Journal of Computer-Aided Molecular Design, **17:** 13–38, 2003.
KLUWER/ESCOM

Fast 3D molecular superposition and similarity search in databases of flexible molecules

Andreas Krämer, Hans W. Horn & Julia E. Rice
IBM Almaden Research Center, 650 Harry Road, San Jose, CA 95120

MS received 10 October 2002; accepted in final form 28 February 2003

Key words: molecular similarity, molecular superposition, database search, flexible 3D superposition, virtual library screening

Summary

We present a new method (fFLASH) for the virtual screening of compound databases that is based on explicit three-dimensional molecular superpositions. fFLASH takes the torsional flexibility of the database molecules fully into account, and can deal with an arbitrary number of conformation-dependent molecular features. The method utilizes a fragmentation-reassembly approach which allows for an efficient sampling of the conformational space. A fast clique-based pattern matching algorithm generates alignments of pairs of adjacent molecular fragments on the rigid query molecule that are subsequently reassembled to complete database molecules. Using conventional molecular features (hydrogen bond donors and acceptors, charges, and hydrophobic groups) we show that fFLASH is able to rapidly produce accurate alignments of medium-sized drug-like molecules. Experiments with a test database containing a diverse set of 1780 drug-like molecules (including all conformers) have shown that average query processing times of the order of 0.1 seconds per molecule can be achieved on a PC.

Introduction

The virtual screening of compound databases is an important tool in modern drug design. Traditionally, two-dimensional or pharmacophore-based methods, which are very fast but have only limited accuracy, have been used for this purpose [1, 2, 3, 4]. In order to make database searches more accurate, the three-dimensional structure and conformational flexibility of the molecules have to be taken into account. In this paper we describe a highly efficient method (fFLASH) to perform a database search for molecules that are similar to a given conformation of a reference or query molecule, based on explicit three-dimensional, flexible superpositions.

3D molecular superposition methods have been successfully utilized to determine binding geometries relative to a reference molecule [5–14]. They play an important role in 3D-QSAR applications, pharmacophore elucidation, and receptor modelling, in situations where structural data of the target protein is not available. The variety of methodologies used for molecular superposition has recently been extensively reviewed by Lemmen et al. [15], and an application of existing superposition methods to virtual database screening has been reported in [16].

Of course, the use of molecular superposition to determine the binding capability of possible ligands has its limitations. The underlying assumption is that other ligands will have the same overall binding mode as the reference molecule. Also, the bound conformation of the reference molecule has to be known, which is generally only the case if crystallographic information about the corresponding protein-ligand complex is available. Therefore, in practical applications the reference molecule should have a non-flexible structure, or its bound conformation has to be inferred using other methods, e.g., deduced from simultaneous, flexible alignments within a set of ligands that are known to be active [17–22]. The bound conformation of the reference molecule can also be determined from distance constraints obtained in NMR (NOE) experiments, a possibility that has been outlined in [23].

fFLASH is based on earlier work on the FLASHFLOOD method [24], however, most parts of FLASHFLOOD's underlying algorithm have been replaced by new developments which have resulted in a significant performance improvement. We are now able to screen a database of about 1800 molecules within a few minutes on a low-end PC. While FLASHFLOOD uses a field-based feature definition that can be tuned by the user to a particular context under investigation, in this paper, molecules are represented in a more traditional way using point-like features that are known to be important in protein-ligand binding, e.g. hydrogen bond donors, acceptors, charges and hydrophobic regions. This feature definition constitutes a simple test scenario for fFLASH's similarity search algorithm. Nevertheless it should be pointed out that more sophisticated feature schemes such as those using certain molecular surface properties [20, 25]) could be used, since features are allowed to explicitly depend on the molecular conformation.

We will show that fFLASH can be used to search databases of non-trivial size efficiently, while producing accurate 3D molecular alignments. fFLASH relies on a fragmentation-reassembly procedure to significantly speed up conformational sampling. Also, in order to enable fast query processing during database searches, as much computational work as possible is done in a database preprocessing step where features are calculated from molecular properties. During preprocessing typically millions of conformations are explored by uniform sampling of dihedral angles, but many of them are discarded either because of internal steric clashes or because the generated conformational change is too small relative to a conformation which has already been represented in the database.

The paper is organized as follows: In the next section we describe the molecular similarity search algorithm in detail and explain the different steps involved. The third section presents a number of experiments aimed to test the capabilities of fFLASH with regard to computational performance and the accuracy of molecular alignments. In the first set of experiments (section 3.4) we perform mutual alignments of medium-sized molecules taken from the FlexS$^{\mathrm{TM}}$ benchmark test set [26]. The second set of experiments (section 3.5) explores the ability to find molecules which are known to be similar to a given query molecule in a large diverse set of 1780 molecules. For this purpose we prepared a database containing 52 known dihydrofolate reductase inhibitors (50 molecules from Crippen [27], methotrexate, and dihydrofolate) as well as 1728 compounds taken from the diversity set published on the web site of the National Cancer Institute (NCI) [28]. A general discussion follows (section 4) and the conclusions are presented in section 5.

Algorithm

The similarity definition used in fFLASH is based on the representation of a molecule by a set of *features* which are located at certain points in space. We distinguish various rotationally invariant *feature types* that classify different local physico-chemical properties of the molecule. These can be atom-based entities like hydrogen bond donors and acceptors or charges, but could also be local properties which are not necessarily related to atoms or functional groups. For instance, they could describe local geometry or quantities which have been derived from a continuous field.

We say that a molecule D is similar to the rigid query molecule Q if a conformation of D exists that can be aligned to Q such that a certain number of features on Q and D match with respect to feature type and location within a certain tolerance. The number of matching features (subsequently called *votes*) together with a Carbo score [29], which characterizes the overall molecular overlap, is used to construct a final *similarity score* that measures the similarity between Q and D. The goal is to find all molecules D in a database that can be aligned to a given rigid query molecule Q such that the similarity score is larger than a given value.

fFLASH's similarity search procedure is illustrated in Figures 1 and 2. It consists of a *database preprocessing* step which is only carried out once, and the actual database query. Here we outline the general scheme, and go into more detail in the subsequent subsections 2.1 to 2.4.

During preprocessing (see Figure 1) all database molecules are partitioned into disjoint fragments which are connected by rotatable bonds (see section 2.1) and which are each expanded into a set of conformations such that the conformational space is sampled up to a certain accuracy (measured by RMSD) (see section 2.2), and there are no steric clashes between nonbonded atoms. The coordinates of all fragment conformations are then explicitly stored in the database. Here, we focus on the special case of a 'linear-chain' fragmentation, where each fragment is connected to one or two other fragments. In the next step we join *pairs* of adjacent fragments whose

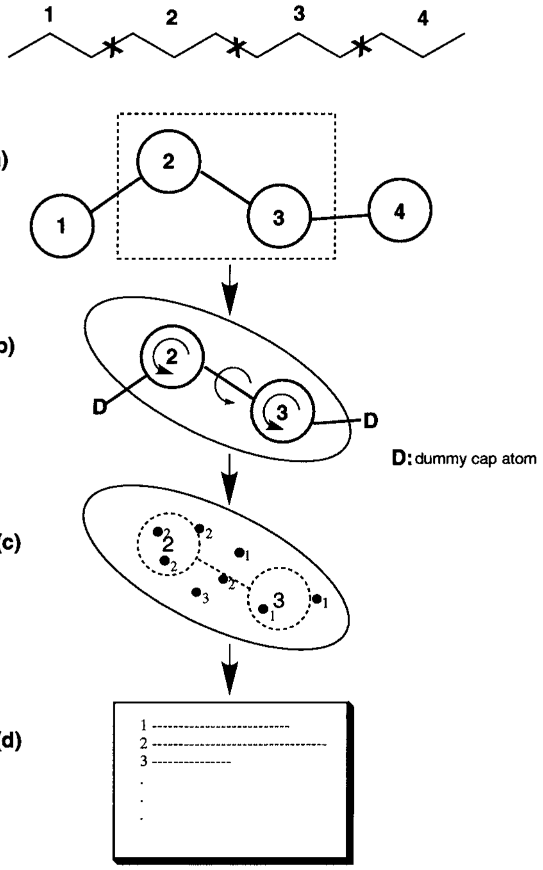


Figure 1. Database preprocessing step (schematically). (a) Database molecules are partitioned into fragments (1,2,3,4), and adjacent fragments are joined to make fragment pairs ([1,2],[2,3],[3,4]). (b) Sampling of fragment pair conformations. (c) Feature computation for each sampled fragment pair conformation (see text for explanation). (d) Features of each fragment pair conformation are stored in a lookup table, where the feature type serves as the lookup index. Different feature types will, in general, have a different number of entries.

conformations are expanded w.r.t. the dihedral angle at the connecting bond, however, we only store the rules (i.e. the rotations and translations applied to the two fragments) to make a fragment pair rather than storing fragment pair coordinates explicitly. Features are then calculated for each fragment pair conformation. The use of overlapping fragment pairs instead of fragments in the feature computation reduces the influence of boundary effects. Information about each feature, including a reference to the corresponding fragment pair and its location in space, is stored in fFLASH's central lookup table using the feature type as the lookup index (see section 3.2 and Table 2 for explicit examples).

When querying the database (see Figure 2), for each feature on the rigid query molecule, entries

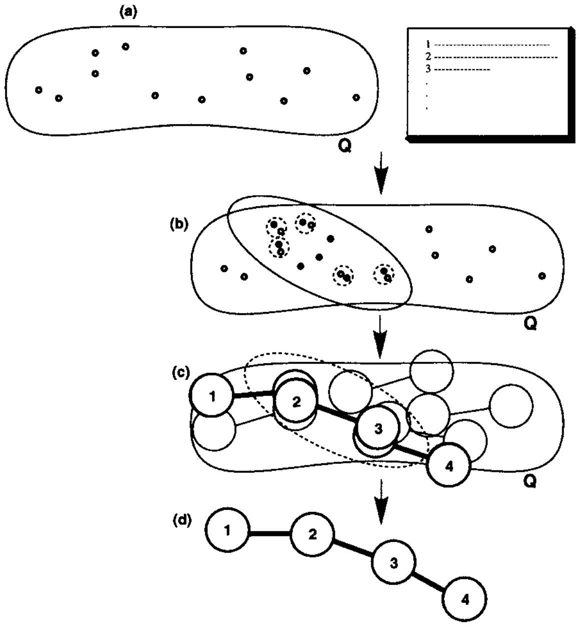


Figure 2. Database query (schematically). (a) Table lookup for each feature on the query molecule. (b) Alignment of matching fragment pair conformations (hypotheses) onto the query molecule *Q* after feature pattern matching. (c) Joining of hypotheses that are pairwise compatible. (d) Finally assembled database molecule (candidate).

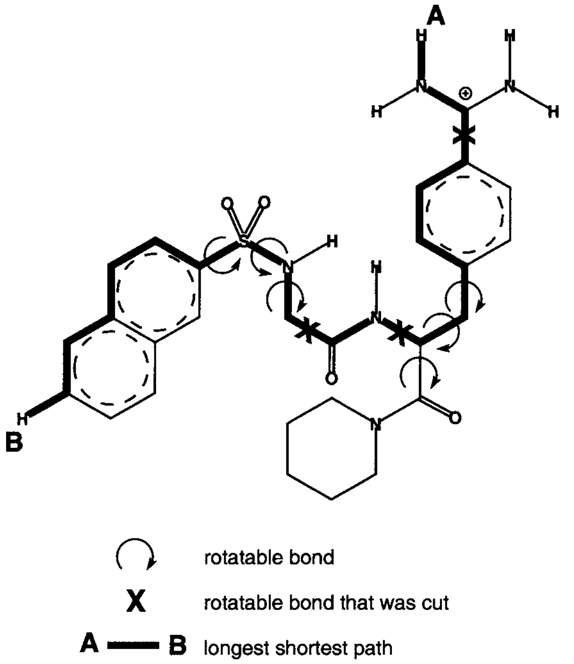


Figure 3. Illustration of the automatic molecular fragmentation procedure. The algorithm determines the path in the molecular graph which contains the largest number of rotatable bonds. The molecule can then, in principle, be cut at each rotatable bond along this path. The number and actual location of the cuts is chosen such that the number of fragment pair conformations is minimal given a certain fragment pair size.

are retrieved from the lookup table (see section 2.3). Using clique detection [30], patterns of features on fragment pairs are geometrically matched with similar feature patterns on the query molecule. This produces multiple alignments of fragment pairs on the query molecule which we will call *hypotheses*. Hypotheses are then assembled to whole molecules (subsequently called *candidates*) using a graph-based algorithm which is based on pairwise compatibility of hypotheses (see section 2.4). All candidates are scored as described above, and a final bump check is performed (see section 3.3).

The size of the molecular fragments is determined by a trade-off between the total number of conformations that have to be processed during a database query, and the selectivity of single pattern matches. Fragment pairs that are too large will generally exhibit too many internal conformations. On the other hand, fragment pairs that are too small will have only a few feature points and therefore produce many unspecific placements on the query molecule, so that a meaningful assembly of complete molecules is not possible. Here, molecules are typically divided into 1 to 3 fragment pairs which contain between 20 and 40 atoms, and contain around 10 feature points. In contrast to other approaches [8], fFLASH uses no special base fragment from which an incremental assembly procedure is started.

The current software prototype of fFLASH has a client/server architecture where the server (implemented in Java) and applications (implemented in C++) run on an IBM RS/6000 workstation and the graphical user interface (GUI) client (implemented in Java) runs on a personal computer (PC). Alternatively GUI, server and applications can be installed on a stand-alone PC or laptop.

In the following we will describe each step of the similarity search algorithm in more detail.

Fragmentation

Fragmentation of a molecule is an automated process. This process results in the identification of a set of rotatable bonds at which the molecule is cut. This set of rotatable bonds is determined such that the total number of fragment pair conformations is minimized given the minimum fragment pair size and the sampling resolution $\Delta\varphi$ of the dihedral angles. The algorithm used for automatic partitioning is illustrated in Figure 3. First we determine all bonds which are rotatable. Here, we distinguish between those bonds which are expected to be more or less freely rotatable (e.g. C–C single bonds between carbon atoms) and those which are expected to exist in two (cis/trans) conformations corresponding to $\Delta\varphi = 0°$ and $180°$ (e.g. C=C double bonds). The criteria used to determine whether a particular bond is considered to be rotatable are listed in Table 1. The algorithm determines a path P through the molecular graph which contains those rotatable bonds at which the molecule may be cut into fragments. This is done by assigning a weight of 1 to each edge of the molecular graph which represents a rotatable bond and a small weight ϵ for each non-rotatable one. We then use Dijkstra's algorithm [31] for finding shortest paths in a graph to determine the pair of atoms A and B where the shortest path between A and B has the maximal length for all atom pairs in the molecule. The parameter ϵ has to be larger than 0 in order to ensure stability of the algorithm, here, we chose $\epsilon = 0.1$. In the limit $\epsilon \to 0$ the length of the path P corresponds to the number of rotatable bonds along P. The procedure described above minimizes the number of rotatable bonds not contained in the path P, and therefore maximizes the number of 'cuttable' bonds. This works well for molecules that have few branches. Finally, an exhaustive search of all possible fragmentations involving cuts at the rotatable bonds along P is done, in order to find the optimal set of rotatable bonds to be cut, such that the total number of fragment pair conformations is minimized, given the minimum fragment pair size, and the sampling resolution $\Delta\varphi$ of the dihedral angles.

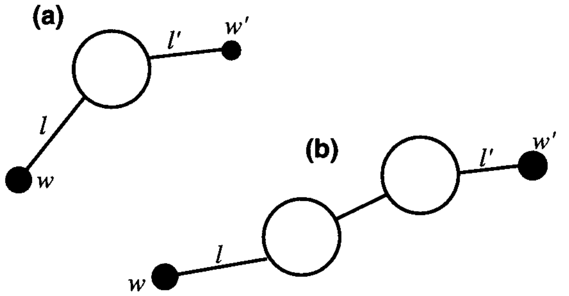


Figure 4. Fragment (a) and fragment pair (b) capped by dummy atoms with weights w, w' at distances ℓ, ℓ'.

Sampling of conformational space

Molecular conformations are generated by uniformly sampling the dihedral angles of all rotatable bonds with a certain angular resolution $\Delta\varphi$. In the preprocessing step this is done for all fragment pairs of

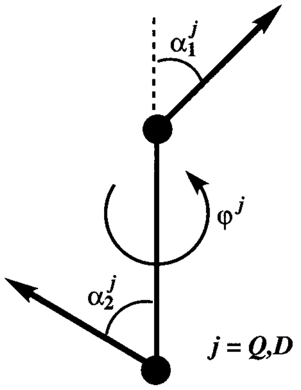


Figure 5. Definition of angles φ^Q, φ^D, α_i^Q, and α_i^D ($i = 1, 2$) in Eqs (4),(5), and (6). The points (•) denote two features located on the query molecule ($j = Q$) or a fragment pair conformation in the database ($j = D$). The arrows are the corresponding vectors $\vec{p}$ (see Eq. (1)).

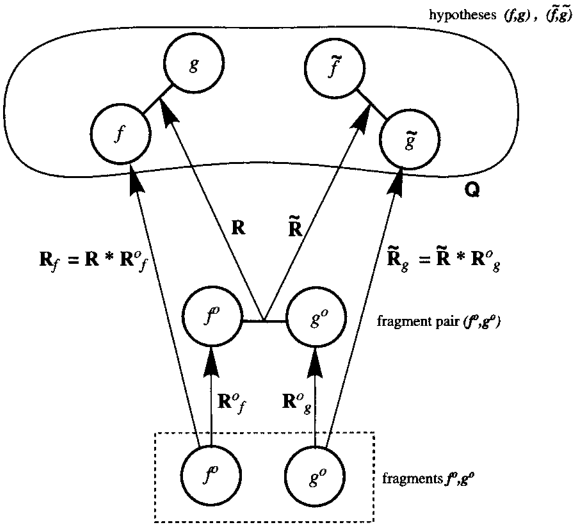

hypotheses (f,g), $(\tilde{f},\tilde{g})$
g
$\tilde{f}$
f
$\tilde{g}$
Q
R
$\tilde{\mathbf{R}}$
$\mathbf{R}_f = \mathbf{R} * \mathbf{R}^o_f$
$\tilde{\mathbf{R}}_g = \tilde{\mathbf{R}} * \mathbf{R}^o_g$
f^o
g^o
fragment pair (f^o,g^o)
$\mathbf{R}^o_f$
$\mathbf{R}^o_g$
f^o
g^o
fragments f^o,g^o

Figure 6. Mapping of fragment conformations (bottom) onto a fragment pair ($\mathbf{R}^0$, middle) and onto hypotheses ($\mathbf{R}$, $\tilde{\mathbf{R}}$, top) aligned to the query molecule (schematically).

the database molecules, and a bump check is performed that removes conformations where two atoms come too close to each other (see section 3.1). Since the number of conformations should be as small as possible, in order to minimize the query processing time, we perform a subsequent ‘coarse graining’ step in which we make sure that the mutual RMSD R (after minimization w.r.t. translations and rotations) of all pairs of conformations is larger than a certain threshold μ. This is achieved by scanning through all conformations generated by the angular expansion procedure, and removing all those whose RMSD w.r.t. any previously encountered conformation is smaller than μ. Since R represents a metric in the space of conformations this can be done efficiently by making use of the triangle inequality. The parameter μ should reflect the expected accuracy (in terms of the RMSD) of the final candidate alignments on the query molecule. We typically use a value of 1 Å here. In order to ensure that all fragment pairs are built from the same set of fragment conformations, the coarse graining procedure is performed first on the fragment level, and then again after sampling the dihedral angle of the connecting bond on a fragment pair.

A complication of the strategy described above arises from the fact that we actually sample conformations of fragments and fragment pairs instead of complete molecules. It may happen that fragment conformations which are very close in conformational space, and thus fall below the RMSD threshold μ, lead to conformations of the complete molecule which are very far apart. In order to deal with this problem, we have to approximate RMSDs of the complete molecule while sampling conformations on the fragment or fragment pair level. This is done by attaching a dummy atom to each connecting bond of the fragment or fragment pair at a certain distance ℓ (see Figure 4). These dummy atoms can be thought of as representing the average centroid of the remainder of the full molecules which has, – say – N atoms, and therefore carries a weight, $w = N$ in the RMSD computations. We choose $\ell = N/6$ Å, which is approximately the distance of the centroid of a stretched hydrocarbon chain of N atoms. The angular resolution, $\Delta\varphi$, determines how completely the conformational space is being sampled given the accuracy parameter, μ. If $\Delta\varphi$ has been chosen to be too large, we have to expect that certain conformations that are important for the alignment with the query molecule may be missed. Here, we typically choose $\Delta\varphi = 60°$.

Lookup Table and pattern matching

The query molecule as well as the fragment pair conformations in the database are represented by a set of features reflecting local physico-chemical properties of the molecules. These features are explicitly calculated and stored for each fragment pair conformation. A feature F can be written as a tuple

$$F = (I, K, \vec{x}, \vec{p}), \qquad (1)$$

where I is an integer representing the feature type, K is a reference to the fragment pair (not present for query features), $\vec{x}$ is the location of the feature in Cartesian space (in the fragment pair or query molecule coordinate system), and $\vec{p}$ is an optional unit vector that may represent any directional information of the feature, e.g. the direction of a hydrogen bond. All fragment pair features are stored in a lookup table as an (index, value list)-pair, where I serves as the lookup index, and $(K, \vec{x}, \vec{p})$ represents an entry in the value list.

When a database query is launched, lookup table entries are retrieved for each query molecule feature resulting in a list of *feature correspondences* $\{c_1, c_2, \ldots\}$. A feature correspondence c is the match of a single feature on the query molecule with a feature on a fragment pair, i.e. both the query feature and the fragment pair feature have the same feature type (or lookup index). c can be expressed as

$$c = (\vec{x}_q, \vec{x}_d, \vec{p}_q, \vec{p}_d, K),$$

where $\vec{x}_q$ and $\vec{x}_d$ are the feature locations on the query molecule and a fragment pair of a particular database molecule, respectively $\vec{p}_q$ and $\vec{p}_d$ are the corresponding unit vectors encoding directional information, and K is the fragment pair index. Different molecules in the database are processed separately, so we have to deal only with fragment pair conformations belonging to one molecule at the same time.

In order to detect common feature patterns on the query molecule and on a particular fragment pair conformation we have to find sets of feature correspondences $c_j = (\vec{x}_{qj}, \vec{x}_{dj}, \vec{p}_{qj}, \vec{p}_{dj}, K_j)$, $j = 1, \ldots, V$, where V is the number of votes, which can be matched simultaneously within a certain geometric tolerance by aligning the fragment pair onto the query molecule. This alignment can be expressed by a rotation matrix $\mathbf{R}$ and a translation vector $\vec{t}$ applied to the fragment pair. $\mathbf{R}$ and $\vec{t}$ are determined by the condition that $|\vec{x}_{qj} - \mathbf{R}\vec{x}_{dj} - \vec{t}|$ and $\angle(\vec{p}_{qj}, \mathbf{R}\vec{p}_{dj})$ are small given a certain tolerance for all $j = 1, \ldots, V$. The alignment of the fragment pair defined by the rotation $\mathbf{R}$ and translation $\vec{t}$ then represents a hypothesis with V votes.

The pattern matching problem described above can be treated within a clique detection approach by mapping the feature correspondences onto the vertices of a graph G, where two vertices are connected by an edge if the corresponding feature correspondences are pairwise compatible. Alignments with V votes then correspond to cliques with V vertices in the graph G. Two feature correspondences $c_i = (\vec{x}_{qi}, \vec{x}_{di}, \vec{p}_{qi}, \vec{p}_{di}, K_i)$ (in the following indexed by $i = 1, 2$) are defined to be compatible if the distances between the features are the same on the query molecule and on the fragment pair within a tolerance $\epsilon = \epsilon_1 + \epsilon_2$,

$$\|\vec{x}_{q1} - \vec{x}_{q2}| - |\vec{x}_{d1} - \vec{x}_{d2}\| < \epsilon, \tag{2}$$

and a pairwise alignment $\vec{d}_q \| \tilde{\mathbf{R}}\vec{d}_d$ (where $\tilde{\mathbf{R}}$ is a rotation matrix) of the difference vectors $\vec{d}_q = \vec{x}_{q1} - \vec{x}_{q2}$ and $\vec{d}_d = \vec{x}_{d1} - \vec{x}_{d2}$ exists, such that

$$\angle(\vec{p_{qi}}, \tilde{\mathbf{R}}\vec{p}_{di}) < \gamma_i \tag{3}$$

where γ_i denotes an angular tolerance. The parameters $\epsilon_1, \epsilon_2, \gamma_1$, and γ_2 are independently assigned to the two feature correspondences c_1 and c_2. It can be shown that Eq. (3) is equivalent to the condition

$$\Delta\varphi_1 + \Delta\varphi_2 \geq \left|\varphi^Q - \varphi^D\right| \tag{4}$$

where

$$\Delta\varphi_i = 2 \arcsin \sqrt{\lambda_i} \tag{5}$$

and

$$\lambda_i = \frac{\cos\left[\alpha_i^Q - \alpha_i^D\right] - \cos\gamma_i}{2 \sin\alpha_i^Q \sin\alpha_i^D} \tag{6}$$

with φ^D, φ^D, α_i^Q, and α_i^D defined as shown in Figure 5. If $\lambda_1 < 0 \vee \lambda_2 < 0$ no pairwise alignment is possible. Otherwise, if $\lambda_1 \geq 1 \vee \lambda_2 \geq 1$ a pairwise alignment is always possible.

fFLASH uses an incremental approach, similar to that given in [32], to compute the list of all cliques in the graph G ordered by the number of votes. The actual fragment pair alignments or hypotheses – represented by the rotation $\mathbf{R}$ and translation $\vec{t}$– are then determined by an RMSD minimization procedure applied to the feature locations. While pairwise distance compatibility of feature correspondences ensures the distance compatibility of the whole feature pattern (except in situations with mirror symmetry), this is not the case for the pairwise angular compatibility of the attached vectors $\vec{p}_{qj}$ and $\vec{p}_{dj}$. Therefore, after clique-detection a post-processing step is applied that removes all feature correspondences from the clique where $\angle(\vec{p}_{qj}, \mathbf{R}\vec{p}_{dj}) \geq \gamma_j$, and reduces the number of votes V accordingly. Note, that at least 3 votes are needed for a unique alignment. For this reason all cliques with $V < 3$ are discarded.

In order to limit the workload in the subsequent processing steps not all hypotheses are passed to the assembly procedure. For each fragment pair conformation, hypotheses that pass an additional shape screen (see section 3.1) are ordered w.r.t. votes, and put into 'buckets', B_k, containing all hypotheses with the same number of votes, V_k. All n buckets B_k with the highest number of votes ($V_1 = V_{\max}$, $V_2 = V_{\max} - 1$, …, $V_n = V_{\max} - n + 1$, where $V_{\max}$ is the maximal number of votes, and

$$\sum_{k=1}^{n-1} m_k + \frac{m_n}{2} \leq p, \tag{7}$$

are handed over to the assembly, where m_k is the number of hypotheses in B_k. This ensures that the total number of hypotheses per fragment pair conformation is always smaller than $2p$. Unless stated otherwise, in the experiments described below, we set $p = 5$.

Assembly algorithm

After lookup and pattern matching each fragment pair conformation results in zero, one, or several hypotheses, i.e., fragment pairs that are aligned to the query molecule. A hypothesis can be written as an ordered pair (*f*, *g*) of fragments where $f = f\{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n\}$ and $g = \{\vec{y}_1, \vec{y}_2, \ldots, \vec{y}_m\}$ are atomic coordinates of the two fragment conformations in the fragment pair aligned to the query molecule. *f* and *g* can be expressed in terms of the rotation matrices $\mathbf{R}_f$ and $\mathbf{R}_g$ which define the rotation of the aligned fragments relative to the fragment coordinates f^0 and g^0 stored in the database,

$$f = \mathbf{R}_f\left(f^0 - \vec{x}^0\right) + \vec{x}, \quad g = \mathbf{R}_g\left(g^0 - \vec{y}^0\right) + \vec{y},$$

where $\vec{x}^0$ ($\vec{y}^0$) is the center of geometry of the stored coordinates f^0 (g^0), and $\vec{x}$ ($\vec{y}$) is the center of geometry of the fragment f (g) aligned onto the query molecule. $\mathbf{R}_f$ and $\mathbf{R}_g$ can be calculated from the rotation $\mathbf{R}$ obtained in the pattern matching step and the known rotations and translations $\mathbf{R}^0$ which map the fragment coordinates stored in the database onto fragment pair coordinates (see Figure 6). The goal is now to find all *candidates*, i.e. superpositions of a complete database molecule with the query molecule, which are consistent with the hypotheses. We are thus looking for sets of hypotheses $H_1 = (f_1, f_2)$, $H_2 = (f_2', f_3)$, …, $H_{N-1} = (f_{N-1}', f_N)$, where the coordinates of overlapping fragments are sufficiently close w.r.t. a certain metric d which we will define below. Here, N is the number of fragments present in the molecule, and f_i and f_i' stand for the same fragment conformation, but with different coordinates according to the alignments of the hypotheses H_{i-1} and H_i.

For the assembly procedure we use the following metric d in the space of the coordinate representations of a fragment conformation f:

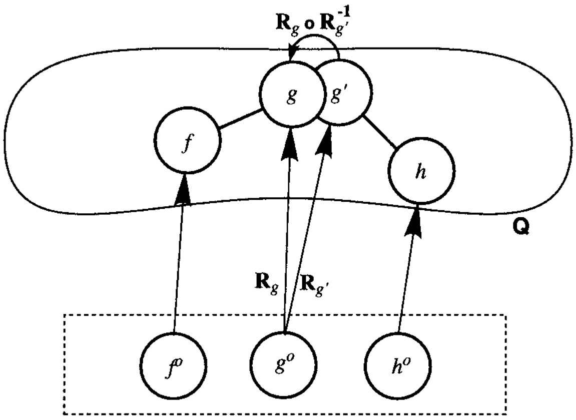


Figure 7. Merging of pairwise compatible hypotheses aligned to the query molecule.

$$d(f, f') = \frac{1}{2}|y_f - y_f'| + \frac{\alpha}{2}\mathrm{acos}\left[\frac{1}{2}\mathrm{Tr}\left(\mathbf{R}_{f'}\mathbf{R}_f^T\right) - 1\right]$$

d consists of a translational part and a rotational part. The translational part is simply the distance of the centers of geometry of f and f' while the rotational part represents the angle of rotation between f and f'. The parameter α determines the relative weight of both contributions and has been chosen to be $\alpha = 3$ Å [24]. Note, that we do not need explicit coordinates in order to evaluate d.

The candidate assembly can be expressed in terms of a graph representation, where hypotheses form the vertices of a *directed* graph G and two hypotheses $H = (f, g)$ and $I = (g', h)$ are connected by a directed edge ($H \rightarrow I$) if the distance (in terms of the metric d) of the overlapping fragment g, g' is smaller than a cutoff parameter δ,

$$d(g, g') < \delta \tag{8}$$

i.e., H and I are pairwise compatible (see Figure 7). We typically choose $\delta = 2$ Å. A candidate then corresponds to a directed path of length N in the graph G. In the assembly procedure we enumerate all those paths in G, which can be done in a straightforward way using recursion. Since we know the number of votes every fragment received as a result of ‘matching’, we can straightforwardly calculate the total number of votes for candidates by adding up all votes along the path. Votes of overlapping fragments are weighted by a factor of 1/2 to avoid counting them twice.

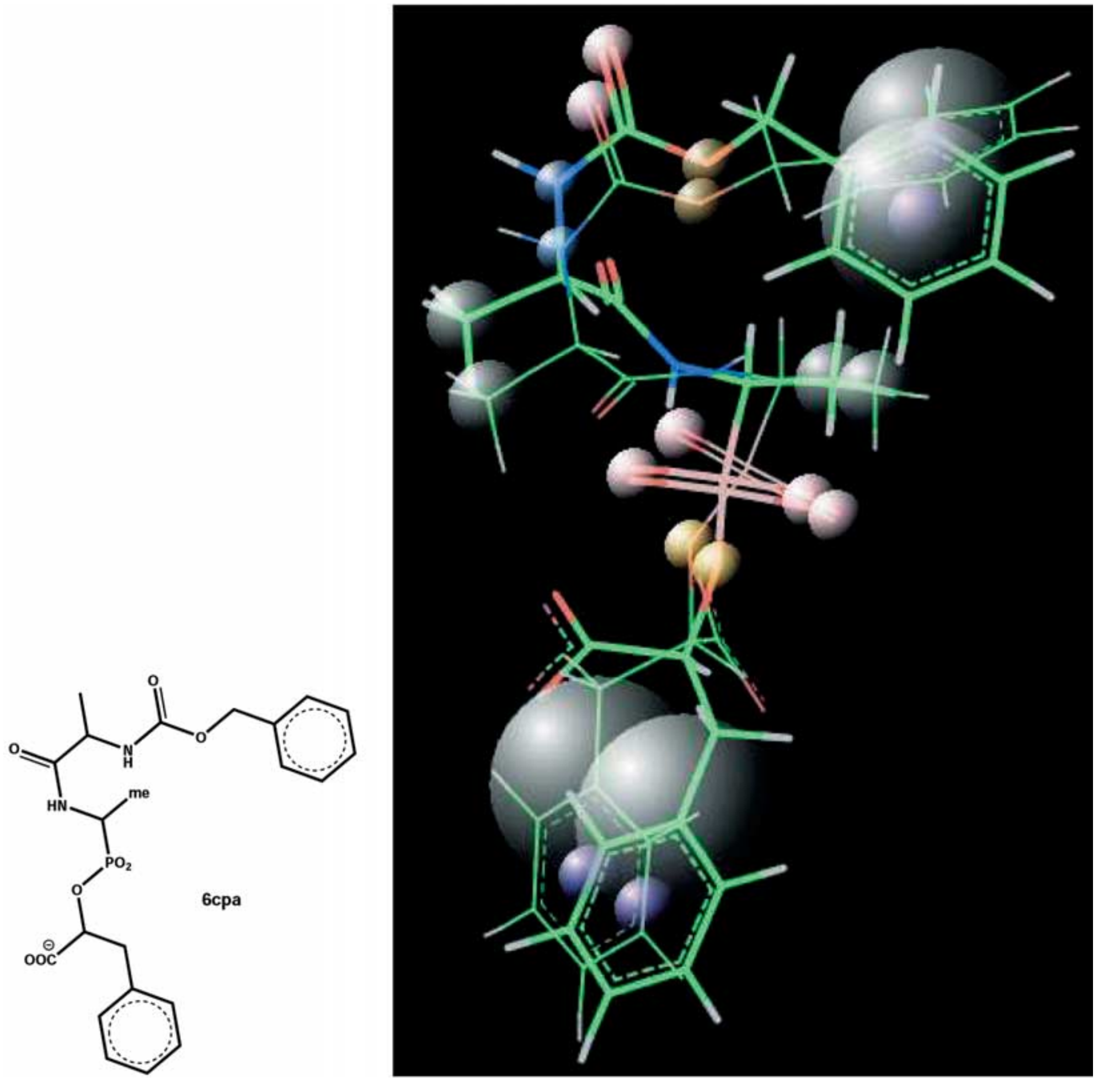


Figure 8. Result of the self query for the carboxypeptidase A inhibitor 6cpa. In this and the subsequent figures the colored spheres represent the following features (cp. Table 2): hydrogen bond donor (light blue), single tetrahedral hydrogen bond acceptor (red), double tetrahedral hydrogen bond acceptor (yellow), double planar hydrogen bond acceptor (pink), single planar hydrogen bond acceptor (gold), single linear hydrogen bond acceptor (blue), acid group (orange), base group (dark gray), hydrophobic aromatic (purple), phosphate (dark pink), OH or SH donor (green), OH or SH acceptor (light red), carboxyl (dark green), and hydrophobic region (medium gray).

The result of this graph-based assembly procedure is a list of candidates which is ordered by the number of votes, and where each candidate is represented by a path in G. In order to obtain the explicit coordinate representation of a candidate, fragments have to be explicitly rotated and translated. This is done by starting from the first fragment f_1 and then adding the other fragments successively. All rotations and translations needed in this process can be calculated straightforwardly from the rotation matrices and centers of geometry which were used above to calculate the distance d of adjacent hypotheses. After the assembly of a candidate in Cartesian space we perform a rigid alignment to the query molecule using RMSD minimization w.r.t. to the matching feature points. Of course, the explicit calculation of candidate coordinates is an expensive procedure. We therefore do an explicit assembly only for the highest scoring candidates that do not fail a bump check (see section 3.1) and pass an additional shape screen (see section 3.1). For this set of candidates, which represent the final

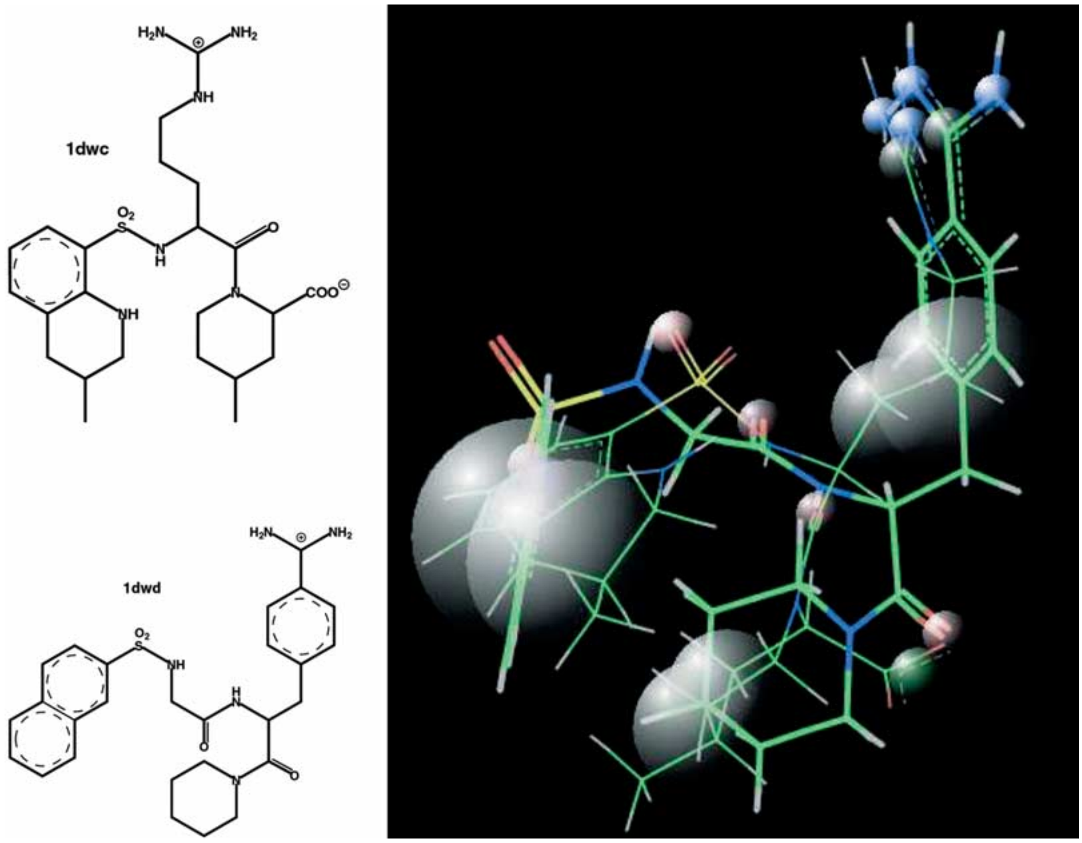


Figure 9. Molecular superposition of the thrombin inhibitors 1dwd and 1dwc, where 1dwc was stored in the database. Votes and Carbo score for this alignment were 0.68 and 0.80, the RMSD w.r.t. the crystal structure was 2.42 Å, and query processing took 2.2 s CPU time. The colored spheres represent different types of features (for an explanation of the colors see caption of Figure 8).

result of the query, we then also calculate the Carbo score [29].

Experiments

We have carried out two types of experiments in order to assess the performance of the algorithm (in terms of CPU time), the quality of molecular alignments, the influence of the sampling resolution of the conformational space, and the sensitivity and selectivity of the molecular similarity search method. The first set of experiments are mutual alignments of pairs of molecules which bind to the same receptor, and where the conformations of the bound states are known. In the second set of experiments we perform queries against a database which contains both a diverse set of 1728 compounds and also a smaller set of 52 molecules which are active towards the same receptor as the query molecule. Throughout these experiments we use a feature definition which is based on SYBYL® atom types [33].

Conformational expansion

As explained in section 2.1 we expand all rotatable bonds of database molecules with an angular resolution $\Delta\varphi$ in order to sample molecular conformations. In this first test of the algorithm we do not expand ring conformations. Table 1 shows the rules used to determine whether a given bond is rotatable or not, or may exist in two different states (‘cis/trans’). Bonds which connect to a terminal group which is approximately rotationally symmetric (CH_3, CX_3 (X=F, Cl), SO_3^-, PO_3^-) are considered to be non-rotatable, here. Special cases include hydroxyl, SH, and carboxyl groups which are treated as being rigid, i.e. the dihedral angle is not expanded explicitly, but features on these groups are defined in a way that their inherent flexibility is taken into account. During the conformational expansion we perform bump checks between all non-bonded atoms. Conformations where the distance between any two non-bonded atoms is smaller than 0.65 times the sum of the respective van der Waals radii (as taken from [34]) are discarded.

Feature definition

Table 2 shows the definition of the different atom type-based feature types and feature indices used in our experiments. In the current implementation we do not consider metal ions. Note that there may be more than one feature at a given point. The feature indices are 1 (hydrogen bond donor), 2 (hydrogen bond acceptor), 3 (base group, positive charge), 4 (acid group, negative charge), 5 (hydrophobic), and 6 (hydrophobic-aromatic). All features are located on non-hydrogen atoms except for benzene rings where the ring center is used, and carboxyl groups where features are placed in the middle of the distance between the oxygen atoms and hydrophobic groups. ‘Hydrophobic atoms’ as defined in Table 2 are combined into hydrophobic groups represented by single features. These hydrophobic groups are obtained as connected components of a distance-compatibility graph, where the vertices correspond to the hydrophobic atoms and an edge is created if the distance between two of those atoms is smaller than a certain cut-off distance, here taken to be 3 Å. The feature location is given by the center of geometry of the atoms within a group. The information about the group size, i.e. the number of atoms contained within, is passed along through the lookup table. Only feature matches of two hydrophobic groups where the difference in the number of atoms is smaller than 3 are considered for pattern matching. Note, that the above definition of hydrophobic groups is inherently conformation-dependent and cannot be derived from molecular topology information alone.

The vector $\vec{p}$ attached to a feature (see Eq. (1)) represents either the approximate direction of a fictitious hydrogen bond or the ring normal in the case of a benzene ring. In the latter case the sign of the vector is not used during processing of the query.

As in the conformational expansion step OH, SH, carboxyl (COO$^-$), amine oxide, SO_3^- , and PO_3^- groups are treated in a special way. Since we do not expand the dihedral angle at an OH or SH group, both an acceptor and a donor feature are placed on the O or S atom, with the vector $\vec{p}$ pointing in the direction of the bond away from the non-hydrogen atom connected to O or S. We account for the inherent flexibility of a hydroxyl or SH group by choosing appropriate angular tolerances (see Eq. (3)) in the pattern matching procedure (see below). Carboxyl groups are treated in a similar way. They are always assumed to be deprotonated (even if a hydrogen atom is explicitly present in the input file) and therefore carry an acid feature and a hydrogen bond acceptor feature. Here also, flexibility is accounted for by the choice of the angular tolerance of the attached vector (see below). PO_3^- and SO_3^- groups only carry a charge (acid) feature on the phosphorus or sulfur atom and no features on the oxygen atoms.

Parameters used in the following experiments are $\epsilon_i = 0.4$ Å for the distance tolerance (Eq. (2)), except for hydrophobic groups, where $\epsilon_i = 0.7$ Å if the matching groups both contain only one atom, and $\epsilon_i = 1.5$ Å otherwise. The angular tolerance (Eq. (3)) is $\gamma_i = 50°$, except for the special cases of carboxyl and hydroxyl-like groups, where γ_i was chosen to be $\gamma_i = 90°$ to reflect the internal flexibility. The cut-off parameter for the assembly of adjacent hypotheses (Eq. (8)) was set to $\delta = 2$ Å. The shape screen, imposed before and after assembly, removes all hypotheses and candidates that contain an atom which is farther than 4 Å away from any atom of the query molecule.

Scoring

For the experiments described below we used a generalized scoring scheme – rather than just counting votes – to allow for different weights of feature correspondences according to the feature types involved. This was done in order to better balance the importance of the various functional groups in the calculation of the score. For instance, a match between two carbonyl groups (>C=O) with two hydrogen bond acceptors should not have the same weight as two independent hydrogen bond acceptors. Similarly, a match between two carbonyl groups should be weighted higher than a match between a carbonyl group and some single hydrogen bond acceptor. All weights were chosen to be equal to 1 except for double tetrahedral and double planar acceptors, and hydroxyl-like donors and acceptors, where the weight was set equal to 0.86. We then define the ‘votes score’ as

$$\text{votes score} = \frac{\sum\limits_{\text{feature correspondences}} w_d w_q}{(\sum w_q^2)^{1/2} (\sum w_d^2)^{1/2}}. \qquad (9)$$

Here, w_q and w_d are the weights of features on the query molecule Q and the candidate D under consideration, and the sums in the denominator run over

Table 1. Definition of rotatable and non-rotatable bonds using SYBYL$^{®}$ atomtypes.

non-rotatable bonds	contained in ring structure,
	leads to endpoint in the molecular graph (e.g. −H, or =O),
	connects to methyl group CH_3,
	connects to CX_3 (X=F, Cl) group,
	connects to NH_2, where N is planar (N.pl3);
	connects to C.2 in carboxyl group
	amide bonds between C.2 and N.am
	connects to O or S in −OH or −SH
	connects to XO_3^- group (X=S, P; S.3 or P.3 with 3 O.cos2 attached)
'cis/trans' bonds (0° and 180° rotations allowed)	bonds between C.2-C.2, C.ar-C.2, C.ar-N.x, C.ar-C.cat, C.2-C.cat, N.pl3-C.cat
rotatable bonds	all other bonds

all features on Q and D. Note, that the votes score as defined above is a number between zero and one. The introduction of feature type-dependent weights requires the additional information about a specific feature type to be passed along in the lookup table. This also allows for a feature type-dependent definition of distance and angular tolerances used in the pattern matching algorithm (see Eqs. (2) and (3)).

Molecular superposition

Mutual superpositions have been performed for a set of 19 molecules each known to bind to one of 7 different receptors where the crystal structure of the protein-ligand complex is available, and from which the 'correct' molecular alignment can be inferred [9]. These molecules (listed in Table 3) are all medium-sized compounds taken from the FlexS^{TM} benchmark set [26]. With these experiments we want to demonstrate the ability of fFLASH to handle molecular flexibility in an efficient manner. Therefore, small, rigid molecules were not considered, because they represent trivial cases for the flexible pattern matching procedure.

For all molecules in our test set that fall into the same receptor class we first performed mutual 'rigid' superpositions, where the fragmentation-reassembly procedure of fFLASH was bypassed, and only the crystallographically known structure of one ligand was put into the database. fFLASH was able to produce the correct alignments (as specified in the FlexS benchmark set [26]) with an accuracy of below 1.5 Å RMSD (w.r.t. non-hydrogen atoms) for all pairs of molecules, except for some cases of H. rhinovirus coating protein ligands, where a reverse orientation of the molecules was found (see example 2 below). In most cases the alignment was better than 0.7 Å RMSD. This demonstrates the suitability of the atom type-based feature definition for the set of molecules considered here. The vote scores for these 'rigid' superpositions are listed in the next-to-last column of Table 3, and the last column shows the ratio between the 'flexible' vote scores (see below) and the 'rigid' vote scores. For self-queries we expect values less or equal to 1; deviations indicate less-than-perfect conformational sampling. Values greater than 1 for non-self-queries indicate the importance of taking molecular flexibility into account.

In the next step, we prepared databases each containing fragment pair conformations of *one* molecule of our test set. During the conformational sampling procedure we made sure that the molecular conformation that corresponds to the crystal structure was *not* contained in the database by adding random offsets in the dihedral angle expansion. This ensures that *self-queries*, where query and database molecules are the same, are non-trivial, and can be used as a test of how well the conformational space is being sampled. All dihedral angles have been sampled with a resolution of $\Delta\varphi = 60^\circ$, except for the case of 7cpa (a carboxypeptidase A inhibitor) where a correct self-alignment (<2.4 Å RMSD) could only be achieved with $\Delta\varphi = 30^\circ$. The molecular partitioning scheme (see section 2.1) was optimized for each molecule in the sense that the number of fragment pair conformations represented in the database was made as small

Table 2. Feature definitions based on SYBYL® atom types [33]. In some cases the number of bonds is given in angled brackets after the SYBYL® atom type. See Eq. (1) for the definition of the vectors $\vec{x}$ and $\vec{p}$.

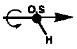
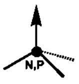
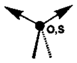
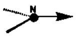
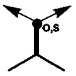
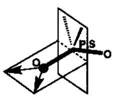
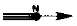
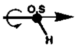
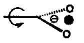
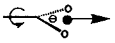
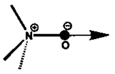
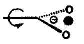
feature type	feature location $\vec{x}$	vector geometry $\vec{p}$	atom type	lookup index
hydrogen bond donor:				1
OH or SH	atom		O.3, S.3 (attached to one H)	
other	atom	atom→H	N.3, N.2, N.ar<3>, N.am, N.pl3, N.4 with H attached	
hydrogen bond acceptor:				2
single tetrahedral	atom		N.3, P.3	
double tetrahedral	atom		O.3, S.3 not attached to H	
single planar	atom		N.2, N.ar<2>	
double planar	atom		O.2, S.2	
sulphone or phosphone (double)	atom		O.2	
single linear	atom		N.1<1>	
-OH or -SH	atom		O.3, S.3 (attached to one H)	
carboxyl			C.2 (attached to two O.co2)	
amine oxide	atom		O.3 (attached to N.4)	
base	atom	-	C.cat, N.ar<3>, N.4	3
acid:				4
carboxyl		-	C.2 (attached to two O.co2)	
PO_3^-	atom	-	P.3 (attached to >2 O.co2)	
SO_3^-	atom	-	S.3 (attached to >2 O.co2)	
amine oxide	atom	-	O.3 (attached to N.4)	
hydrophobic:				
generic hydrophobic (see text)	atom	-	C.3, C.2, C.1 (not neighbor of O,N,S,P)	5
	ring center	-	6-ring with all C.ar	
hydrophobic-aromatic	ring center	ring normal	6-ring with all C.ar	6

Table 3. Results of molecular superposition experiments (see text). 'R' refers to a reverse alignment of query and database molecules. 'X' means that no candidate was found. The CPU times shown were obtained on a 1.8 GHz PC with Pentium 4 processor.

Query molecule	Database molecule	Fragment pairs	Conf's expanded	Conf's processed	Vote score	Carbo score	RMSD (min RMSD) (Å)	CPU time (s)	Vote score (rigid)	Vote (flexible)/ vote (rigid)
Thrombin inhibitors										
1dwc	1 dwc	2	$3.4 \cdot 10^6$	246	0.88	0.92	1.32 (1.28)	2.7	1.0	0.88
1dwd					0.68	0.80	2.42 (2.28)	2.2	0.41	1.66
1dwc	1 dwd	3	$3.4 \cdot 10^6$	86	0.59	0.83	1.93 (1.93)	0.5	0.41	1.44
1dwd					0.91	0.83	2.26 (1.86)	0.6	1.0	0.91
H. rhinovirus coating protein										
2r04	2r04	2	$2.0 \cdot 10^7$	333	1.0	0.95	1.37 (1.37)	1.5	1.0	1.0
2r06					1.0	0.94	2.56 (1.49)	1.5	0.66	1.52
2rr1					0.95	0.90	R	1.1	0.70 R	1.36
2rs3					0.95	0.88	R	1.3	0.70 R	1.36
2r04	2r06	2	$5.5 \cdot 10^5$	179	1.0	0.91	1.90 (1.80)	0.3	0.66	1.52
2r06					1.0	0.91	1.96 (1.28)	0.4	1.0	1.0
2rr1					0.95	0.89	R	0.3	0.63 R	1.51
2rs3					0.95	0.88	R	0.3	0.63 R	1.51
2r04	2rr1	2	$2.0 \cdot 10^7$	334	0.95	0.94	R	1.6	0.70	1.36
2r06					0.95	0.91	R	2.1	0.63	1.51
2rr1					1.0	0.92	1.64 (1.21)	1.5	1.0	1.0
2rs3					1.0	0.91	1.29 (1.29)	1.3	1.0	1.0
2r04	2rs3	2	$1.2 \cdot 10^8$	749	0.95	0.90	R	2.2	0.70 R	1.36
2r06					0.95	0.93	R	2.3	0.63 R	1.51
2rr1					1.0	0.93	2.16 (1.66)	1.8	1.0	1.0
2rs3					1.0	0.93	1.92 (1.64)	1.7	1.0	1.0
Fructose biphosphatase ligands										
4fbp	4fbp	2	216	9	0.89	0.99	0.52 (0.52)	0.2	1.0	0.89
t0039					0.63	0.96	0.57 (0.57)	0.2	0.73	0.86
4fbp	t0039	2	432	6	0.74	0.95	0.61 (0.61)	0.1	0.73	1.01
t0039					0.90	0.98	0.61 (0.61)	0.1	1.0	0.90
Dihydrofolate reductase inhibitors										
1dhf	1dhf	2	31104	27	0.89	0.89	1.62 (1.54)	0.1	1.0	0.89
4dfr					0.70	0.87	1.73 (1.73)	0.1	0.68	1.03
1dhf	4dfr	2	31104	32	0.70	0.91	2.69 (2.30)	0.2	0.68	1.03
4dfr					0.86	0.86	1.84 (1.84)	0.2	1.0	0.86
Thermolysin inhibitors										
1tlp	1tlp	2	$6.0 \cdot 10^7$	722	0.76	0.94	1.76 (1.48)	92.0	1.0	0.76
1tmn					0.54	0.86	3.11 (2.94)	9.4	0.57	0.95
2tmn					×	×	×	1.1	0.37	×
3tmn					0.51	0.68	3.86 (3.86)	5.3	0.63	0.81
1tlp	1tmn	2	$3.6 \cdot 10^8$	1305	0.52	0.70	3.54 (3.43)	11.8	0.57	0.91
1tmn					0.92	0.89	2.20 (1.85)	20.7	1.0	0.92
2tmn					×	×	×	0.8	0.42	×
3tmn					0.72	0.75	4.10 (3.70)	3.8	0.85	0.85
1tlp	2tmn	1	1296	25	0.37	0.65	1.88 (1.10)	0.1	0.37	1.0
1tmn					0.42	0.69	2.00 (1.24)	0.1	0.42	1.0
2tmn					1.0	0.96	1.42 (0.97)	0.1	1.0	1.0

Table 3. Continued.

Query molecule	Database molecule	Fragment pairs	Conf's expanded	Conf's processed	Vote score	Carbo score	RMSD (min RMSD) (Å)	CPU time (s)	Vote score (rigid)	Vote (flexible)/ vote (rigid)
3tmn					0.46	0.72	1.34 (1.22)	0.1	0.46	1.0
1tlp	3tmn	2	46656	20	0.58	0.76	1.36 (1.25)	0.1	0.63	0.92
1tmn					0.85	0.82	1.24 (1.24)	0.1	0.85	1.0
2tmn					×	×	×	0.1	0.46	×
3tmn					1.0	0.90	1.25 (1.00)	0.1	1.0	1.0
Carboxypeptidase A inhibitors										
1cbx	1cbx	1	216	12	1.0	0.96	0.88 (0.88)	0.1	1.0	1.0
6cpa					0.46	0.61	1.03 (1.03)	0.1	0.46	1.0
7cpa					0.44	0.53	1.13 (1.13)	0.1	0.44	1.0
1cbx	6cpa	2	$3.6 \cdot 10^{8}$	384	×	×	×	0.3	0.46	×
6cpa					0.79	0.88	1.61 (1.37)	11.6	1.0	0.79
7cpa					0.67	0.81	1.60 (1.39)	10.6	0.87	0.77
1cbx	7cpa	3	$1.3 \cdot 10^{15}$	1187	×	×	×	0.8	0.44	×
6cpa					×	×	×	24.3	0.87	×
7cpa					0.77	0.87	2.40 (2.32)	46.6	1.0	0.77
Elastase inhibitors										
1ela	1ela	2	$2.0 \cdot 10^{7}$	262	1.0	0.95	1.46 (0.89)	3.0	1.0	1.0
1ele					0.83	0.92	1.98 (1.08)	2.2	0.87	0.95
1ela	1ele	2	15552	121	0.72	0.83	1.24 (1.22)	0.5	0.87	0.83
1ele					0.84	0.77	2.15 (1.20)	0.5	1.0	0.84

as possible, while the fragment pairs were still large enough to allow for a complete assembly. Databases generated this way were queried with all molecules in the test set that bind to the same receptor as the database molecule. As seen in Table 3 most of the molecules have been partitioned into 3 fragments or 2 fragment pairs. Only in the case of 1dwc vs. 1dwd (row 3 in Table 3) could fully assembled candidates be obtained from 3 fragment pairs for a non-self query. In this case the parameter p (Eq. 7) was set to $p = 7$ in order to collect enough hypotheses.

Table 3 summarizes the results of the mutual molecular superpositions. Candidates were first ordered by the votes score and then by Carbo score for candidates with identical votes score. Votes score, Carbo score, and the RMSD w.r.t. to the correct alignment are shown for the highest scoring candidate. The smallest RMSD value out of the 10 highest scoring candidates is also given in parentheses. Figure 8 shows the result of a self-query of the carboxypeptidase A inhibitor 6cpa. The molecule had been partitioned into two fragment pairs which reduces the original number of conformations screened in the dihedral angle expansion from $3.6 \cdot 10^{8}$ to 384 explicitly represented fragment pair conformations. The correct molecular alignment is obtained within an accuracy of 1.37 Å RMSD, and almost all possible feature correspondences (pairs of spheres in Figure 8) are found despite the fact that the exact conformation of the query molecule was not contained in the database. This indicates that a torsional angle resolution of $\Delta\varphi = 60°$ is obviously sufficient for molecules of this size. We expect that larger molecular structures require smaller values of $\Delta\varphi$ in order to avoid missing important molecular conformations. As mentioned above, the largest molecule considered here (7cpa with 74 atoms) is the only structure that required a setting of $\Delta\varphi = 30°$ in order to obtain the self-alignment correctly.

In the following we will discuss the results shown in Table 3 in more detail.

1. *Thrombin inhibitors* (see Figure 9). The two ligands (1dwc and 1dwd) have similar size (71 and 69 atoms), functionally similar structure elements (guanidine vs. benzamidine), but backbones of rather different topologies. We find convincing mutual alignments between 1dwd and 1dwc with RMSDs of 1.93 and 2.28 Å which are not simple substructure matches.

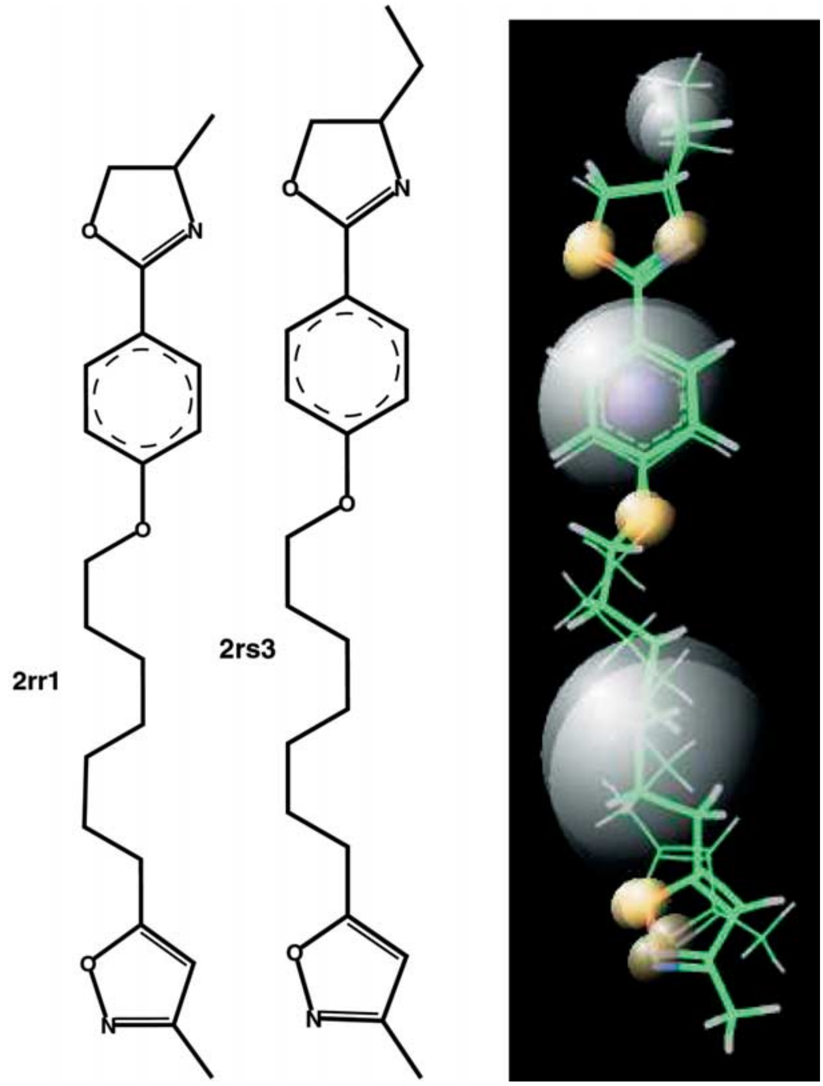


Figure 10. Molecular superposition of the human rhinovirus coating protein inhibitors 2rr1 and 2rs3, where 2rs3 was stored in the database. Votes and Carbo score for this alignment were 1.0 and 0.93, the RMSD w.r.t. the crystal structure was 2.16 Å, and query processing took 1.8 s CPU time. The colored spheres represent different types of features (for an explanation of the colors see caption of Figure 8).

2. *Human rhinovirus coating protein ligands* (see Figure 10). All 4 ligands considered are structurally very similar – a stretched, flexible hydrocarbon chain capped with two 5-ring heterocyclic termini. The ligands fall into two groups representing binding modes, which differ by a reverse orientation of the entire molecules. fFLASH finds excellent alignments ($<$1.8 Å) between ligands that have the same binding mode.

3. *Fructose biphosphatase inhibitors* (see Figure 11). The chemical structures of the two ligands consist of an N-heterocyclic system bound to a sugar phosphate backbone. The alignments found show a substructure match on the backbone and a non-trivial match involving a reverse orientation of the 5-rings present in both structures (purine in 4fbp and imidazole in t0039) with an RMSD of $<$ 0.6 Å.

4. *Dihydrofolate reductase (DHFR) inhibitors* (see Figure 12). Both ligands, dihydrofolate (1dhf) and methotrexate (4dfr), have a pteridine substructure in common that participates in the ligand's binding modes. fFLASH finds the correct relative orientation of these substructures resulting from the match of the characteristic hydrogen bond donor/acceptor pattern (see pairs of spheres in Figure 12 showing feature correspondences). The overall minimum RMSDs are 1.73

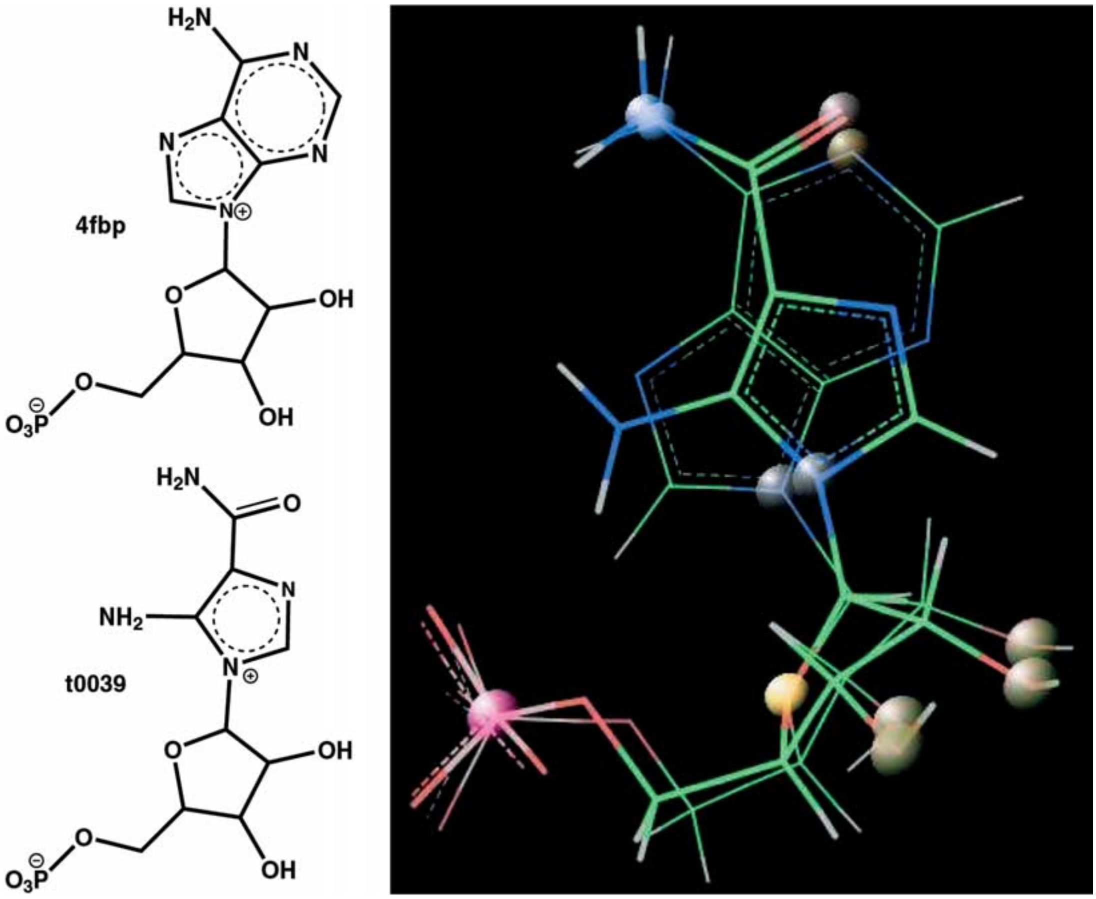


Figure 11. Molecular superposition of the fructose biphosphatase ligands 4fbp and t0039, where 4fbp was stored in the database. Votes and Carbo score for this alignment were 0.63 and 0.96, the RMSD w.r.t. the crystal structure was 0.57 Å, and query processing took 0.2 s CPU time. The colored spheres represent different types of features (for an explanation of the colors see caption of Figure 8).

(4dfr as query molecule) and 2.30 Å (1dhf as query molecule). These relatively large values reflect the fact that all features on the molecule, including those on the hydrophilic tail, are assumed to be of equal importance. fFLASH therefore interpolates between the pteridine ring pattern match and the matches of the benzene ring and the carboxyl groups for the overall superposition, while in reality the alignment of the molecules is determined mainly by the pteridine ring match.

5. *Thermolysin inhibitors* (see Figure 13). The 4 ligands all possess a similar binding mode, but are very different in size. All queries against 1tlp (especially the self-query) suffer a performance toll from a high load of hypotheses caused by the sugar substructure (which is only present in 1tlp) with its many hydrogen bond donors and acceptors. The (non-self) queries of the smallest ligand 2tmn (26 atoms) against the larger molecules pose a problem because of the size mismatch – no candidates are produced because of missing hypotheses.

6. *Carboxypeptidase inhibitors* (see Figure 14). The 3 ligands 1cbx, 6cpa, and 7cpa are of very different size (25, 58, and 74 atoms) but all are chemically and structurally similar, with 1cbx being a substructure of 6cpa, and 6cpa being a substructure of 7cpa. Again no candidates are found in queries of the smaller ligands against the larger ones because of the size mismatch. Queries of the bigger ligands 6cpa, and 7cpa suffer a performance toll from a high load of hypotheses caused by matches on unspecific regions of the molecules such as amide and benzyl groups. The query of the largest structure 7cpa against 6cpa produces an excellent alignment with an RMSD of 1.39 Å.

7. *Elastase inhibitors* (see Figure 15). Both ligands, 1ela and 1ele, are of peptidic nature and possess

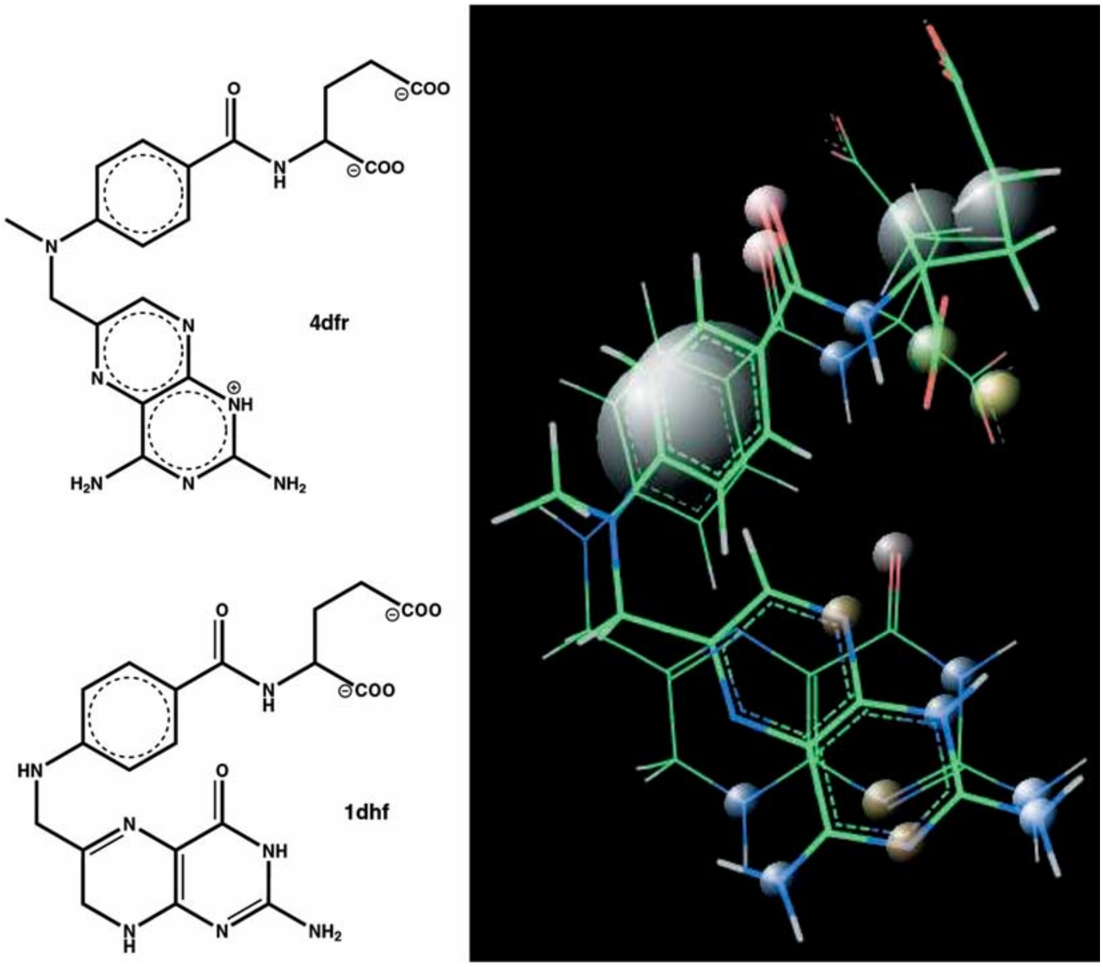


Figure 12. Molecular superposition of the dihydrofolate reductase inhibitors 1dhf and 4dfr, where 1dhf was stored in the database. Votes and Carbo score for this alignment were 0.70 and 0.87, the RMSD w.r.t. the crystal structure was 1.73 Å, and query processing took 0.1 s CPU time. The colored spheres represent different types of features (for an explanation of the colors see caption of Figure 8).

similar binding modes, with 1ele being primarily a substructure of 1ela with minor variations in backbone substituents. All alignments (self and mutual) obtained are excellent with RMSDs ranging from 0.89 to 1.22 Å.

fFLASH exhibits an inherent asymmetry in mutual molecular alignments w.r.t. the size of the structures. The query molecule should be larger than or of approximately equal size as the database molecule, otherwise candidates can most likely not be assembled since hypotheses are missing. Substructures or approximate substructures are found only if the substructure itself is the database molecule. This becomes apparent in the cases 5 and 6 above. The CPU times for query processing (nineth column of Table 3) are in most cases very encouraging, being mostly of the order of a few seconds, in many cases even only fractions of a second. Factors contributing to a slower performance are, of course, the number of conformations that have to be processed, and also the size of the matching patterns. Self-queries generally take a longer time than mutual superpositions. Many relatively unspecific feature correspondences (see the sugar rest of 1tlp in case 5) drastically slow down the clique detection algorithm since the underlying graph (see section 2.3) exhibits more edges. An almost ideal case is the methotrexate-dihydrofolate alignment with its very specific, rigid pattern match on the pteridine ring. This, together with the relatively small number of fragment pair conformations that have to be processed, leads to a very short processing time of about 0.1 second.

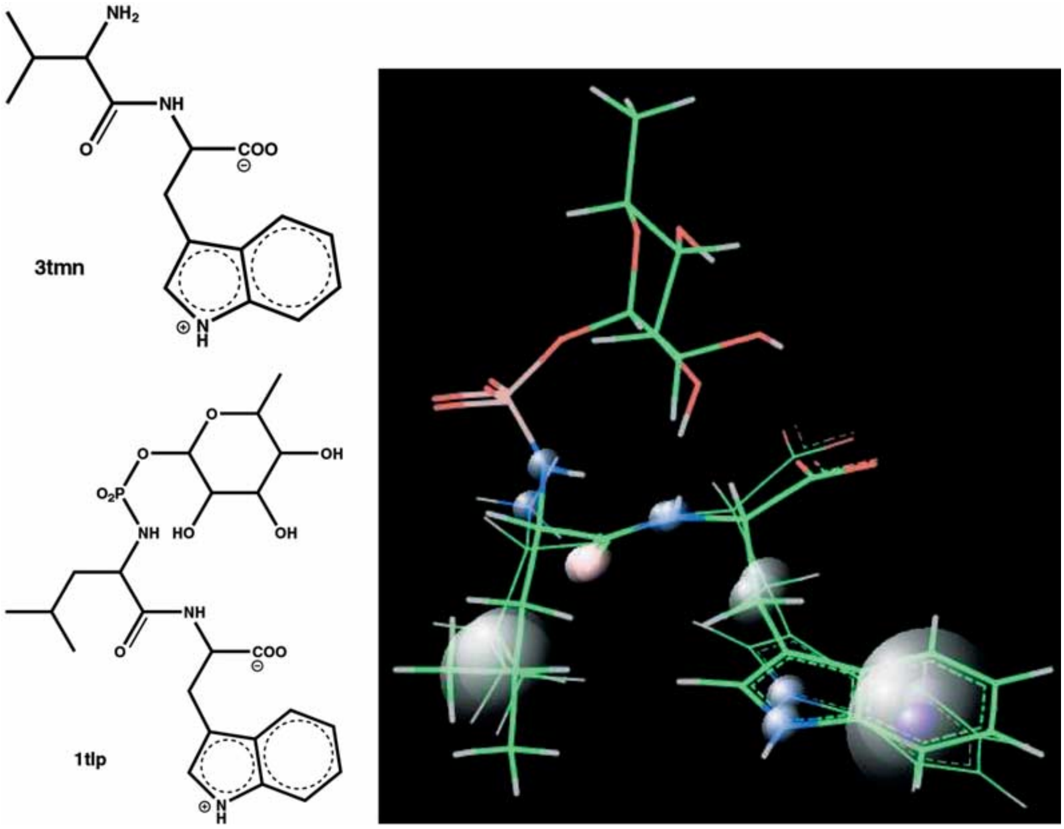


Figure 13. Molecular superposition of the thermolysin inhibitors 1tlp and 3tmn, where 3tmn was stored in the database. Votes and Carbo score for this alignment were 0.58 and 0.76, the RMSD w.r.t. the crystal structure was 1.36 Å, and query processing took 0.1 s CPU time. The colored spheres represent different types of features (for an explanation of the colors see caption of Figure 8).

Database search

We have performed queries against a test database containing 1780 molecules, 1728 compounds from the NCI diversity set [28][1], 50 known DHFR inhibitors taken from Crippen [27], methotrexate (4dfr), and dihydrofolate (1dhf). The general structure of the molecules taken from [27] is shown in Figure 16. All database molecules have been partitioned by an automatic procedure that is based on the partitioning algorithm described in section 2.1. The total number of fragment pair conformations, which are explicitly represented, is 32872. Database preprocessing, which includes the generation and storage of all fragment pair conformations from an arbitrary starting conformation for each molecule as well as feature computation and creation of fFLASH's central lookup table, took about 35 minutes. The size distribution of molecules in the test database is given in Figure 17(a), and the distribution of the number of molecular conformations is shown in Figure 17(b). In the latter diagram we distinguish between the number of conformations obtained from the uniform dihedral angle expansion, and the number of fragment pair conformations explicitly represented in the database. Figure 17(b) clearly illustrates the performance impact of our 'coarse graining' scheme (see section 2.2) for sampling the conformational space: in the best case, a 10^6 fold savings in the number of conformations processed is achieved.

Figure 18(a) shows the votes scores and Carbo scores of the highest scoring candidate for each database molecule using dihydrofolate (1dhf) as the query molecule. In this experiment the total query processing time was approximately 3 min, i.e. on the average 0.1 s per database molecule. Points in Figure 18(a) represent the molecules from the NCI diversity set, while triangles correspond to DHFR inhibitors from Crippen's dataset [27]. Dihydrofolate and methotrexate are shown as squares. 45% of the database molecules (including 6 DHFR inhibitors, i.e. 11%) are not shown

[1] Salts/complexes with metal ions and disjoint molecular clusters were excluded.

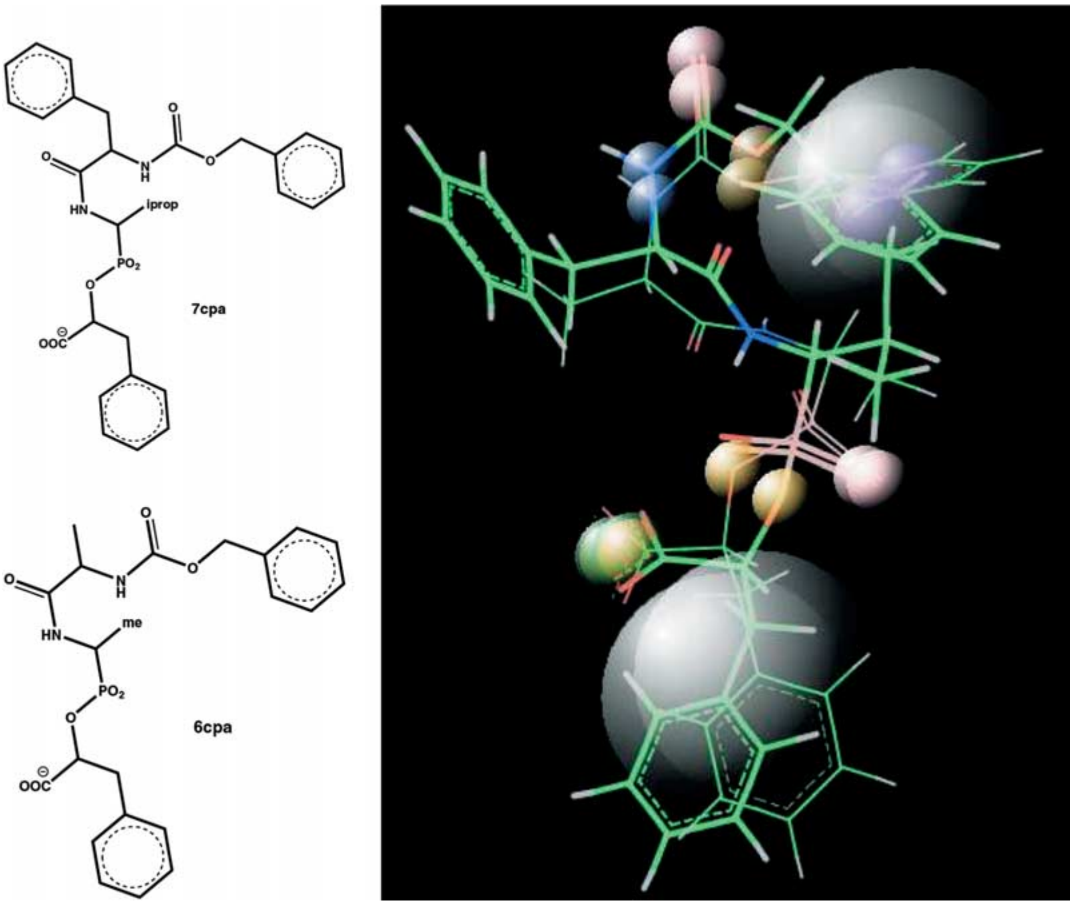


Figure 14. Molecular superposition of the carboxypeptidase A inhibitors 7cpa and 6cpa, where 6cpa was stored in the database. Votes and Carbo score for this alignment were 0.67 and 0.81, the RMSD w.r.t. the crystal structure was 1.60 Å, and query processing took 10.6 s CPU time. The colored spheres represent different types of features (for an explanation of the colors see caption of Figure 8).

in the plot since they do not get enough support to be assembled as candidates. Visual inspection of the data shows that dihydrofolate (self query) and methotrexate as well as about 50% of the other DHFR inhibitors receive the highest votes and Carbo scores, and are clearly separated from the 'background noise' produced by the NCI diversity set. Reference experiments using other query molecules (see Table 4) did not alter the location of the point cloud from the diversity set, indicating as expected that the statistics of the background noise in the score plot (Figure 18(a)) are independent of the particular query molecule. Of course, the query processing time strongly depends on the size of the query molecule (or, more precisely, the number of its features). The average processing time per database molecule is shown for a number of different query molecules in Table 4.

Table 4. Average processing time per molecule for queries against the test database (see text). CPU times refer to a 1.8 GHz PC with Pentium 4 processor.

Query molecule	Number of		CPU time (s)
	atoms	features	
1dhf	51	19	0.095
4dfr	54	19	0.080
1dwc	71	20	0.11
1ela	64	17	0.067
1tlp	69	27	0.45
1tmn	67	18	0.070
2r04	51	11	0.036
4fbp	35	15	0.10
6cpa	58	22	0.22
7cpa	74	23	0.30

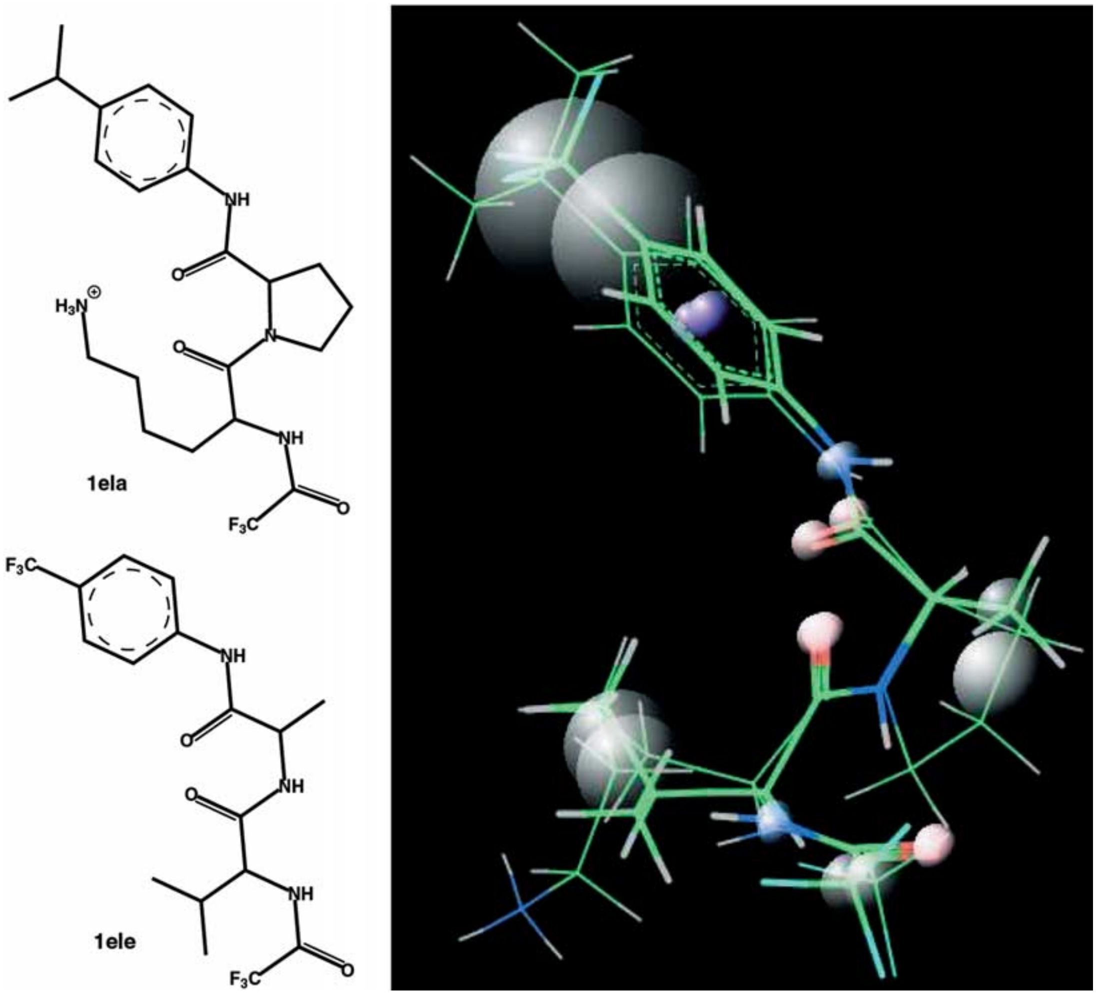
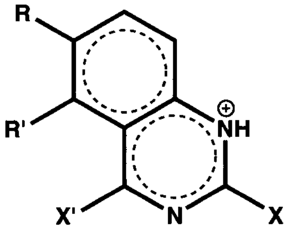


Figure 15. Molecular superposition of the elastase inhibitors 1ele and 1ela, where 1ela was stored in the database. Votes and Carbo score for this alignment were 0.83 and 0.92, the RMSD w.r.t. the crystal structure was 1.98 Å, and query processing took 2.2 s CPU time. The colored spheres represent different types of features (for an explanation of the colors see caption of Figure 8).



X = NH_2, OH, SH, H, ...
X'= NH_2, OH, SH, ...
R = SO_2-napht, S-napht, ...
R'= H, CH_3, ...

Figure 16. General structure of the DHFR inhibitors from the Crippen dataset [27].

A quantitative analysis of the score plot (Figure 18(a)) can be done in the following way: We note that the votes score roughly correlates linearly with the Carbo score for the diversity set and we therefore perform a least squares fit with a linear function for the corresponding points. This is the straight line with slope 0.55 in Figure 18(a). ‘Hits’ of a database query, i.e. those molecules considered to be similar to the query molecule, should fall close to the upper right corner of the plot. We construct a rectangle with sides parallel to the coordinate axes and the lower left corner located on the straight line at a certain ‘cutoff’ value of the Carbo score (see example shown in Figure 18(a)). In the following, all points that fall into that rectangle are considered to be hits of the database

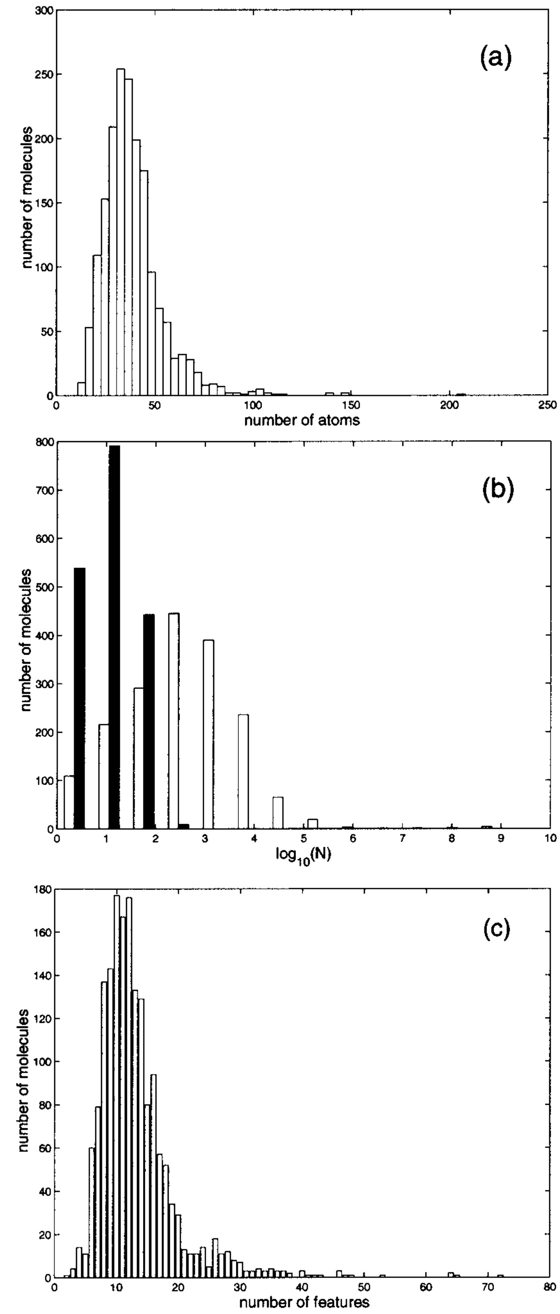


Figure 17. Characterization of the subset of the NCI diversity set [28] used in the fFLASH test database.
(a) Distribution of the number of atoms per molecule.
(b) Distribution of the number of conformations per molecule. White bars: N = number of conformations after dihedral angle expansion with $\Delta\varphi = 60°$. Black bars: N = number of fragment pair conformations after 'course graining'. Among the 1728 compounds, there are 1156 that consist of 1 fragment pair, 554 molecules that are partitioned into 2 fragment pairs, 15 molecules with 3 fragment pairs, and 3 molecules with 4 fragment pairs.
(c) Distribution of the number of features per molecule.

query. Figure 18(b) shows the percentage of hits relative to the size of the diversity set (continuous line) as well as relative to the number of DHFR inhibitors (broken line) as a function of the Carbo score cutoff parameter. This illustrates the balance of selectivity and sensitivity, i.e. the signal-to-noise ratio of the similarity search method. The lines in this plot may be interpreted as false positives (diversity set) and true positives (DHFR inhibitors). It is seen that if 50% of the DHFR inhibitors are found (50% false negatives at a Carbo score cutoff value of 0.69), we also find about 3% false positives.

In order to illustrate that the distinction between DHFR actives and the diversity set is non-trivial, we binned the query results w.r.t. the number of matches that were observed within topologic regions of the query molecule 1dhf as defined in Figure 19(a). The results (Figures 19(b) and (c)) clearly show that the distinction is based on the presence of an extended feature pattern in region 1 (the NCNCN motif in the pteridine substructure of 1dhf): about 80% of all DHFR actives show 4 or more matches here, as opposed to only about 20% from the NCI diversity set.

It is interesting to analyze some of the higher scoring molecules in the diversity set in more detail. Figure 20 shows the candidate alignment found for the point marked by an arrow and an asterisk in the score plot (Figure 18(a)). This molecule from the diversity set (NCI index 127917) shows a number of remarkably similar features to dihydrofolate 1dhf. Namely, parts of the characteristic donor/acceptor pattern on the pteridine ring important for DHFR binding are found as well as a common benzene ring, and an overall shape match.

Discussion and outlook

We have shown that fFLASH is capable of rapidly screening a database of flexible, drug-like molecules for candidates that are similar to a given rigid query molecule, while producing accurate three-dimensional molecular superpositions. We may compare the quality of alignments produced by fFLASH with those published for the FlexS method [9]. FlexS finds all molecular alignments shown in Table 3 except for 3tmn vs. 1tlp and 3tmn vs. 1tmn, while fFLASH does not find the correct superposition of 6cpa vs. 7cpa which is reproduced correctly by FlexS. The values for the smallest RMSD among the 10 best scoring candidates published in [9] are on the average 0.23 Å

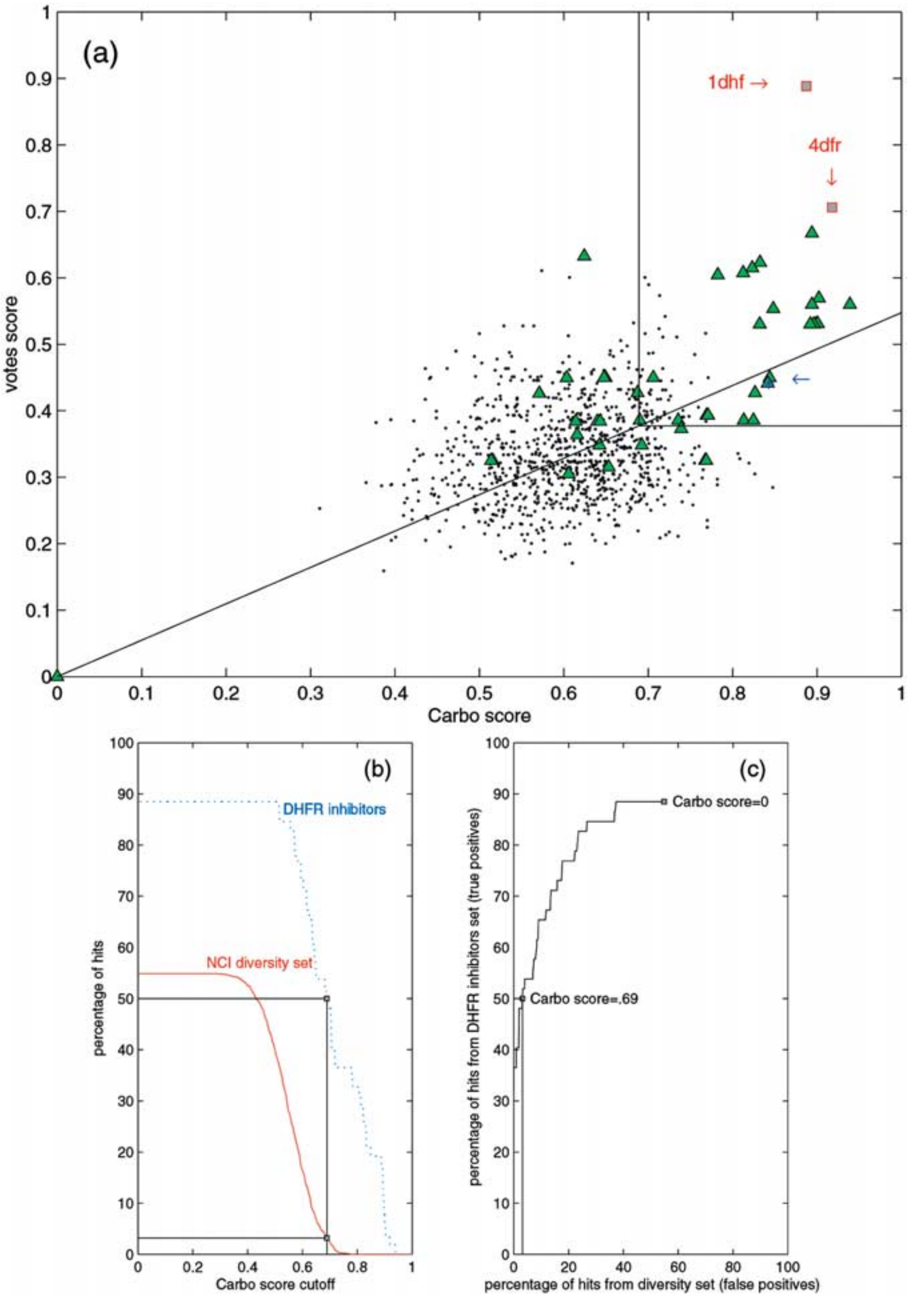


Figure 18. Result of a dihydrofolate (1dhf) query against the fFLASH test database. Diagram (a) shows the Carbo and votes scores of the highest scoring candidate for each molecule that passes the assembly phase. Molecules from the NCI diversity set are shown as dots (·), DHFR inhibitors from the Crippen set [27] are represented by triangles (△), and methotrexate (4dfr) and the dihydrofolate self query are shown as squares (□). The arrow points to the high scoring candidate (∗) from the diversity set (NCI index 127917) for which the molecular alignment is shown in Figure 20. The straight line and the rectangle in the upper right corner refer to the method described in the text to quantitatively define 'hits' of a database query. Diagram (b) shows the percentage of hits as a function of the Carbo score cutoff value for the NCI diversity set (continuous line) and for the DHFR inhibitors (broken line). Note that a Carbo score cutoff of e.g. 0.2 refers to all candidates having a Carbo score of 0.2 or higher.

Diagram (c) shows an ROC (receiver-operator curve) plotting the *true* vs the *false* positive hits.

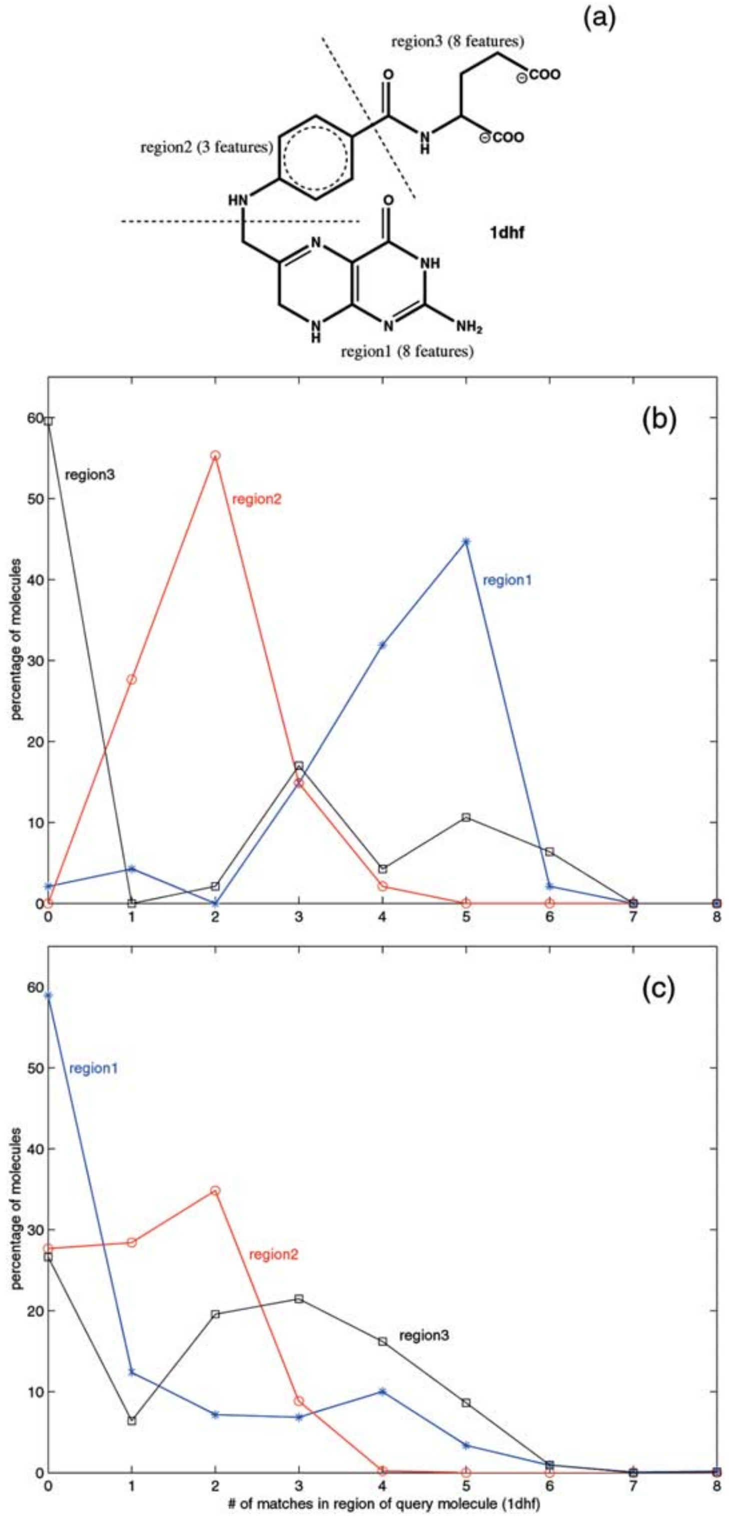


Figure 19. Regional binning of candidates from a dihydrofolate (1dhf) query against the fFLASH test database.
(a) Definition of the topologic regions considered for query molecule 1dhf.
(b) Percentage of candidates from the DHFR actives set that have matches within regions 1, 2 and 3.
(c) Percentage of candidates from the NCI diversity set that have matches within regions 1, 2 and 3.

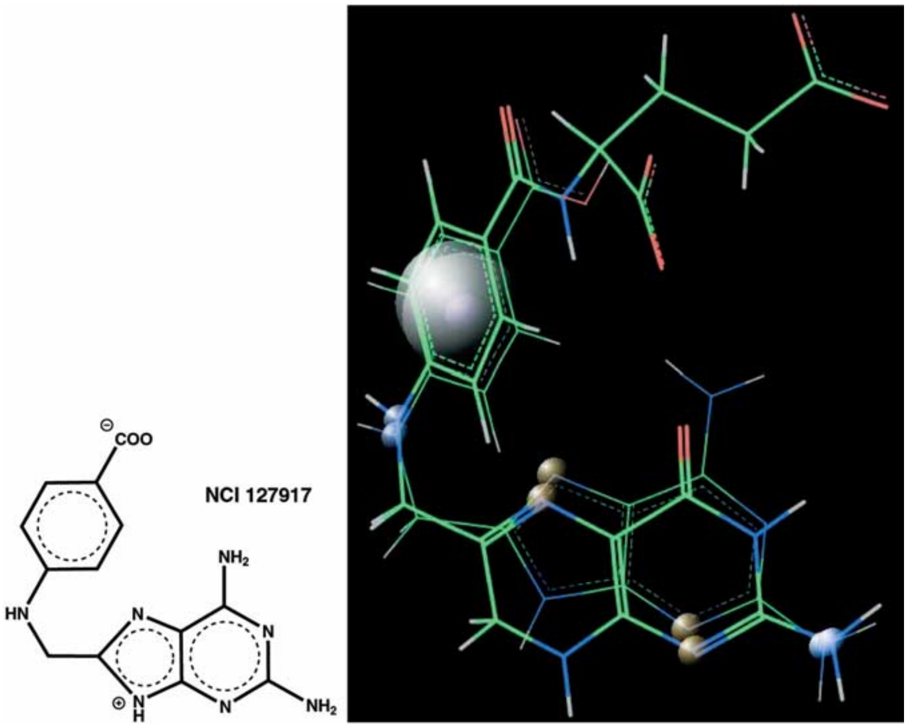


Figure 20. Molecular alignment for the high scoring candidate from the diversity set (NCI index 127917) shown as an asterisk (∗) in Figure 18(a). The colored spheres represent different types of features (for an explanation of the colors see caption of Figure 8).

smaller than the corresponding RMSDs obtained with fFLASH. This might be attributed to the fact that FlexS utilizes a user-defined anchor fragment (e.g. the pteridine ring in the methotrexate-dihydrofolate alignment). The choice of this anchor fragment may already reflect the importance of certain molecular regions for the superposition, and thus may lead to slightly more accurate molecular alignments. In contrast to that, all molecular features are considered to be equally important in the fFLASH algorithm. For a database screening application, the actual RMSD value does not play a significant role as long as qualitatively correct molecular alignments are detected that result in high scores.

In a drug discovery environment fFLASH could be utilized to efficiently generate focussed libraries from a large compound database for subsequent assays by collecting hits from one or several query structures. The produced 3D alignments could also be exploited by electronic screening methods (e.g. 3D QSAR applications), or could help to identify new leads by providing chemical insight through visual inspection. In the following, we make a few remarks about the direction of future developments of fFLASH that will increase its applicability, and also address some of the current limitations.

fFLASH currently uses a linear chain-like fragmentation approach which generally seems to be sufficient for medium-sized, drug-like molecules. Nevertheless, a more general scheme should be able to handle closures of large ring structures, and, more importantly, be able to deal with branched molecules. In section 2.2 we already mentioned the issue of asymmetry between query and database molecules w.r.t. their size. This asymmetry can be reduced by allowing for the assembly of partial candidates, and appending the missing fragment 'tails' in an arbitrary conformation. Likewise, the sensitivity of the method can be increased by bridging gaps within a candidate if certain fragment pairs do not get enough support in the pattern matching procedure.

The user should be able to actively manipulate features on the query molecule by introducing weights, or switching features on and off according to their importance in the receptor-ligand interaction. This way, the user may distinguish between parts of the query molecule that are solvent-exposed, and regions that bind to the protein. The resulting molecular alignments could be further improved in a postprocessing step by relaxing the conformational constraints, and performing a continuous optimization of the RMSD of feature matches. In another direction, fFLASH can straightforwardly be generalized to start from a receptor model instead of a rigid query molecule in order to allow for docking-type applications [35, 36].

In the present paper we made use of a relatively simple, atom type-based feature definition. This feature definition was shown to yield good results for the set of medium-sized drug-like molecules considered here. In addition, we have performed tests with larger molecules (HIV protease inhibitors) from the benchmark set [26] and found correct alignments in some cases. However, *local* shape similarity, e.g. the packing of hydrophobic side-chains, plays an important role for these structures that is not properly resolved in the atom-based feature definition used here. Therefore we argue that other, more appropriate feature schemes which also encode local shape properties should be used in those situations. This is, in principle, possible because features are explicitly allowed to depend on fragment pair conformations, a capability which was already used in our 'geometric' definition of hydrophobic regions.

A first step in the direction of shape-dependent features has already been undertaken by the construction of a feature scheme that distinguishes between (atom type-based) features in a locally flat or non-flat environment. This, of course, increases the selectivity of the molecular similarity search method since we now have a larger set of feature indices, and therefore receive fewer feature correspondences. A convenient side-effect is that query processing times are reduced by a factor of 2–3, because the work-load for the clique detection algorithm is smaller. Using this feature scheme we were able to reproduce all but 4 alignments reported in Table 3, and obtained results qualitatively similar to those shown in Figure 17 when querying dihydrofolate against our test database.

A generalization of fFLASH for the application to combinatorial libraries is under way.

Conclusion

We have described a new method (fFLASH) for the virtual screening of compound databases that is based on explicit three-dimensional molecular superpositions that take the torsional flexibility of the database molecules fully into account. Molecules are represented by an arbitrary number of point-like features that are allowed to explicitly depend on the conformation of molecular fragments. The method uses an extensive database preprocessing step during which features are pre-calculated and stored in a lookup table in order to minimize computer time needed for a database query. Using a conventional molecular feature definition (hydrogen bond donors and acceptors, charges, and hydrophobic groups) we have shown that fFLASH is able to rapidly produce accurate alignments for a test set of pairs of medium-sized drug-like molecules which are known to bind to the same receptor. Experiments with a test database containing a diverse set of 1728 drug-like molecules from the NCI diversity set as well as 52 dihydrofolate reductase actives have shown that average query processing times of the order of 0.1 seconds per molecule (including all conformers) can be achieved on a PC. These results suggest that fFLASH could be applied efficiently to focused library design using much larger datasets. Since the fFLASH algorithm is naturally parallelizable (all database molecules are processed independently of each other), fFLASH could be straightforwardly scaled up to run on a farm of – say 50 – PCs. Extrapolating our current results, such a system would be capable of searching a database of several hundred thousand molecules in a few minutes.

Acknowledgements

We would like to thank Jed Pitera, William Swope and Mike Pitman for valuable discussions.

References

1. Matter, H. and Rarey., M., In Jung, G. (Ed.) Combinatorial Organic Chemistry. John Wiley & Sons, New York, NY, 1999.
2. Humblet, C. and Dunbar Jr., J.B., In Venuti, M.C. (Ed.) Annual Reports in Medicinal Chemistry. Vol. 28, Chapter VI. Topics in Drug Design and Discovery. Academic Press, London, 1993, pp. 275–284.
3. Willet, P., J. Mol. Recognition, 8 (1995) 290.
4. Brown, R.D. and Martin, Y.C., J. Chem. Inf. Comput. Sci., 36 (1996) 572.
5. Klebe, G., Mietzner, T., and Weber, F., J. Comput. Aided Mol. Des., 8 (1994) 751.
6. Kearsley, S.K. and Smith, G.M., J. Comput. Aided Mol. Des., 8 (1994) 565.
7. Klebe, G., Mietzner, T., and Weber, F., J. Comput. Aided Mol. Des., 13 (1999) 35.
8. Lemmen, C., and Lengauer, T., J. Comput. Aided Mol. Des., 11 (1997) 357.
9. Lemmen, C., Lengauer, T., and Klebe, G., J. Med. Chem., 41 (1998) 4502.

10. Lemmen, C., Hiller, C., and Lengauer, T., J. Comput. Aided Mol. Des., 12 (1998) 491.
11. Miller, M.D., Sheridan, R.P., and Kearsley, S.K., J. Med. Chem., 42 (1999), 1505.
12. Grant, J.A., Gallardo, M.A., Pickup, B.T., J. Comput. Chem., 17 (1996) 1653.
13. McMartin, C. and Bohacek, R.S., J. Med. Chem., 42 (1999) 1505.
14. Handschuh, S., Wagener, M. and Gasteiger, J., J. Chem. Inf. Comput. Sci., 38 (1998) 220.
15. Lemmen, C., and Lengauer, T., J. Comput. Aided Mol. Des., 14 (2000) 215.
16. Lemmen, C., Zimmermann, M., and Lengauer, T., Perspectives in Drug Discovery and Design, 20 (2000) 43.
17. Kearsley, S.K, J. Comput. Chem., 11 (1990) 1187.
18. Diamond, R., Protein Sci., 1 (1992) 1279.
19. Mestres, J., Maggiora, G.M. and Rohrer, D.C., J. Mol. Graph., 15 (1997) 114.
20. Cosgrove, D.A., Bayada, D.M. and Johnson, A.P., J. Comput. Aided Mol. Des., 14 (2000) 573.
21. Labute, P., William, C., Feher, M., Sourial, E. and Schmidt, J.M., J. Med. Chem. 44 (2001) 1483.
22. Mills, J.E.J., de Esch, I.J.P., Perkins, T.D.J. and Dean, P.M., J. Comput. Aided Mol. Des., 15 (2001) 81.
23. Roberts, G.C.K., Drug Discovery Today, 5 (2000) 230.
24. Pitman, M.C., Huber, W.K., Horn, H., Krämer, A., Rice, J.E. and Swope, W.C., J. Comput. Aided Mol. Des., 15 (2001) 587.
25. Lawton, J., Tudor, M. and Wipke, W.T., In Parrill, A.L. and Reddy, M.R. (Eds.) Rational Drug Design: Novel Methodology and Practical Applications. ACS Symposium Series 719, Oxford University Press, 1999, pp. 239–254.
26. FlexS-77 dataset collected by C. Lemmen, G. Klebe, M. Böhm, first published in [9] (http://www.biosolveit.de)
27. Compounds 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 15, 16, 17, 18, 19, 21, 22, 23, 24, 25, 29, 30, 31, 32, 34, 35, 36, 38, 39, 41, 42, 43, 45, 46, 47, 48, 50, 52, 54, 56, 57, 60, 62, 63, 64, 65, 66 from Crippen, G.M., J. Med. Chem. 23 (1980) 599.
28. The NCI diversity set can be found at http://dtp.nci.nih.gov/docs/3d database/structural information /structural data.html.
29. Carbo, R., Leyda, L. and Arnaua, M., Int. J. Quant. Chem., 17 (1980) 1185. In the present work the Carbo function is used with three-dimensional Gaussian densities $\exp(-(\vec{r} - \vec{r}_0)^2/2\tau^2)$, where $\vec{r}_0$ is an atom center and $\tau = 1$ Å [24].
30. See e.g. in Chartrand, G., Introductory Graph Theory, Dover, 1985.
31. See e.g. in Sedgewick, R.,Algorithms in C++, Addison-Wesley, New York, 1992.
32. Pardalos, P.M., University of Florida, 1997.
33. Clark, M., Cramer III, R.D. and Van Opdenbosch, N., J. Comput. Chem. 10 (1989) 982.
34. Bondi, A., J.Phys.Chem., 68 (1964) 441.
35. Mason, J.S., Morize, I., Menard, P.R., Cheney, D.L., Hulme, C. and Labau-diniere, R.F., J. Med. Chem. 42 (1999) 3251.
36. Eksterowicz, J.E., Evensen, E., Lemmen, C., Brady, G.P., Lanctot, J.K., Bradley, E.K., Saiah, E., Robinson, L.A., Grootenhuis, P.D.J. and Blaney, J.M., J. Mol. Graphics Modelling 20 (2002) 469.