is evidenced by the failure of all three methods to produce reasonable results for these cases (Table 1).

Carboxypeptidase-a Inhibitors. The carboxypeptidase ligand 1cbx is a fragment of the larger molecules 6cpa and 7cpa and can easily be mapped to both ligands. It turns out that the larger 7cpa ligand has three groups, of equal similarity to 1cbx. As only one mapping is used for the superposition, the mapped group is arbitrarily chosen. Our method only returns the first solution (out of three possible; Figure 8).

The subgraph isomorphism identifies the larger molecules 6cpa and 7cpa as very similar and produces a very good alignment (RMSD avg. < 1.5; Figure 9).

Elastase Inhibitors. The ligands inhibiting elastase, 1ela and 1ele, are of similar structure where 1ela is a substructure of 1ele (Figure 10). It has to be noted that the lysine side chain in 1ele cannot be aligned, since the corresponding group in 1ela is significantly shorter and does not provide any information with respect to the conformation of the side chain. Thus, our solution as well as the other reported placements in Table 1 are either fortuitous or simply due to sampling around solutions with an optimal score and picking out the one with the best RMSD.

Overall Comparison between fFlash, FlexS, and GMA. In the comparison between fFlash, FlexS, and GMA (Table 1), the FlexS results seem better than those of the other two methods in terms of RMSD, while GMA is on average better than fFlash (Table 1). It is interesting to note that after a reviewer suggestion we calculated the graph-based similarities between the template and query, and it turns out that the relative size of the mapping (%matched in Table 1) is a very good indicator of the probability of any method finding the correct structure. Comparisons with less than 50% similarity practically never lead to good alignments compared to the crystal structure. The similarity to the correlation

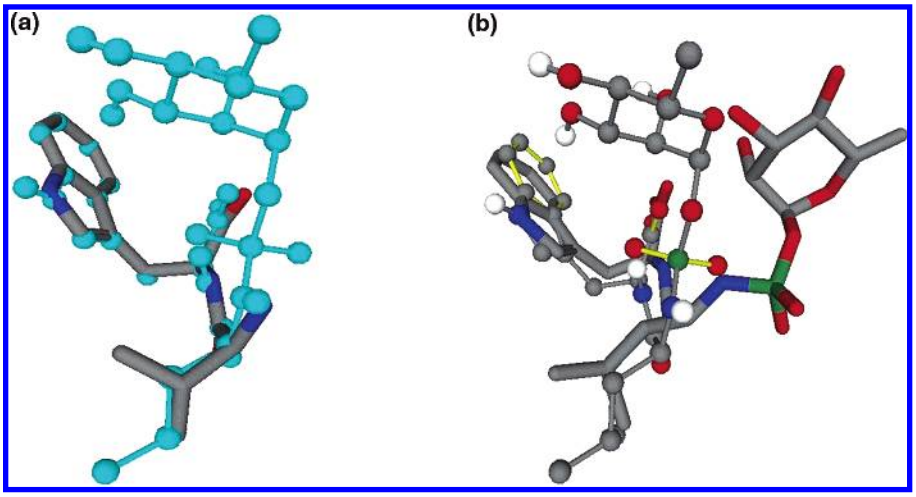

Figure 7. (a) 1tlp (light-blue ball-and-stick) aligned on the significantly smaller ligand 3tmn. The mapped subgraph is scored with a RMSD < 0.3. (b) Comparison of the crystal structure of the 1tlp ligand (thick stick) vs the aligned pose (ball-and-stick). The carbohydrate part of the ligand cannot be aligned and lies free in space, resulting in a RMSD > 3.0 Å.

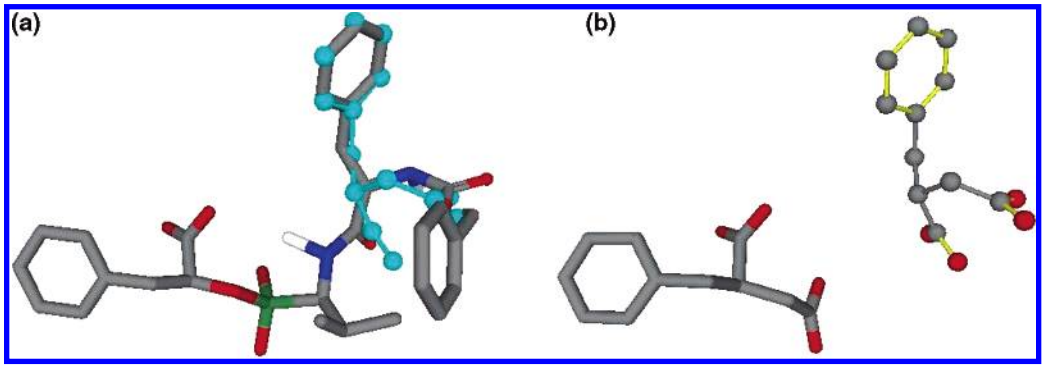

Figure 8. (a) 1cbx ligand (light-blue ball-and-stick) aligned on 7cpa ligand (thick sticks). (b) Comparison between aligned pose (ball-and-stick) and crystal structure. The mapped structure is aligned on the wrong branch of 7cpa, RMSD > 8.5 Å.

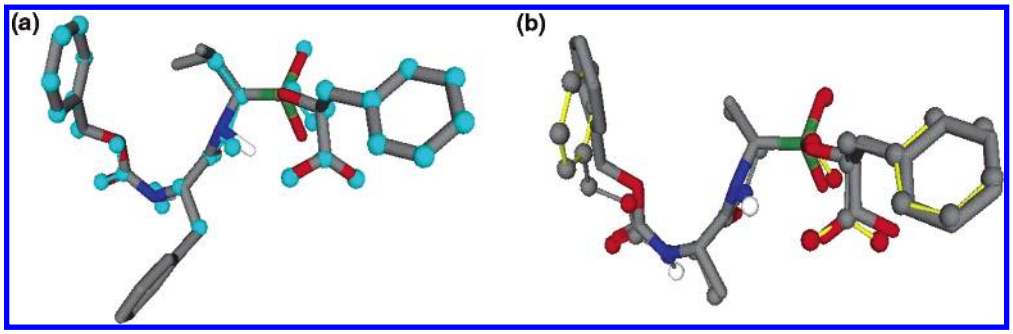

Figure 9. (a) 6cpa ligand (light-blue ball-and-stick) aligned on 7cpa ligand (thick sticks). (b) Comparison between aligned pose (ball-and-stick) and crystal structure for ligand 6cpa (thick sticks), RMSD = 0.68 Å.

between sequence similarity and the quality of protein homology models is striking. The difference among the three methods is not very large and appears marginally significant. A sign test comparing the results of FlexS with GMA indicates no significant difference over all of the results (p value = 0.4) and a weak significance (p value = 0.16) for a better performance of FlexS if the self-alignments are not included in the comparison. Correspondingly, the difference between GMA and fFlash (excluding self-alignments) has a p value of 0.2 for better performance of GMA. We attribute the difference between FlexS and the other two methods at least partly to the fact that the authors did not report RMSD values for the top-ranking solution, giving only the values for the best RMSD over all of the suggested solutions and the best RMSD obtained in the top 10 scoring solutions. The score evaluates the overlap with the template molecule, while the RMSD measures the deviation between a proposed conformation for the query molecule and its structure in the crystal and can be used to assess the quality of the alignment only when the crystal structure of the protein with the query molecule is known. To obtain a more detailed picture of the comparison between FlexS and GMA, we performed all comparisons used in the benchmark in FlexS and compared to the statistics of the number of alignments that had a RMSD below 1.5 Å: In ~280 comparisons between different molecules, FlexS found a good solution (<1.5 Å) in the first 10 ranks in 44% of cases, while the solution from GMA was below 1.5 Å in 30% of the cases. Since GMA returns only a single solution, this corresponds to rank 1. The difference obtained by including the 10 best ranks as opposed

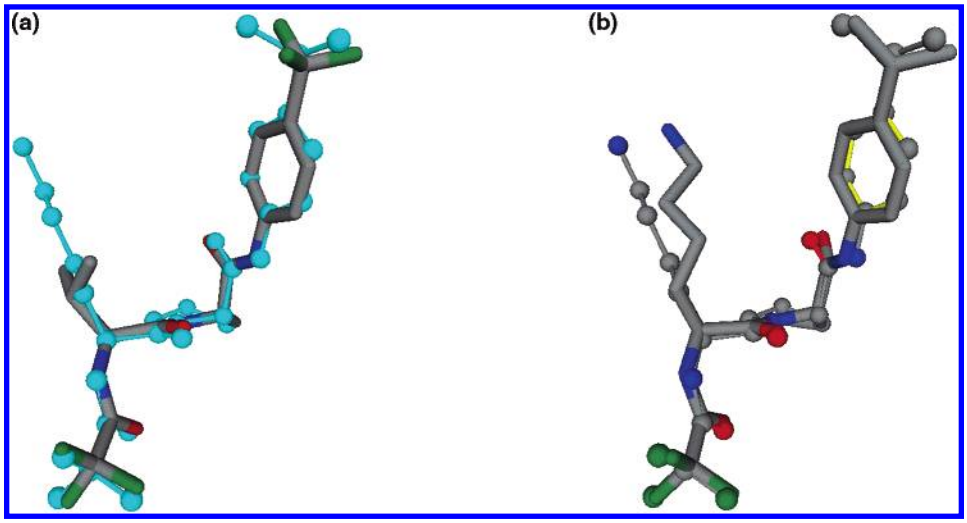

Figure 10. (a) 1ele ligand (light-blue ball-and-stick) aligned on 1ela ligand (thick sticks). (b) Comparison between aligned pose (ball-and-stick) and crystal structure for ligand 1ele (thick sticks), RMSD = 0.79 Å.

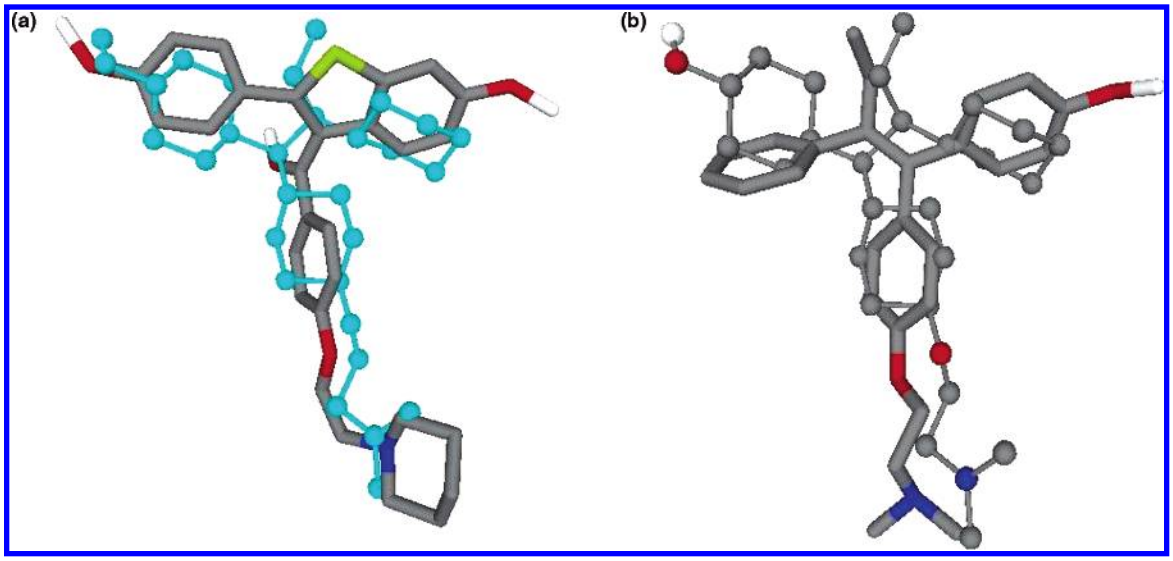

Figure 11. (a) 3ert ligand (light blue) aligned on 1err (thick sticks). (b) Comparison between aligned pose (ball and stick) and crystal structure for ligand 3ert (thick sticks), RMSD = 4.9 Å.

to using only the highest rank is seen in the results of fFlash where all three measures were reported (best rank, best RMSD out of top 10, and best RMSD overall). While GMA outperforms fFlash at the best rank level, it is only slightly better or comparable with the best RMSDs from the 10 best-ranked solutions. In many cases, simple sampling of the unaligned part of the ligands can lead to significant improvement in the obtained RMSD, while it has little to do with the quality of the alignments. Furthermore, for screening, the quality of the structural alignment at rank 1 is of interest, while lower ranks are in general not used. FlexS uses a sampling and clustering approach that should be even more effective in providing good solutions around the actual alignments, and we expect the effect to be significantly more pronounced than that observed in fFlash.

Finally, it is important to note that the good results obtained by GMA may indicate the relative simplicity of the benchmark set; however, the data sets in this benchmark are relatively typical for QSAR applications.

Comparison to Further Methods. A number of methods have been recently suggested for molecular superposition and their properties demonstrated on relatively limited test sets. Labute et al.[11] have introduced a field-based method and tested it on four alignment pairs. Two of those are the pair 4dfr/1dls and the pair 3cpa/1cbx. 1dls contains the same ligand as 1dhf and is therefore comparable to the corresponding entry in Table 1. The reported RMSD for the superposition of 4dfr on 1dhf is 1.4 Å, which is comparable to the FlexS result and better than the one obtained by GMA.

The comparison between 3cpa and 1cbx is a relatively simple case which leads to an RMSD of 1.3 Å in their mapping and 1.1 Å with GMA. The other two examples are superpositions of 1err (raloxifene) on 3ert (4-hydroxytamoxifene) and 1err on 1ere (estradiol). The result obtained by GMA in the first case shows very good overlap between the two ligands but is, nevertheless, wrong compared to the crystal structure superposition (RMSD > 4.5 Å).

The main reason is that a single torsion angle is rotated by approximately 180°, placing the two ring systems on top of each other, though in an inverted orientation (see Figure 11). In Labute et al.,[11] it is mentioned that this type of conformation is possible as it is also obtained in docking experiments of raloxifene in the estradiol receptor. Labute et al. obtain (after parameter fitting on all four examples) a

very good superposition also in this case. In the second case (1err on 1ere), the superposition is even more challenging, since estradiol is chemically significantly different from raloxifene. While a reasonable superposition is obtained with GMA, again it shows a reverse conformation with respect to the crystal structure. Again, the reason is the partial pseudosymmetry of 1err and the pseudosymmetry of ergosterol. In three of the four examples demonstrated in ref 11, their method outperforms GMA. It is, however, important to note that they have specifically trained the scoring function on this same selection, so as to minimize the RMSD to the crystal structure alignment. In all cases where the results of GMA and their method differ, the suggestion by GMA is a perfectly valid superposition and in some cases actually shows better overlap than the crystal structure, both in steric terms as well as in terms of functionality. Tuning parameters in order to obtain the crystal structure for a few selected cases appears to be a very specific approach that needs to be tested for its generality. However, the accuracy of the approach on independent test sets has not been reported.

Flame[12] has been demonstrated on two of those examples (3cpa/1cbx and 1err/3ert) and additionally on two thrombin ligands (1dwd and 1dwe). In both 1err/3ert and 1dwd/1dwe, specific parameter adjustment was necessary in order to obtain an overlay similar to the one in the crystal structure. In the first case (1err/3ert), the standard parameters (i.e., without tuning) lead to the same type of rotation of the two ring systems that we observe with GMA. In the case of 1dwd/1dwe, GMA leads to an RMSD of 2.45 Å with respect to the crystal structure alignment. The parameter tuning used in Flame to obtain better RMSDs is even more specific than that described by Labute et al.[11] as it is applied on every pair of alignments independently. This type of case-dependent parameter tuning used in Flame[12] is of rather limited applicability as it requires the corresponding crystal structures of the complexes with the two ligands. When it is used, it needs to be validated by showing improved screening performance or some comparable external criterion.

All in all, the comparison to those two approaches is made difficult by the smallness of the benchmarks used by the authors, as well as the fact that they appear somewhat arbitrarily chosen. For example in both studies, only one direction of the alignment is tested, even though it is clear that differences in flexibility and molecular size often lead to significantly different results in the two directions. Further, the overlap between the test sets of those two studies is only partial, although the authors of Flame are clearly aware of the previous work. This arbitrary choice of benchmark data makes an objective comparison of the relative weaknesses and strengths of new methods rather difficult, since a common reference frame is missing.

From the examples discussed by the authors of those two studies, it appears that GMA leads to results that are comparable to their naïve (not fitted) results. On the down side, we have to note that graph-based methods such as GMA appear less tunable, since initial tests to learn weights and other parameters of the graph alignment part failed to significantly improve the results. Thus, it appears that graph-based approaches are not necessarily the best starting point for tunable alignment methods. On the other hand, a technical advantage of the graph-based alignment is that the obtained results are based on globally optimal mappings, while pure 3D methods (have to) rely on optimization methods that can at best guarantee local optimality. The higher robustness (lower tunability) of GMA may partly be due to the reliance on these optimal mappings [since the tunability may at least partly be explained by kinetic effects, making a particular optimum more accessible (broader) for the optimization as opposed to improving its score] and certainly makes the analysis of the results simpler: changes in the smoothness of the score hypersurface express themselves in changes in the efficiency of finding the solution. Thus, when the obtained alignments change, it is due to a change in the position of the global optimum.

Time Comparison with Other Methods. GMA's performance was measured on an AthlonXP 3000+ CPU. So, for a comparison of the methods (fFlash vs GMA), all the times from Krämer et al.[10] were divided by 1.67 as the times from the fFlash data set were prepared on a P4 1.8 GHz machine. This approximately corrects for the CPU frequency difference. The remaining difference of 1 order of magnitude is significant. For small ligands, for example, fructose, fFlash's and GMA's time performances are similar, with a small time advantage for fFlash (Table 1). On data sets with a larger conformational space, fFlash efficiency rapidly decreases (e.g., thermolysin and carboxypeptidase-a).

The reported CPU time requirement for FlexS for a single run (which produces a number of candidate solutions and ranks them) is in the minute range; however, those times were measured almost 10 years ago, making it somewhat difficult to compare. However, even nowadays, the newest version of FlexS requires approximately 2 min per superposition on average.[19]

The method by Labute et al. has been reported by the authors to be of comparable efficiency to that of FlexS.

On the basis of the reported CPU times for a screen of a library of 90 000 molecules (9.5 h on 100 CPUs), we estimate an average of ∼30 s per alignment for Flame.[12]

On the basis of those times, we obtain an improvement in efficiency between 1^{10} and $3^{9,11,12}$ orders of magnitude compared to state of the art methods. An interesting side effect of the high efficiency of the method is that in our applications alignments do not need to be stored (e.g., for visualization); rather, they can be created online upon demand.

The quality of the obtained alignments appears comparable to previously reported results, and we are currently investigating the performance of the method in screening. Initial results suggest that the obtained enrichment factors are very good, while the advantage in using the 3D structure (as opposed to graph-based screening) is highly case-dependent. However, further tests are needed for conclusive evidence on this point.

CONCLUSION

We have evaluated a simple approach for flexible superposition between molecules, which is based on the identification of a mapping between the chemical graphs of the molecules. The mapping allows the use of the RMSD between the mapped molecules as an objective function for optimization. The optimization is performed in torsion space. The reliance of the method on the maximum common subgraph of the molecules being compared emphasizes

chemical structure similarity versus interaction similarity in the comparison. While this is not always desirable, it does have a number of advantages: it is consistent with the MCS-based 2D searches used for virtual screening in drug design[20] and provides reasonable structural superpositions for the obtained hits at minimal additional cost. These provide additional information on the similarity of the compounds and allow a visual representation which yields additional insight. For example, the problem of topologically inconsistent mappings only became obvious from the poor alignments obtained in such cases. Topologically inconsistent mappings are problematic as a basis for 2D similarity measures, since they lead to artificially high similarity values of compounds consisting of similar building blocks with a completely different connectivity.

Unlike other approaches that combine 2D and 3D features,[10,21] GMA has no problems handling macrocyclic structures. In Feature Trees,[21] all cycles have to be condensed to single nodes or broken while according to Krämer et al.[10] fFlash cannot currently handle macrocyclic molecules. Finally, as one reviewer suggested, another potential benefit of the approach is that it can easily accommodate distance constraints among atoms as compared to field-based approaches.

We have shown that graph-based alignment leads on average to results that are of comparable quality to those obtained with a number of commonly used methods. We have concentrated on unbiased alignment (which makes no use of the X-ray structure of the complexes) and in providing the user with a single alignment (as opposed to providing a number of possible solutions), because in general this corresponds to the typical application in virtual screening.

ACKNOWLEDGMENT

The authors acknowledge partial funding by the DFG project AP 101/1-1. The authors further wish to thank the anonymous reviewers for their insightful comments, which have contributed to the improvement of this manuscript.

REFERENCES AND NOTES

(1) Gund, P. Evolution of the Pharmacophore Concept in Pharmaceutical Research. In *Pharmacophore, Perception, Development, and Use in Drug Design (IuI Biotechnology series 2*); Guner, O. F., Ed.; International University Line: La Jolla, CA, 1999; pp 3−12.

(2) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.

(3) Stahl, M.; Mauser, H.; Tsui, M.; Taylor, N. R. A Robust Clustering Method for Chemical Structures. *J. Med. Chem.* **2005**, *48*, 4358−4366.

(4) Raymond, J. M.; Willett. P. Maximum Common Subgraph Isomorphism Algorithms for the Matching of Chemical Structures. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 521−533.

(5) Barker, E. J.; Buttar, D.; Cosgrove, D. A.; Gardiner, E. J.; Kitts, P.; Willett, P.; Gillet, V. J. Scaffold Hopping Using Clique Detection Applied to Reduced Graphs. *J. Chem Inf. Model.* **2006**, *46*, 503−511.

(6) Available from Daylight Chemical Information Systems Inc., 27401 Los Altos, Suite #360, Mission Viejo, CA 92691.

(7) Bystroff, C. An Alternative Derivation of the Equations of Motion in Torsion Space for a Branched Linear Chain. *Protein Eng.* **2001**, *14*, 825−828.

(8) Deo, A. S.; Walker, I. D. Overview of Damped Least-Squares Methods for Inverse Kinematics of Robot Manipulators. *J. Intell. Robotic Syst.* **1995**, *14*, 43−68.

(9) Lemmen, C.; Lengauer, T.; Klebe, C. FLEXS: A Method for Fast Flexible Ligand Superposition. *J. Med. Chem.* **1998**, *41*, 4502−20.

(10) Kramer, A.; Horn, H. W.; Rice, J. E. Fast 3D Molecular Superposition and Similarity Search in Databases of Flexible Molecules. *J. Comput.-Aided Mol. Des.* **2003**, *17* (1), 13−38.

(11) Labute, P.; Williams, C.; Feher, M.; Sourial, E.; Schmidt, J. M. Flexible Alignment of Small Molecules. *J. Med. Chem.* **2001**, *44*, 1483−90.

(12) Cho, S. J.; Sun, J. FLAME: A Program to Flexibly Align Molecules. *J. Chem. Inf. Model.* **2006**, *46*, 298−306.

(13) Raymond, J.; Gardiner, E.; Willett, P. RASCAL: Calculation of Graph Similarity Using Maximum Common Edge Subgraphs. *Comput. J.* **2002**, *45*, 631−644.

(14) Kearsley, S. K.; Smith, G. M. An Alternative Method for the Alignment of Molecular Structures: Maximizing Electrostatic and Steric Overlap. *Tetrahedron Comput. Methodol.* **1990**, *3*, 615−633.

(15) Kabsch, W. A Solution for the Best Rotation to Relate Two Sets of Vectors. *Acta Crystallogr.* **1976**, *32*, 922−923.

(16) Lemmen, C.; Claus Hiller, C.; Thomas Lengauer, T. RigFit: A New Approach to Superimposing Ligand Molecules. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 491−502.

(17) Clark, M.; Cramer, R. D., III; Van Opdenbosch, N. Validation of the General Purpose Tripos 5.2 Force Field. *J. Comput. Chem.* **1989**, *10*, 982−1012.

(18) Badger, J.; Krishnaswamy, S.; Kremer, M. J.; Oliveira, M. A.; Rossmann, M. G.; Heinz, B. A.; Rueckert, R. R.; Dutko, F. J.; McKinlay, M. A. Three-Dimensional Structures of Drug-Resistant Mutants of Human Rhinovirus 14. *J. Mol. Biol.* **1989**, *207*, 163−174.

(19) BioSolveIT GmbH − FlexS. http://www.biosolveit.de/FlexS/ (accessed Oct 19, 2006).

(20) Raymond, J. W.; Willett, P. Effectiveness of Graph-Based and Fingerprint-Based Similarity Measures for Virtual Screening of 2D Chemical Structure Databases. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 59−71.

(21) Rarey, M.; Dixon, J. S. Feature Trees: A New Molecular Similarity Measure Based on Tree Matching *J. Comput.-Aided Mol. Des.* **1998**, *12*, 471−90.

CI600387R