Automatic and Efficient Decomposition of Two-Dimensional Structures of Small Molecules for Fragment-Based High-Throughput Docking

Peter Kolb* and Amedeo Caflisch*

Department of Biochemistry, University of Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland

Received July 17, 2006

The computer program DAIM (Decomposition and Identification of Molecules) has been developed to automatically break up compounds in small-molecule libraries for fragment-based docking as well as database analysis. Here, DAIM is evaluated on 130 ligands derived from known crystal structures of ligand−protein complexes. The decomposition and a new fingerprint-based identification technique are used to select anchor fragments for docking. The docking results show that the DAIM selection is superior to size-based or random selection of fragments. To evaluate the usefulness for analyzing the fragment composition of a large library, DAIM is applied to a collection of about 1.85 million commercially available compounds. Interestingly, it is found that the set of most frequent cyclic and acyclic fragments originating from the decomposition of the 1.85 million molecules shows a large overlap with the most frequent fragments in a library of 5120 known drugs. DAIM has been successfully used in the in silico screening for inhibitors of β-secretase and EphB4 kinase by fragment-based high-throughput docking. Possible future applications for de novo ligand design are briefly discussed.

1. Introduction

The ever-increasing understanding of human diseases at a molecular level is spurring considerable interest in small-molecule inhibitors of enzymes and receptors. Because of the large number of 3D-structures of pharmacologically relevant protein targets, structure-based computer-aided approaches are useful and widely employed tools in drug discovery.[1] High-throughput docking has recently emerged as a very cost-effective and efficient alternative to in vitro screening campaigns to discover lead compounds.[2−7] Furthermore, docking and accurate methods to calculate the binding free energy are being used for the in silico evaluation of chemical modifications of initial hits to guide the synthesis of molecules with more favorable binding constants.[8−10]

Efficient docking algorithms include fragment-based approaches in which (almost) rigid molecular fragments are automatically placed in the binding site and used as “anchors” to guide the docking of the compounds they originate from. Fragment-based docking is a divide-and-conquer approach, which was introduced 15 years ago in the context of computational ligand design.[11−15] In vitro fragment-based search strategies have been developed[16−18] using NMR[19,20] and X-ray crystallography,[21] and successful screening campaigns have been reported for several targets, including kinases and DNA gyrase.[22−24] Although small molecular fragments usually bind unspecifically with IC_{50} values in the low millimolar range, they sometimes exhibit significant “efficiency” (i.e., binding energy per atom[18,25]). Furthermore, the probability of a match between a protein binding site and a small molecule decreases exponentially with the complexity of the molecule.[26] Hence, as small molecular fragments bind in multiple sites, they are especially useful to chart the characteristics of a protein surface. Because of their small number of atoms, they can be chemically modified to improve potency and other properties (e.g., solubility) to a greater extent without increasing too much in molecular weight.

In this work, an approach for the efficient decomposition of molecules into mainly rigid fragments is presented and evaluated. The method has been implemented in a computer program called DAIM (Decomposition and Identification of Molecules) and is part of our fragment-based docking approach SEED/FFLD. For every compound in a library, DAIM identifies the three fragments most appropriate for fragment-based docking. The fragment triplets are then docked by the program SEED (Solvation Energy for Exhaustive Docking).[27,28] The most favorable poses of the three “anchor” fragments guide the docking of the molecule inside the binding site by FFLD (Fast Flexible Ligand Docking).[29,30] Recently, the entire suite of programs has successfully been applied to identify inhibitors of β-secretase.[5,6] Here, three analyses are performed to further assess the usefulness of the DAIM decomposition and anchor selection procedure for docking and design. First, it is investigated whether, using solely the 2D-structure, DAIM can predict the ligand fragments that are involved in the highest number of contacts in X-ray structures of ligand−protein complexes. Second, the binding mode obtained using the fragment triplet selected by DAIM is compared to the results of docking using all other fragment triplets. Finally, a collection of about 1.85 million available chemical compounds is decomposed by DAIM, and the resulting fragments are compared with those originating from a database of known drugs to evaluate overlap and differences.

2. Methods

2.1. Fragment-Based Docking. The fragment-based docking approach consists of three steps (Figure 1): (1) decomposition of each molecule of the library into mainly rigid fragments, (2) fragment docking with evaluation of electrostatic solvation, and (3) flexible docking of each molecule of the library using the positions of its fragments as anchors. Step 1 is performed by the program DAIM, which is described in detail in the following subsection. The programs used in steps 2 and 3 will be briefly described here.

Step 2: The docking approach implemented in the program SEED determines optimal positions and orientations of small- to medium-sized molecular fragments in the binding site of a

* To whom correspondence should be addressed. Phone: (41 44) 635 55 21 (P.K.; A.C.). Fax: (41 44) 635 68 62 (P.K.; A.C.). E-mail: caflisch@bioc.unizh.ch (A.C.); pkolb@bioc.unizh.ch (P.K.).

10.1021/jm060838i CCC: $33.50 © 2006 American Chemical Society
Published on Web 11/21/2006

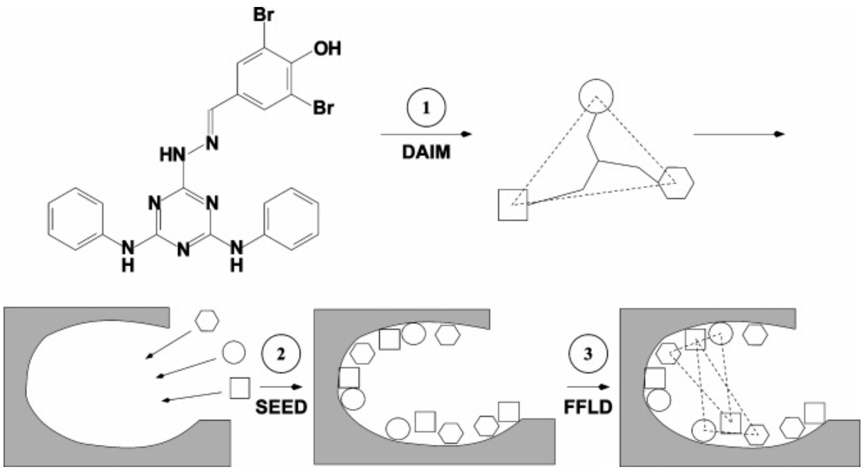


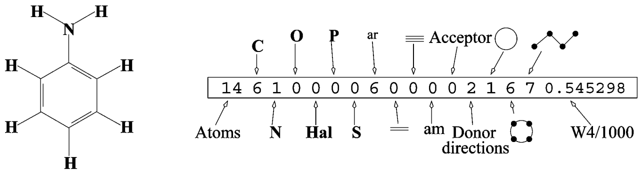
Figure 1. Schematic depiction of the fragment-based docking procedure. The programs used for steps 1, 2, and 3 are DAIM, SEED, and FFLD, respectively.

protein.[27,28] Apolar fragments are docked into hydrophobic regions of the receptor while polar fragments are positioned such that at least one intermolecular hydrogen bond is formed. Each fragment is placed at several thousand different positions with multiple orientations (for a total of on the order of 10^6 conformations), and the binding energy is estimated only if there are no severe clashes (usually about 10^5 conformations). The binding energy is the sum of the van der Waals interaction and the electrostatic energy. The latter consists of the screened receptor-fragment interaction, as well as the desolvation penalty of receptor and fragment.[31]

Step 3: The flexible-ligand docking approach FFLD uses a genetic algorithm and a very efficient but approximate scoring function.[29,30] FFLD requires the positions of three (not necessarily different) fragments to place a flexible ligand unambiguously in the binding site. Solvation effects are implicitly accounted for as the binding modes of the fragments are determined with electrostatic solvation in SEED.

2.2. DAIM Program. DAIM is a versatile tool to decompose molecules and prioritize the resulting fragments according to their suitability as anchors for fragment-based docking. During the decomposition, a simple in-house-developed fingerprint is computed for each molecule and fragment. This fingerprint is used to compare molecules or fragments and is the basis for the selection of the fragment triplet needed by FFLD. Moreover, DAIM determines the ligand flexibility for FFLD by identifying the rotatable bonds between fragments.

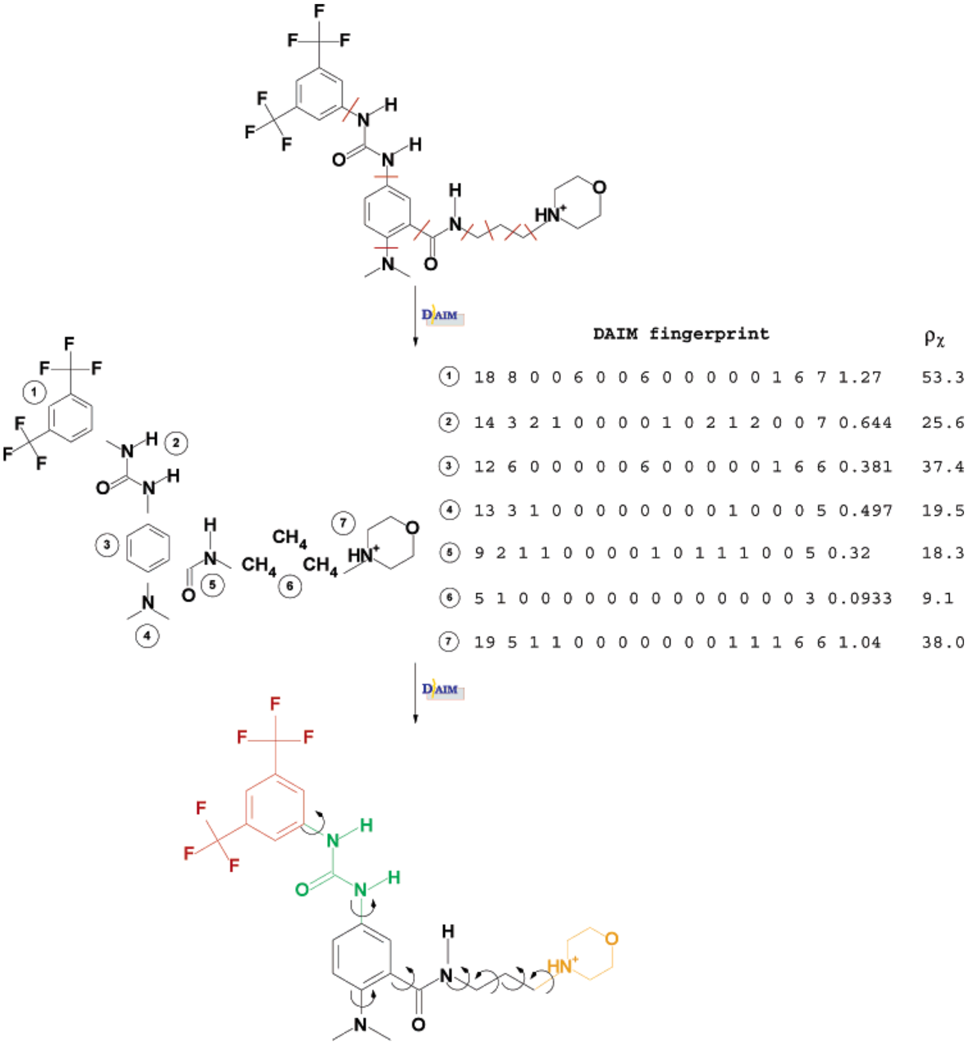
The decomposition of a molecule proceeds in four phases: ring identification, initial fragment definition, functional group merging, and completion of the valences. (i) Rings are identified by successively enumerating all neighbors (i.e., directly covalently bound atoms) of every atom, similar to a breadth-first search. A neighbor with an already assigned number indicates a ring closure, and the corresponding ring size is the sum of the two numbers. (ii) A fragment is defined as a set of atoms connected by unbreakable bonds. The basic definition of unbreakable bonds includes terminal, double, triple, and aromatic bonds and bonds in rings. Nonrotatable and unbreakable bonds are distinguished in DAIM; a nonrotatable bond is always unbreakable, whereas the reverse is not true (e.g., a double bond is nonrotatable and unbreakable, whereas an amide bond is unbreakable but can assume more than one conformation). In this study, an extended definition of unbreakable bonds is used since with the basic definition mentioned above single bonds of groups that form chemical entities would be cut (e.g., in a sulfonamide group, the bond between sulfur and nitrogen is formally a single bond and would thus be cut). This extended list includes amide, phosphate group, and sulfonamide bonds, as well as the single bonds in conjugated systems, and the single bond connecting an amidine group. (iii) To form chemically relevant fragments and avoid the generation of many small groups, small functional groups (e.g., $-OH$, $-CH_3$, $-CX_3$ [where X can be any halogen], $-SO_3$, $-CHO$, $-NO_2$, $-NH_2$, and $-SH$) are merged with the fragment they are connected to. Unbreakable bonds and functional groups (points ii and iii, respectively) can be defined by the user. (iv) In the final step, missing atom neighbors are added. An atom will lack a neighbor atom where the bond connecting them has been cut. These missing neighbors are replaced by hydrogen atoms to reconstitute the correct valence for every atom. A methyl group is used to fill valences where a hydrogen atom would result in an unwanted additional hydrogen bond direction (e.g., a hydrogen replacing a carbon atom bound to an sp^3 nitrogen).



Field	Description
field 1	Total number of atoms
field 2	Number of carbon atoms
field 3	Number of nitrogen atoms
field 4	Number of oxygen atoms
field 5	Number of halogen atoms
field 6	Number of sulfur atoms
field 7	Number of phosphorus atoms
field 8	Number of aromatic bonds
field 9	Number of double bonds
field 10	Number of triple bonds
field 11	Number of amide bonds
field 12	Number of hydrogen bond acceptors
field 13	Number of hydrogen bond donor directions
field 14	Cyclomatic number (ref 50)
field 15	Total number of heavy atoms in rings
field 16	Longest sequence of atoms in the molecule
field 17	Modified Wiener Index 4 divided by 1000 (ref 51)

Figure 2. The DAIM fingerprint for aniline. The Wiener index 4 (field 17, ref 51) has been modified to take into account the covalent radii of the atoms instead of their maximum principal quantum numbers. Furthermore, the modified Wiener Index is scaled by a factor of 1000 so that its weight is of the same order of magnitude as those of the other fields, which is necessary when evaluating the similarity between two molecules.

2.2.1. DAIM Fingerprints. The DAIM fingerprint is a simple structural key and is generated for each fragment obtained by DAIM decomposition. Its main aim is to provide a numerical identifier for a chemical structure to allow fast comparisons of molecules or fragments. The DAIM fingerprint consists of 17 fields, which count atomic and chemical features (Figure 2). It uses only chemical elements and does not require a fragment dictionary, which might have to be updated for every new project. Furthermore, the entries of a fingerprint consisting of chemical element counts can be combined to estimate molecular descriptors, such as the log P (octanol/water partition coefficient), which can be calculated by



	DAIM fingerprint	ρ_χ
1	18 8 0 0 6 0 0 6 0 0 0 0 0 1 6 7 1.27	53.3
2	14 3 2 1 0 0 0 0 1 0 2 1 2 0 0 7 0.644	25.6
3	12 6 0 0 0 0 0 6 0 0 0 0 0 1 6 6 0.381	37.4
4	13 3 1 0 0 0 0 0 0 0 0 1 0 0 0 5 0.497	19.5
5	9 2 1 1 0 0 0 0 1 0 1 1 1 0 0 5 0.32	18.3
6	5 1 0 0 0 0 0 0 0 0 0 0 0 0 0 3 0.0933	9.1
7	19 5 1 1 0 0 0 0 0 0 0 1 1 1 6 6 1.04	38.0

Figure 3. Compound **1** of ref 5 is used as an example of a DAIM decomposition and triplet selection. (Top) Compound **1** is shown with the covalent bonds that are cut by DAIM marked with red lines. (Middle) The fragments identified by DAIM are shown together with their DAIM fingerprints and chemical richness ρ_χ. Note that ρ_χ is evaluated by summing over all values in the fingerprint but neglecting hydrogen atoms or CH_3 groups added by DAIM (e.g., the CH_3 group on the nitrogen of the morpholine in fragment 7). (Bottom) The fragment triplet suggested for docking by DAIM is shown in color. The trisubstituted benzene is considered "central" and is not suggested as anchor (see text). Curly arrows denote rotatable bonds.

atom-additive methods.[32,33] Most importantly, DAIM fingerprints are evaluated rapidly (139 min for the 4.3 million compounds of the November 2005 version of the ZINC database[34] on a single Athlon 2.1 GHz CPU) and do not require the extensive search procedures necessary for hashed fingerprints (e.g., the Daylight fingerprints[35]). Furthermore, as has been shown by Bender and Glen,[36] simple atom count fingerprints perform very well in similarity searches, achieving almost as high enrichment rates as Unity fingerprints.[37] The DAIM fingerprint can be used for assessing the similarity of molecules or fragments thereof, but not directly for searches of arbitrary substructures, because it contains only limited information about connections. It represents a balance between detailed description (i.e., high number of entries) and computational efficiency. The DAIM fingerprints are employed to decompose large libraries into a set of unique fragments and to define three anchor fragments for docking by the SEED[27,28]/FFLD[29,30] procedure (see below).

2.2.2. Selection of Fragments as Anchors. The most suitable anchor fragments for fragment-based docking are those that form highly favorable interactions with the protein upon binding. In other words, these fragments define the strongest constraints for the pose of the ligand. DAIM selects fragments in a three-step selection process. In the first step, the "chemical richness" ρ_χ of a fragment is evaluated by summing over all values in the fingerprint but neglecting hydrogen atoms or CH_3 groups which have been added by DAIM to fill valences. The assumption behind this simple sum is that the fingerprint includes fields that consider both size features and functional groups. The size of a fragment determines in which pockets of a binding site it can fit. Functional groups are likely to form directional interactions and thus determine the orientation of the fragment. The fragments with the largest sum over all entries in the DAIM fingerprint are likely to contain many such functional groups.

All fragments with a value of ρ_χ lower than ten are discarded for reasons of computational efficiency. The value of ten as acceptance threshold was chosen to exclude small apolar fragments such as methane ($\rho_\chi = 9.09$), which is very frequent. It allows, however, the selection of methanol ($\rho_\chi = 14.18$). In the second step, fragments with several substituents are eliminated. These "central" fragments, if highly substituted, will not form significant interactions with the protein for steric reasons. For a cyclic fragment, the number of substituents (n_{subst}) and the number of rings (n_{rings}) are counted, and the cyclic fragment is considered "central" if $n_{\text{subst}} \geq k_r \times (n_{\text{heavy atoms in ring}} - n_{\text{rings}})$. Using a k_r value of $^1/_{1.75}$, a disubstituted benzene is not deselected, whereas a trisubstituted one is considered "central" and, therefore, not used as anchor. An acyclic fragment is deselected if $n_{\text{subst}} \geq k_l \times n_{\text{heavy atoms}}$. The k_l value of 0.5 used in this study permits the selection of terminal amide groups (i.e., connected to one other fragment) but rejects amide groups originating from within the chain (i.e., connected to two fragments). Finally, the fragments are ranked according to their ρ_χ value, and the top three fragments are chosen as anchors. Figure 3 shows a

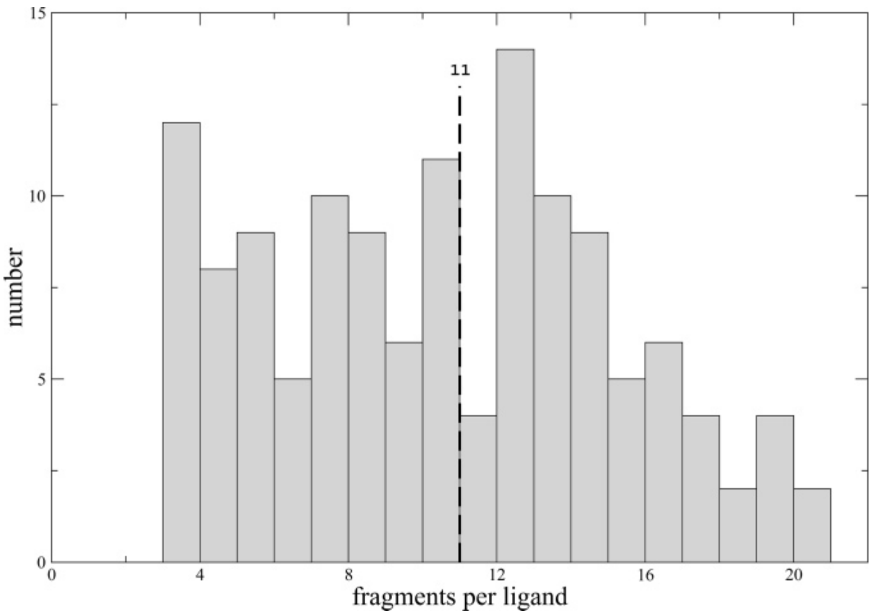


Figure 4. Distribution of the number of fragments per ligand for the 130 test cases. The black dashed line denotes the median value. The average value is 11.1.

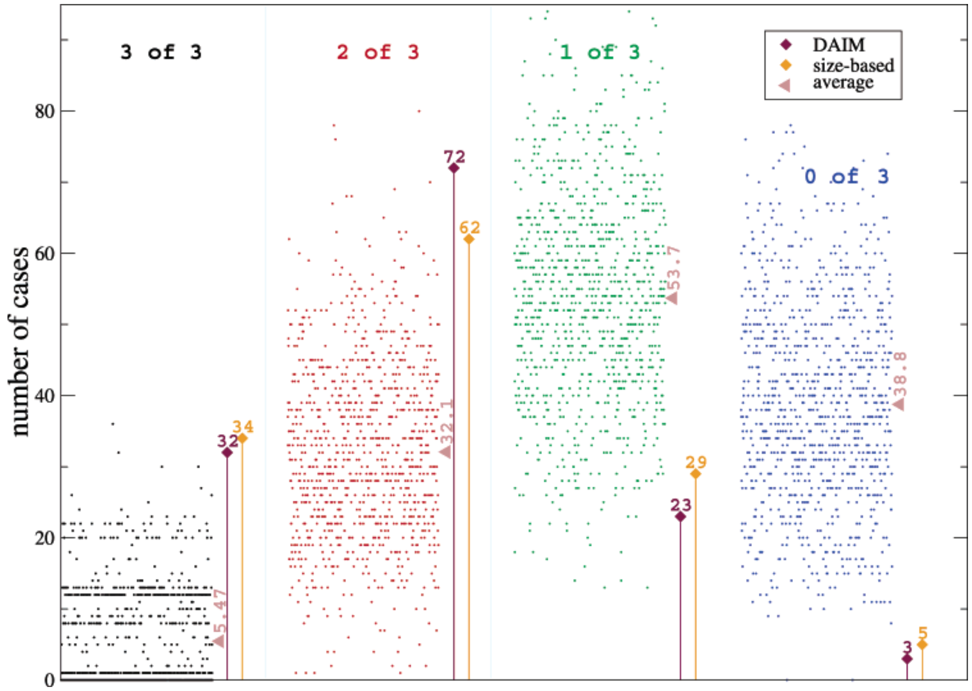


Figure 5. Comparison of the DAIM prediction (purple diamonds), the size-based prediction (orange diamonds), and 1000 random predictions (dots) for the fragments with the highest numbers of contacts. Results are separated into four categories: all correct (i.e, the three fragments suggested by DAIM are the ones with the most contacts in the X-ray structure), two out of three correct, one correct, and none correct. For 32 out of 130 testcases, DAIM was able to correctly predict the three fragments forming the highest number of contacts with the protein. In 72 cases, two out of three were predicted correctly. Brown triangles show the average of the 1000 random runs in every category.

β-secretase inhibitor,[5] its DAIM-defined fragments, and their chemical richness.

2.3. Preparation of the Test Sets. The ligand−protein database (LPDB),[38] a collection of 262 complexes (as of February 2005), was used as test set. Complexes in which the ligand had more than 21 rotatable bonds were not considered, because it had been shown previously that molecules with such a high degree of flexibility cannot readily be treated by the docking approach.[30] The ligands were decomposed into all possible fragments with DAIM. Only ligands with at least four fragments were kept for further analysis, because four is the smallest number of fragments that can be combined to more than one triplet. This left 130 ligand−protein complexes for the contact analysis.

For the docking analysis, the subset of the 130 complexes with ligands of 10 or fewer rotatable bonds was used to evaluate the docking quality, because most compounds in commercial libraries have less than 10 rotatable bonds. This subset initially contained 48 ligand−protein complexes. The number was then reduced to 36 by excluding all test cases in which no triplet yielded a solution with a root-mean-square deviation (rmsd) from the X-ray structure smaller than 2 Å. In both analyses, the results obtained using the DAIM selection of fragments were compared to selections based

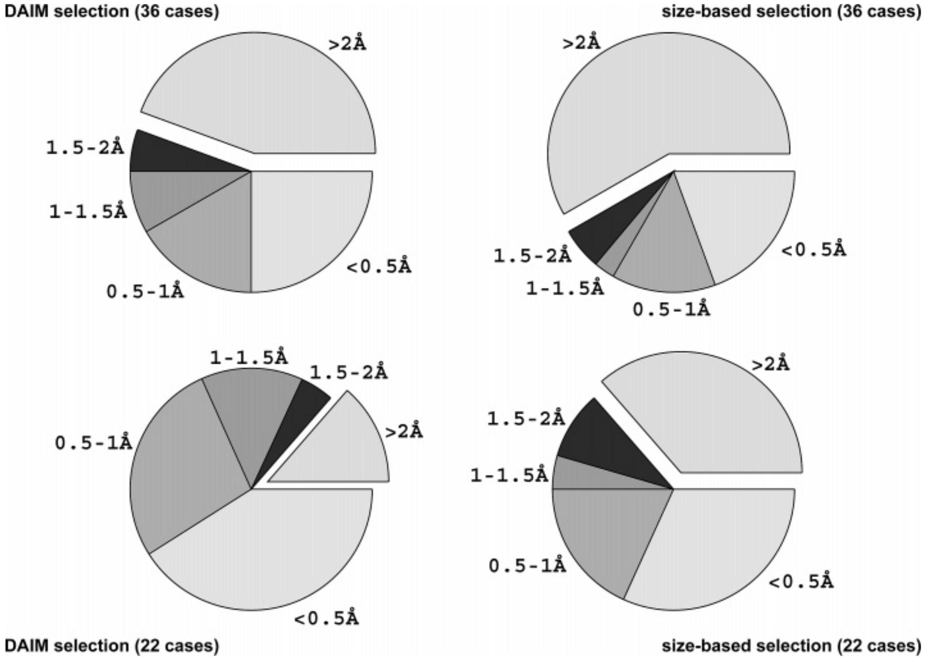


Figure 6. Pie chart of the distribution of the rmsds after docking for the subsets of 36 (top row) and 22 (bottom row) test cases. In the first subset, for the DAIM selection (top left), more than one-third of the docking runs generated at least one pose with rmsd below 1 Å and more than 50% were below 2 Å rmsd from the X-ray structure. For the size-based selection (top right), more than 50% had an rmsd above 2 Å. The advantage of using the DAIM selection is even more evident for the second subset of 22 test cases.

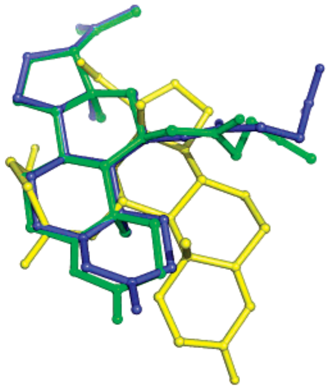


Figure 7. Example of a docking calculation which illustrates the superiority of DAIM selection vs size-based selection of fragment triplets. The ligand is progesterone 11α-hemisuccinate (only heavy atoms are shown), which has seven DAIM fragments, and the protein (not shown) is the Fab' fragment of the DB3 antibody (PDB code 1dbm[44]). Green, the minimized pose of the X-ray structure; blue, the pose closest to the X-ray structure when using the DAIM triplet in the docking calculation (rmsd of 1.06 Å); yellow, the pose closest to the X-ray structure when using the triplet of the size-based selection in the docking calculation (rmsd of 5.46 Å). Figure generated with PyMOL.

either on fragment size alone (taking the number of atoms as criterion) or a random choice of three fragments.

All minimizations were carried out with CHARMM[39] using the CHARMm22[40] force field, applying a distance-dependent dielectric function. The protein was kept rigid in all calculations. Partial charges were computed with the modified partial equalization of orbital electronegativity (MPEOE) method developed by No et al.,[41,42] as implemented in the program Wit!P (A. Widmer, Novartis Pharma AG, Basel, unpublished). In all complexes, the ligands were minimized in the protein binding site. The minimized structures were then used as references in the calculation of the rmsd.

2.4. Assessment of the DAIM Selection of Ligand Fragments in Contact with the Protein.

2.4.1. Computation of Fragment−Protein Contacts. A fragment−protein contact was defined as any heavy atom intermolecular distance shorter than 4.5 Å. For each of the 130 ligand−protein complexes, the three fragments with the highest numbers of contacts were identified. For eight complexes there were four and for one complex there were five “most contacting” fragments because of degeneracy in the number of contacts.

2.4.2 Comparison with Random Results. To ensure that the triplet identified by DAIM is better than chance selection, a comparison set was created by random selection of triplets. This selection was repeated 1000 times with different seed numbers for the random selection process. All random selections were uniformly distributed.

2.5. Assessment of the Usefulness of the DAIM Triplet for Docking. To eliminate conformational bias due to the binding mode,[30] each ligand was first minimized in the isolated state (i.e., outside of the binding site). Then, each ligand was docked by FFLD using as anchor fragments every possible fragment triplet in turn. This required 3 × $\binom{n}{3}$ docking runs for each ligand, where *n* is the number of fragments. The factor of 3 in front of the binomial reflects three independent FFLD calculations with different seeds for the random number generator for each choice of anchor triplet. For each fragment triplet of every compound, the 150 poses with the best FFLD scores (50 poses from each FFLD run) were clustered by using a leader algorithm with a similarity cutoff of 0.7.[27,43] The representative of each cluster was selected for further CHARMM minimization according to the protocol for high-throughput docking.[5,6] Afterward, the rmsds of all minimized poses were calculated with respect to the X-ray pose, which had been minimized previously. Special attention was paid to cases of ligands containing symmetric groups, for which the rmsd was calculated manually. Finally, each of the $\binom{n}{3}$ triplets was assigned the smallest rmsd of all poses resulting from the respective docking as a figure of merit. The rmsd and not the ranking of a pose was used so that the method used for scoring has no influence on the outcome of the tests.

3. Results and Discussion

The usefulness of the DAIM selection of fragments for docking was assessed in two separate tests. The first test concerned the ability of DAIM to identify the ligand fragments

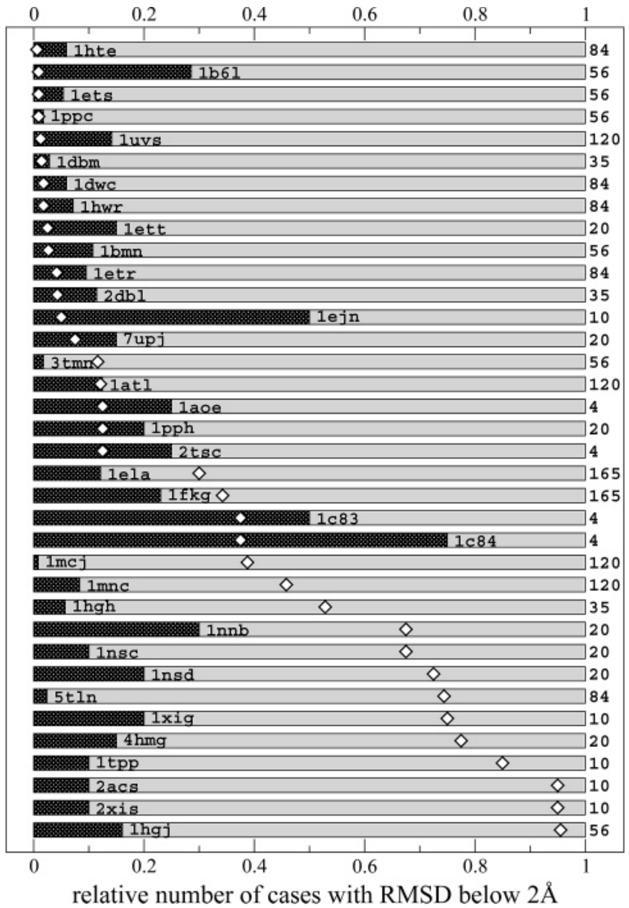


Figure 8. Distributions of the rmsds of the results of the docking calculations when using every possible triplet combination. On the *x*-axis, the number of possible triplets has been normalized to the interval between 0 and 1. The dark gray area of each bar represents the fraction of the triplets whose application as anchors yielded at least one pose with an rmsd below 2 Å. White diamonds display the rank of the calculations using the DAIM triplet. In 20 of 36 cases, the white diamond is in the dark gray area, indicating successful docking (within 2 Å rmsd) using the DAIM selection of the fragment triplet. Numbers to the right of each bar give the total number of possible fragment triplets.

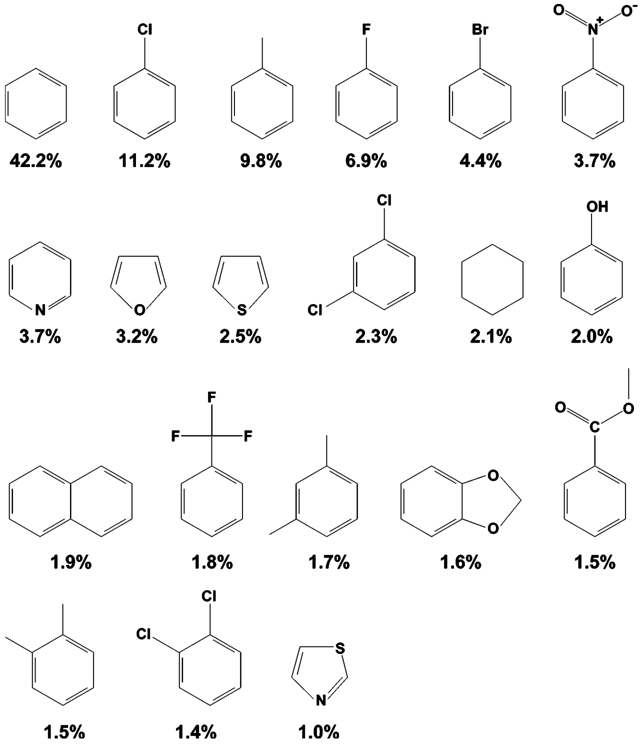


Figure 9. The cyclic fragments that occur in at least 1% of the molecules in the November 2005 version of the ZINC database.

involved in the highest number of contacts with the protein using only the 2D structure of the ligand. In the second test, the docking results obtained with the DAIM selection of three anchor fragments were compared to the docking results obtained using all possible fragment triplets.

The 130 ligands with less than 22 rotatable bonds were decomposed into 1438 fragments, with a maximum number of 21 fragments per ligand and a median of 11 (Figure 4). A total of 109 unique fragments were identified. The three most frequent fragments were methane, formate, and *N*-methylformamide with an occurrence of 697, 301, and 158 times, respectively. Benzene was the most common cyclic fragment, with a frequency of 117.

3.1. Contact Analysis. The fragment triplets chosen by DAIM, using the chemical richness ρ_χ (defined as the sum over the DAIM fingerprint entries, i.e., employing only the 2D structure of the ligand), reproduce the ranking according to the number of contacts in the X-ray structures significantly better than randomly chosen triplets (Figure 5). However, this result is mainly due to the fact that the sum of the DAIM fingerprint entries favors bulky fragments. The selection based solely on the number of atoms in a fragment performs almost as well as the DAIM selection (Figure 5). In fact, the fragment molecular weight correlates with its ρ_χ value, with a correlation coefficient of 0.95 for the 109 unique fragments of the 130 ligands. As a basis of comparison, the correlation between ρ_χ and the sum of the number of hydrogen bond acceptors and donor directions (fields 12 and 13 in Figure 2) is 0.47.

3.2. Docking Analysis. It is necessary to assess the usefulness of the chemical richness ρ_χ as a fragment selection criterion irrespective of possible shortcomings of the other programs and scoring functions used in the fragment-based docking approach. Hence, two subsets of the 48 test cases with less than 11 rotatable bonds were prepared, based on the results of the docking calculations. In the first subset, only the 36 test cases which could be successfully redocked with at least one fragment triplet were considered. This set was defined by excluding the 12 ligand−protein complexes for which no docked structure with an rmsd from the X-ray structure smaller than 2 Å was obtained. Because anchor fragments are used in FFLD, the second subset was defined by excluding ligand−protein complexes where the anchors did not match the positions of the respective fragments in the X-ray structures. Therefore, the second subset was derived from the first subset by considering only the 22 test cases that had at least two anchors within 2 Å of the X-ray positions of at least two fragments of both the DAIM triplet and the size-based triplet. For both subsets, the DAIM triplet selection was again compared to the size-based selection and a random selection. To neglect the inaccuracies of the scoring function used to rank the poses, all poses, and not only the pose with the best energy, were compared to the X-ray structure.

3.2.1. Comparison with the Size-Based Triplet Selection. Figure 6 shows a graphical representation of the rmsds of the poses closest to the X-ray pose, irrespective of their rank in a scoring function. For the first subset, the result is clearly better for the dockings based on the DAIM triplet than for the triplets obtained by choosing the largest fragments (20 and 15 cases out of 36 below 2 Å, respectively). Remarkably, more than one-

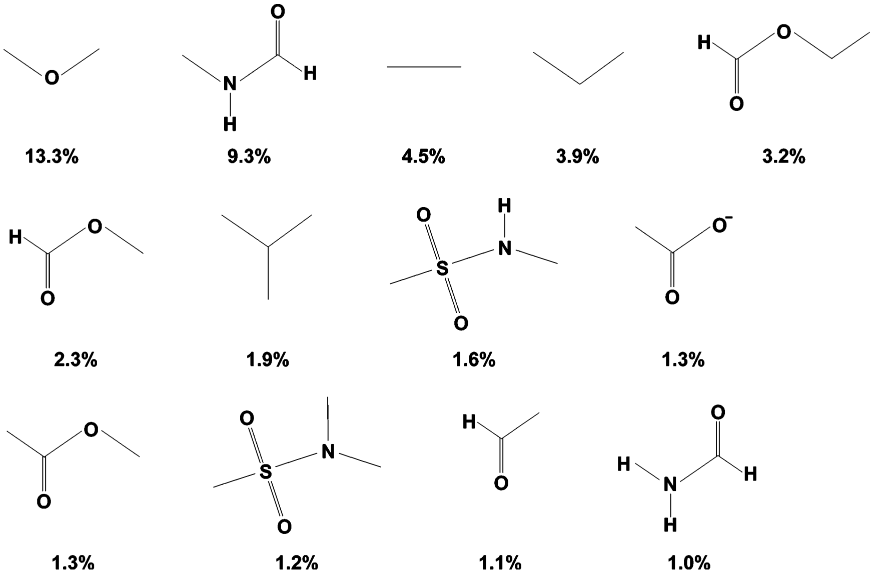


Figure 10. The acyclic fragments that occur in at least 1% of the molecules in the November 2005 version of the ZINC database.

third of the docking runs using the DAIM triplet generated at least one pose with an rmsd from the X-ray structure of less than 1 Å. The superiority of DAIM is even more pronounced when looking at the second subset. Here, docking with the DAIM triplet generated poses within 2 Å of the X-ray structure in 19 out of 22 cases, whereas the size-based selection yielded a pose within 2 Å only in 14 out of 22 cases. As an example, Figure 7 shows a ligand (progesterone 11α-hemisuccinate) that was docked in a pose close to the X-ray structure into the Fab' fragment of the DB3 antibody (PDB code 1dbm[44]) using the triplet suggested by DAIM, but was misplaced when using the triplet suggested by the size-based selection.

3.2.2. Comparison with a Random Triplet Selection. Every possible fragment triplet was used once in a separate docking run. The solutions obtained from these separate docking runs were ranked according to the rmsd from the X-ray structure. Figure 8 depicts the normalized rankings. From top to bottom, each dark gray/light gray bar represents one ligand−protein complex out of the first subset of 36 ligand-protein complexes. The dark gray area of each bar represents the fraction of the total number of fragment triplets whose usage as SEED/FFLD anchors resulted in at least one pose with an rmsd from the X-ray structure below 2 Å. Given these fractions, the expectancy value to obtain results below 2 Å rmsd amounts to 5.87 out of 36 ligand−protein complexes. This value indicates that by selecting a random triplet for each docking calculation, one would obtain on average six cases with at least one pose with an error below 2 Å rmsd from the X-ray structure. Furthermore, by random selection of triplets, the probability to generate poses with an rmsd below 2 Å in 20 or more cases (the number obtained by always using the triplet suggested by DAIM) is less than 0.28%. For the second subset of 22 test cases, chance selection (note that not all triplets might have SEED anchors within 2 Å of the X-ray position) yields an expectancy value of 4.24 and a probability of 1.44% to do equally well or better than the docking using the DAIM triplet. Thus, using the DAIM suggestion of a fragment triplet, poses are generated close to the X-ray structure significantly more often than when using a random triplet. Finally, it has to be emphasized that it is not feasible in high-throughput docking of large libraries to use all fragment triplets for each compound (the numbers to the right of each bar in Figure 8) because the computational cost would increase by a factor of between 1 and 2 orders of magnitude.

3.3. Analysis of Molecular Libraries. DAIM can be employed to automatically analyze the composition and diversity of molecular libraries. As an example, a representative subset of 1.85 million unique compounds of the November 2005 version of the ZINC database[34] was decomposed by DAIM into 186 789 unique fragments (in 29 h on a single Pentium IV 3.2 GHz CPU). It is interesting to compare the DAIM decomposition of the ZINC database with a previous analysis, based on atomic frameworks,[45,46] of a database of 5120 known drugs (which is not publicly available) to assess the relevance of the fragments generated by DAIM. It has to be noted, however, that there are differences between the hierarchical approach used by Bemis and Murcko[45,46] and DAIM. As an example, DAIM does not treat rings connected by a linker as one scaffold ("framework"), but always separates them (e.g., benzylbenzene, the third most frequent framework in known drugs[45] (frequency of 68/5120) is decomposed into two benzene rings by DAIM). For this reason, benzene, which is the most frequent fragment in both databases (Figure 9), has a much larger frequency in ZINC (42.2%) than in known drugs[45] (8.5%). In fact, all molecules with a benzene ring contribute to the benzene count in DAIM. On the other hand, naphthalene and pyridine have comparable frequencies (1.88% and 3.66%, respectively, in ZINC and 0.59% and 0.82% in the known drugs). The main difference between ZINC and the known drugs is the occurrence of aromatic heterocyclic five-rings (Figure 9), of which there is only one appearance among the 41 most frequent frameworks in known drugs[45] (imidazole, frequency of 19/5120). Conversely, no steroid-derived scaffolds are present in ZINC, despite the fact that there are five among the 41 most frequent frameworks in known drugs.[45] The most frequent acyclic fragments identified by DAIM (Figure 10) appear also among the most frequent acyclic substituents ("side chains") in known drugs.[46] This similarity is probably due to the easy synthetic accessibility of certain functional groups.

In spite of the differences between the DAIM decomposition and the approach used in the previous analysis of known drugs,[45,46] it appears that frameworks and side chains in

Table 1. High-Throughput Docking Campaigns in which DAIM was Used To Decompose Libraries of Commercially Available Compounds

target	N_{mol}[a]	N_{frag}[b]	N_{iv}[c]	N_{inh}[d]	ref
β-secretase	306 022	4917	10	1	5
β-secretase	2476	188	20	0	5
β-secretase	391	14	42	1	5
β-secretase	306 022	4917	24	3	6
β-secretase	10 067	469	64	7	6
EphB4[e]	728 202	35 513	30	3	Kolb et al., in preparation

[a] Number of molecules docked by DAIM−SEED−FFLD. [b] Number of unique fragments resulting from the decomposition of N_{mol} molecules. [c] Number of molecules tested in vitro. [d] Number of identified inhibitors with an $IC_{50} \leq 100\ \mu M$. [e] Erythropoietin producing human hepatocellular carcinoma receptor tyrosine kinase B4.

commercially available molecular libraries reflect the chemical features present in drug molecules.

4. Conclusions

DAIM is an automatic and efficient procedure for decomposing molecules and prioritizing the resulting fragments as a first step in fragment-based high-throughput docking. Here, DAIM has been tested on a set of ligands with known binding modes to assess its usefulness for docking. For the prediction of the ligand fragments with the highest number of contacts with the protein, the DAIM rules (i.e., elimination of “central” fragments and ranking according to the chemical richness (ρ_χ)) perform equally well as the prediction based on the fragment size alone. On the other hand, the triplets of fragments suggested by DAIM on the basis of ρ_χ are crucial for the correct placement of a ligand, because the docking solutions obtained by prioritizing fragments with DAIM are more often close to the X-ray structure than the ones obtained with the size-based selection. Because the two sets of 130 triplets overlap only at 80%, it can be concluded that DAIM is able to identify small fragments with a very high contact/size ratio.

To evaluate the usefulness of DAIM for the analysis of large collections of molecules, DAIM was used to decompose 1.85 million compounds from the ZINC database.[34] The decomposition took only 29 h and yielded a list of most frequent cyclic and acyclic fragments very similar to the one obtained by an analysis of atomic frameworks in 5120 known drugs,[45,46] despite the differences in the approach used as well as size and composition of the two libraries. This similarity indicates that most molecules in commercially available libraries contain the scaffolds and functional groups most frequently observed in known drugs.

Recently, we have successfully applied DAIM to decompose compounds from several libraries for a total of more than one million molecules, which has resulted in the discovery of several low molecular weight micromolar inhibitors of β-secretase[5,6] and the receptor tyrosine kinase EphB4 (erythropoietin producing human hepatocellular carcinoma receptor B4, Table 1). It has to be emphasized, however, that the usage of DAIM is not limited to docking together with SEED/FFLD. The ρ_χ-based ranking produced by DAIM can also be employed to select the best anchor fragment for use with other docking programs that require only one anchor (e.g., FlexX[47]).

Other possible applications of DAIM include de novo ligand design and hit improvement. For de novo design, DAIM could be employed to decompose (virtual) libraries into small molecular fragments and use the fingerprint to prioritize them for docking. These databases of fragments could then be used for design by growing or pharmacophore searches or even experimental techniques such as NMR and X-ray crystallography (provided that the fragments are available). An advantage of using DAIM-generated fragments, instead of filtering existing libraries for compounds with very low molecular weight, is that DAIM keeps track of the covalent bonds “cleaved” in the decomposition. This information is useful for estimating the ease of synthesis of de novo designed molecules. Finally, the ability of DAIM to compare fragments via their fingerprints can be exploited for suggesting bioisosteric replacements. In this approach, one or more functional groups of a given inhibitor are replaced by different functionalities of similar size and physicochemical properties but more favorable ADMET (absorption, distribution, metabolism, excretion, and toxicity) properties.[48,49]

Availability of DAIM. The program DAIM as well as its documentation and test cases are available for free to not-for-profit institutions.

Acknowledgment. We are grateful to N. Majeux, F. Dey, D. Huang, M. Cecchini, S. Muff, and A. Cavalli for helpful discussions. S. Audétat is acknowledged for comments on the manuscript. The calculations were performed on Matterhorn, a Beowulf cluster running Linux, and we thank C. Bolliger, Th. Steenbock, and A. Godknecht for setting up and maintaining the cluster. We also thank A. Widmer (Novartis Pharma, Basel) for providing the molecular modeling program Wit!P, which was used for visual and numerical analysis of results and the initial code for ring determination. This work was supported by the Swiss Commission for Technology and Innovation (KTI/CTI).

References

(1) Jorgensen, W. L. The many roles of computation in drug discovery. *Science* **2004**, *303*, 1813−1818.

(2) Desai, P. V.; Patny, A.; Sabnis, Y.; Tekwani, B.; Gut, J.; Rosenthal, P. J.; Srivastava, A.; Avery, M. Identification of novel parasitic cysteine protease inhibitors using virtual screening. 1. The ChemBridge Database. *J. Med. Chem.* **2004**, *47*, 6609−6615.

(3) Desai, P. V.; Patny, A.; Gut, J.; Rosenthal, P. J.; Tekwani, B.; Srivastava, A.; Avery, M. Identification of novel parasitic cysteine protease inhibitors by use of virtual screening. 2. The Available Chemical Directory. *J. Med. Chem.* **2006**, *49*, 1576−1584.

(4) Cozza, G.; Bonvini, P.; Zorzi, E.; Poletto, G.; Pagano, M. A.; Sarno, S.; Donella-Deana, A.; Zagotto, G.; Rosolen, A.; Pinna, L. A.; Meggio, F.; Moro, S. Identification of ellagic acid as potent inhibitor of protein kinase CK2: A successful example of a virtual screening application. *J. Med. Chem.* **2006**, *49*, 2363−2366.

(5) Huang, D.; Lüthi, U.; Kolb, P.; Edler, K.; Cecchini, M.; Audétat, S.; Barberis, A.; Caflisch, A. Discovery of cell-permeable nonpeptide inhibitors of β-secretase. *J. Med. Chem.* **2005**, *48*, 5108−5111.

(6) Huang, D.; Lüthi, U.; Kolb, P.; Cecchini, M.; Barberis, A.; Caflisch, A. In silico discovery of β-secretase inhibitors. *J. Am. Chem. Soc.* **2006**, *128*, 5436−5443.

(7) Kolb, P.; Cecchini, M.; Huang, D.; Caflisch, A. Fragment-Based High-Throughput Docking. *Virtual Screening in Drug Discovery*; CRC Press: Boca Rato, FL, 2005; pp 349−378.

(8) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discovery* **2004**, *3*, 935−949.

(9) Ersmark, K.; Feierberg, I.; Bjelic, S.; Hamelink, E.; Hackett, F.; Blackman, M.; Hulten, J.; Samuelsson, B.; Åqvist, J.; Hallberg, A. Potent inhibitors of the plasmodium falciparum enzymes plasmepsin I and II devoid of cathepsin D inhibitory activity. *J. Med. Chem.* **2004**, *47*, 110−122.

(10) Ersmark, K.; Nervall, M.; Hamelink, E.; Janka, L.; Clemente, J.; Dunn, B.; Blackman, M.; Samuelsson, B.; Åqvist, J.; Hallberg, A. Synthesis of malarial plasmepsin inhibitors and prediction of binding modes by molecular dynamics simulations. *J. Med. Chem.* **2005**, *48*, 6090−6106.

(11) Miranker, A.; Karplus, M. Functionality maps of binding sites: a multiple copy simultaneous search method. *Proteins: Struct., Funct., Genet.* **1991**, *11*, 29−34.

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

(13) Caflisch, A.; Miranker, A.; Karplus, M. Multiple copy simultaneous search and construction of ligands in binding sites: Application to inhibitors of HIV-1 aspartic proteinase. *J. Med. Chem.* **1993**, *36*, 2142–2167.

(14) Caflisch, A. Computational combinatorial ligand design: Application to human α-thrombin. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 372–396.

(15) Caflisch, A.; Wälchli, R.; Ehrhardt, C. Computer–aided design of thrombin inhibitors. *News Physiol. Sci.* **1998**, *13*, 182–189.

(16) Rees, D. C.; Congreve, M.; Murray, C. W.; Carr, R. Fragment-based lead discovery. *Nat. Rev. Drug Discovery* **2004**, *3*, 660–672.

(17) Fattori, D. Molecular recognition: the fragment approach in lead generation. *Drug Discovery Today* **2004**, *9*, 229–238.

(18) Hartshorn, M.; Murray, C.; Cleasby, A.; Frederickson, M.; Tickle, I.; Jhoti, H. Fragment-based lead discovery using X-ray crystallography. *J. Med. Chem.* **2005**, *48*, 403–413.

(19) Shuker, H.; Hajduk, P.; Meadows, R.; Fesik, S. W. Discovering high affinity ligands for proteins: SAR by NMR. *Science* **1996**, *274*, 1531–1534.

(20) Hajduk, P.; Shepperd, G.; Nettesheim, D.; Olejniczak, E.; Shuker, S.; Meadows, R.; Fesik, S. W. Discovery of potent nonpeptide inhibitors of stromelysin using SAR by NMR. *J. Am. Chem. Soc.* **1997**, *119*, 5818–5827.

(21) Gill, A. New lead generation strategies for protein kinase inhibitors - fragment based screening approaches. *Mini-Rev. Med. Chem.* **2004**, *5*, 301–311.

(22) Böhm, H.-J.; Böhringer, M.; Bur, D.; Gmünder, H.; Huber, W.; Klaus, W.; Kostrewa, D.; Kühne, H.; Lübbers, T.; Meunier-Keller, N.; Müller, F. Novel inhibitors of DNA gyrase: 3D structure based biased needle screening, hit validation by biophysical methods, and 3D guided optimization. a promising alternative to random screening. *J. Med. Chem.* **2000**, *43*, 2664–2674.

(23) Furet, P.; Bold, G.; Hofmann, F.; Manley, P.; Meyer, T.; Altmann, K.-H. Identification of a new chemical class of potent angiogenesis inhibitors based on conformational considerations and database searching. *Bioorg. Med. Chem. Lett.* **2003**, *13*, 2967–2971.

(24) Gill, A.; Frederickson, M.; Cleasby, A.; Woodhead, S.; Carr, M.; Woodhead, A.; Walker, M.; Congreve, M.; Devine, L.; Tisi, D.; O'Reilly, M.; Seavers, L.; Davis, D.; Curry, J.; Anthony, R.; Padova, A.; Murray, C.; Carr, R.; Jhoti, H. Identification of novel p38α map kinase inhibitors using fragment-based lead generation. *J. Med. Chem.* **2005**, *48*, 414–426.

(25) Kuntz, I. D.; Chen, K.; Sharp, K. A.; Kollman, P. A. The maximal affinity of ligands. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 9997–10002.

(26) Hann, M. M.; Leach, A. R.; Harper, G. Molecular complexity and its impact on the probability of finding leads for drug discovery. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 856–864.

(27) Majeux, N.; Scarsi, M.; Apostolakis, J.; Ehrhardt, C.; Caflisch, A. Exhaustive docking of molecular fragments with electrostatic solvation. *Proteins: Struct., Funct., Genet.* **1999**, *37*, 88–105.

(28) Majeux, N.; Scarsi, M.; Caflisch, A. Efficient electrostatic solvation model for protein-docking. *Proteins: Struct., Funct., Genet.* **2001**, *42*, 256–268.

(29) Budin, N.; Majeux, N.; Caflisch, A. Fragment-based flexible ligand docking by evolutionary optimization. *Biol. Chem.* **2001**, *382*, 1365–1372.

(30) Cecchini, M.; Kolb, P.; Majeux, N.; Caflisch, A. Automated docking of highly flexible ligands by genetic algorithms: a critical assessment. *J. Comput. Chem.* **2004**, *25*, 412–422.

(31) Scarsi, M.; Apostolakis, J.; Caflisch, A. Continuum electrostatic energies of macromolecules in aqueous solutions. *J. Phys. Chem. A* **1997**, *101*, 8098–8106.

(32) Broto, P.; Moreau, G.; VanDycke, C. Molecular structures–perception, auto-correlation descriptor and SAR studies–system of atomic contributions for the calculation of the normal-octanol water partition-coefficients. *Eur. J. Med. Chem.* **1984**, *19*, 71–78.

(33) Ghose, A. K.; Crippen, G. M. Atomic physicochemical parameters for three-dimensional structure-directed quantitative structure-activity relationships I. partition coefficients as a measure of hydrophobicity. *J. Comput. Chem.* **1986**, *7*, 565–577.

(34) Irwin, J. J.; Shoichet, B. K. ZINC–a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45*, 177–182.

(35) Daylight Chemical Information Systems, Inc. *Fingerprints - Screening and Similarity*. http://www.daylight.com/dayhtml/doc/theory/theory.finger.html.

(36) Bender, A.; Glen, R. C. Discussion of measures of enrichment in virtual screening: Comparing the information content of descriptors with increasing levels of sophistication. *J. Chem. Inf. Model.* **2005**, *45*, 1369–1375.

(37) Hert, J.; Willett, P.; Wilton, D. J. Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177–1185.

(38) Roche, O.; Kiyama, R.; Brooks, C. L., III. Ligand-protein database: Linking protein-ligand complex structures to binding data. *J. Med. Chem.* **2001**, *44*, 3592–3598.

(39) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comp. Chem.* **1983**, *4*, 187–217.

(40) Momany, F. A.; Rone, R. Validation of the general purpose QUANTA 3.2/CHARMm force field. *J. Comput. Chem.* **1992**, *13*, 888–900.

(41) No, K.; Grant, J.; Scheraga, H. Determination of net atomic charges using a modified partial equalization of orbital electronegativity method. 1. Application to neutral molecules as models for polypeptides. *J. Phys. Chem.* **1990**, *94*, 4732–4739.

(42) No, K.; Grant, J.; Jhon, M.; Scheraga, H. Determination of net atomic charges using a modified partial equalization of orbital electronegativity method. 2. Application to ionic and aromatic molecules as models for polypeptides. *J. Phys. Chem.* **1990**, *94*, 4740–4746.

(43) 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.

(44) Arevalo, J. H.; Taussig, M. J.; Wilson, I. A. Molecular basis of crossreactivity and the limits of antibody-antigen complementarity. *Nature* **1993**, *365*, 859–863.

(45) Murcko, M. A.; Bemis, G. W. The properties of known drugs. 1. molecular frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.

(46) Murcko, M. A.; Bemis, G. W. Properties of known drugs. 2. side chains. *J. Med. Chem.* **1999**, *42*, 5095–5099.

(47) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996**, *261*, 470–489.

(48) Patani, G. A.; LaVoie, E. L. Bioisosterism: a rational approach in drug design. *Chem. Rev.* **1996**, *93*, 3147–3176.

(49) Ertl, P. Cheminformatics analysis of organic substituents: identification of the most common substituents, calculation of substituent properties, and automatic identification of drug-like bioisosteric groups. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 374–380.

(50) Euler, L. Elementa doctrinae solidorum. *Novi. Comm. Acad. Sci. Imp. Petropol.* **1752**, *4*, 109–140.

(51) Yang, F.; Wang, Z.-D.; Huang, Y.-P. Modification of the Wiener Index 4. *J. Comput. Chem.* **2004**, *25*, 881–887.

JM060838I