Journal of Computer-Aided Molecular Design, **13:** 167–183, 1999.
KLUWER/ESCOM

Comparison of two implementations of the incremental construction algorithm in flexible docking of thrombin inhibitors

Ronald M.A. Knegtel[a], Denis M. Bayada[a], Richard A. Engh[b], Wolfgang von der Saal[c], Vincent J. van Geerestein[a] & Peter D.J. Grootenhuis[a,*]

[a]*Department of Molecular Design and Informatics, N.V. Organon, P.O. Box 20, 5340 BH Oss, The Netherlands;* [b]*Max-Planck-Institut für Biochemie, D-8152 Martinsried, Germany;* [c]*Boehringer Mannheim GmbH, Sandhofer Strasse 116, D-68305 Mannheim, Germany*

Received 15 May 1998; Accepted 28 August 1998

Key words: flexible docking, incremental construction algorithm, scoring function, structure-based drug design, thrombin

Summary

A set of 32 known thrombin inhibitors representing different chemical classes has been used to evaluate the performance of two implementations of incremental construction algorithms for flexible molecular docking: DOCK 4.0 and FlexX 1.5. Both docking tools are able to dock 10–35% of our test set within 2 Å of their known, bound conformations using default sampling and scoring parameters. Although flexible docking with DOCK or FlexX is not able to reconstruct all native complexes, it does offer a significant improvement over rigid body docking of single, rule-based conformations, which is still often used for docking of large databases. Docking of sets of multiple conformers of each inhibitor, obtained with a novel protocol for diverse conformer generation and selection, yielded results comparable to those obtained by flexible docking. Chemical scoring, which is an empirically modified force field scoring method implemented in DOCK 4.0, outperforms both interaction energy scoring by DOCK and the Böhm scoring function used by FlexX in rigid and flexible docking of thrombin inhibitors. Our results indicate that for reliable docking of flexible ligands the selection of anchor fragments, conformational sampling and currently available scoring methods still require improvement.

Introduction

Automated docking of rigid small molecules to their biomolecular receptors has proven to be a useful tool in the discovery of novel lead compounds of pharmaceutical relevance [1–4]. When large 3D databases of compounds are searched for possible lead compounds, only a limited amount of CPU time can be spent on each molecule. For this reason, most database searches by molecular docking have been performed with a single conformation for each ligand, thus reducing the sampling problem to that of orienting a rigid molecule in a binding site. An interesting extension was recently reported by Lorber and Shoichet [5] in which ensembles of ligand conformers, superimposed on rigid fragments, were used for docking at computational costs comparable to those required for single conformer docking. Several attempts have been made to include ligand flexibility in molecular docking, but until recently none of these approaches could perform a thorough conformational search of the bound ligand in an amount of time suitable for database searching [6–9]. Substantially faster methods, based on the incremental construction algorithm of Leach and Kuntz [10], have recently been reported [11–13] and represent a more time-efficient approach to flexible docking. Currently, the only publicly available software intended for flexible docking of molecule databases are FlexX 1.5 [11] and DOCK 4.0 [14], which are two different implementations of the same incremental construction algorithm.

[*] To whom correspondence should be addressed. Present address: CombiChem Inc., 9050 Camino Santa Fe, San Diego, CA 92121, U.S.A.

A useful molecular docking tool serves two purposes: reconstruction of the correct conformation of the bound ligand and favorable scoring of true ligands with respect to inactive compounds. The incremental construction algorithm attempts to reconstruct the bound ligand conformation by first placing a rigid anchor fragment in the binding site. Additional ligand fragments are subsequently added to complete the ligand structure by applying a greedy algorithm [11]. This algorithm saves only a limited number of the best scoring partial solutions to continue to the next round of ligand reconstruction. Although this algorithm has been shown to be able to reconstruct correct ligand conformations within ∼1–3 min of CPU time [11, 12], there are several conceptual problems that may negatively influence its performance in practical applications. The automatic selection of a suitable, rigid anchor fragment is not a trivial problem and excludes ligands lacking such a substructure [15]. Furthermore, the incremental reconstruction of ligand conformations using the greedy algorithm is prone to the propagation of errors and could miss final conformations that contain suboptimal partial solutions. Finally, both programs use discrete torsion angle libraries to generate conformers, which do not always include dihedral angles observed in experimentally determined conformations of bound ligands [16].

Since the greedy algorithm selects only the best partial solutions to continue to the next round of ligand reconstruction, flexible docking is likely to be more demanding on the quality of the scoring function used to evaluate (partial) docking solutions. A widely used empirical scoring function is the one developed by Böhm [17] and implemented in FlexX [11]. It was derived by fitting experimental binding free energies of 45 protein–ligand complexes to simple functions of hydrogen bond and salt bridge geometry, buried surface area and the number of ligand rotatable bonds as observed in crystal structures of the respective complexes. Although this scoring function has been shown to predict binding constants for some protein–ligand complexes with reasonable accuracy [17], it is not apparent that the functional form of this scoring function is ideal for flexible docking applications. For instance, although the binding affinities for static complexes are reasonably reproduced, it is not evident that the ligand conformation in such complexes represents the global minimum of the scoring function or if multiple, similar or better minima exist which would complicate flexible docking.

In DOCK 4.0 the AMBER [18] inter- and intramolecular force field energy is used to rank putative complexes [19]. This score only represents the interaction energy, neglecting entropic contributions to the binding free energy. In addition, despite the use of distance-dependent dielectrics, the scoring is often still dominated by charge–charge interactions. DOCK 4.0 allows, however, for user-defined scaling of the attractive part of the van der Waals term of different functional groups and this modified force field score is referred to as the chemical score [39]. Chemical types such as hydrophobic, hydrogen bond donor or acceptor etc. are defined on the basis of the Sybyl 6.3 (Tripos Associates, St. Louis, MO, U.S.A.) atom type and the direct neighbors of each atom in the ligand and receptor. Based on this information, each atom is assigned to a single chemical type. In the expression for the van der Waals energy E_{vdW} (Equation 1) between two atoms i and j the attractive term governed by the Lennard-Jones parameter B_{ij} is scaled by a parameter $c_{kl}(i,j)$. This factor denotes the scaling applied to interactions between two chemical types (polar, hydrophobic, etc.) k and l to which atoms i and j have been assigned:

$$E_{\mathrm{vdW}} = \sum_{ij} \frac{A_{ij}}{r_{ij}^{12}} - \frac{c_{kl}(i,j)B_{ij}}{r_{ij}^{6}} \tag{1}$$

This functionality enables the user to make the interaction between groups which are expected to interact unfavorably (for instance, polar–hydrophobic interactions) repulsive in nature, and vice versa. Other empirical scoring functions designed for use in database docking have been described in the literature (for reviews see References 20 and 21), but unfortunately these are not always available to the scientific community for independent evaluation or are not suitable for application in flexible docking of multiple ligands.

The efficacy of recently developed flexible docking tools has mainly been demonstrated by the developers of such programs for the reconstruction of individual protein-ligand crystal structures [11, 12, 22]. The recently reported CASP 2 docking trials [22] involved only a small set of seven single ligand–protein complexes dominated by trypsin inhibitors. For these systems the protein conformation is already maximally adapted to the ligand, which is not the case in most practical applications. FlexX docked two out of seven highest ranking inhibitors within 2 Å of their experimentally determined conformations [16]. The authors argued that the main design objective of FlexX is to

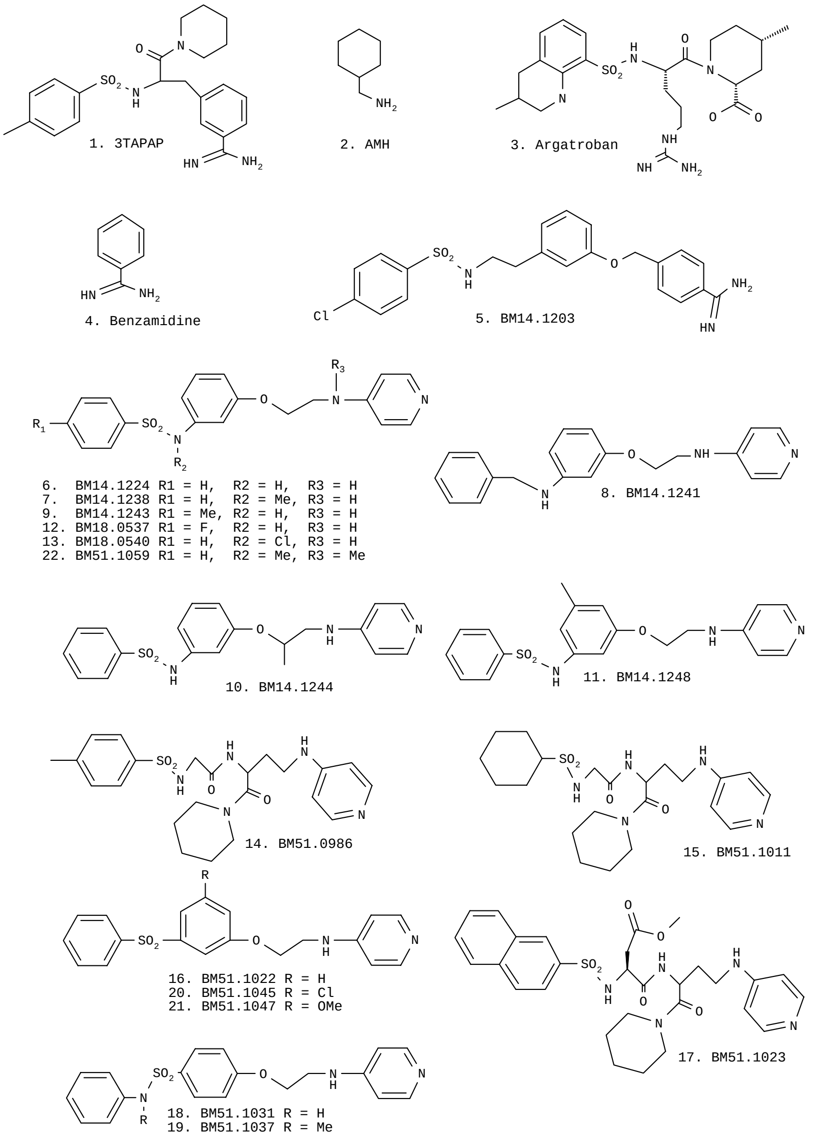

Figure 1. Chemical structures of the 32 thrombin inhibitors used for flexible docking. Inhibitors containing 4-amino pyridine fragments are protonated at the heterocyclic nitrogen atom when bound to thrombin.

generate a limited set of possible bound ligand conformations of which the highest ranking conformer is not necessarily the correct docking. This implies, however, that the selection of the best ligand placement has to be performed by the user and is therefore prone to bias. In addition, if one accepts that the correct solution is not always the highest ranking docking, searching large databases and subsequent visual inspection of multiple solutions for each molecule becomes impractical.

Small database searches using a modified version of DOCK 4.0 with limited back-tracking have been published for dehydrofolate reductase (DHFR) [12]. It is not evident, however, that methotrexate analogues provide a typical example of a flexible ligand since most of the specific interaction with DHFR is determined by a single, rigid ring system. A flexible docking program can be expected to dock a chemically wide variety of ligands and not all active compounds may contain a suitable ring fragment to be used as an initial anchor. Besides the chemical nature of the ligands, there are additional factors that may be important in flexible docking such as the sensitivity to

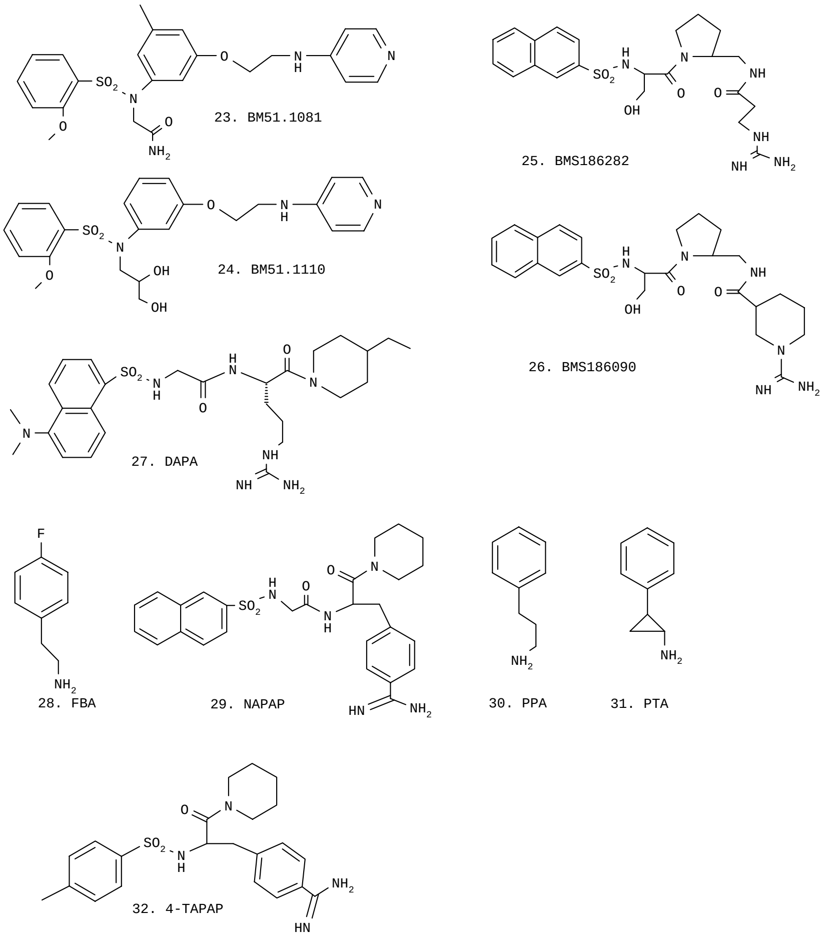

Figure 1. (continued).

the amount of conformational sampling, scoring and the quality of the receptor structure.

In order to evaluate and compare the performance of different implementations of the incremental construction algorithm and scoring functions, we have used FlexX 1.5 and DOCK 4.0 to dock a small database of 32 thrombin inhibitors, for which the crystal structures of the corresponding complexes with thrombin or trypsin have been determined, to two different thrombin crystal structures. Thrombin is a well-documented and realistic test-case (also see Reference 23) with a range of known inhibitors of different sizes, degrees of flexibility, possible anchor fragments and known bound conformations. In order to put the results obtained with flexible docking into perspective, they are compared with rigid body dockings of single, rule-based ligand conformers (which is still often used for database docking) and docking of a set of diverse conformers for each inhibitor. For the generation and selection of sets of diverse conformers, new methodology integrating bond-path weighted conformational sampling, 3D molecular descriptors and Kohonen mapping was applied. In addition, the influence of different receptor conformations at low and high resolution, increased sampling and the scoring method on the accuracy of reconstructed thrombin–inhibitor complexes is evaluated. As a measure of the quality of a docking solution, the root mean square deviation (rmsd) of the docked inhibitor from its conformation in the crystal structure of the thrombin–ligand complex is used. Although there is no accurate definition of what values for the rmsd still correspond to a ‘correct’ docking, due to the dependence of the rmsd on ligand size and flexibility, an upper limit threshold of 2 Å was chosen to determine the number of correctly docked conformations. Similar thresholds of 2–3 Å rmsd have been used previously to interpret docking results [22].

Methods

Thrombin and inhibitor structures

Coordinates of thrombin and inhibitors were obtained from the Brookhaven Protein Databank PDB [24] or from our unpublished database of thrombin–inhibitor complex structures determined during the course of our anti-thrombosis programme [25, 26], and are listed in Table 1. Inhibitor structures are depicted in Figure 1. Inhibitors 2, 3, 27, 28, and 30 do not contain a rigid ring system that can be placed as an anchor fragment in the specificity pocket of thrombin. These molecules therefore present a difficult case for the incremental construction algorithm. Thrombin structures used in this work were 1ETS [27] and 1DWC [28], which are complexes of bovine ϵ-thrombin with NAPAP and human α-thrombin with argatroban at high (2.3 Å) and low (3 Å) resolution, respectively. Both proteins have identical active site residues. Due to the difference in resolution, bound ligand and protein source, small changes in side chain conformations are observed in the active site. The 1PPH structure of trypsin [29] was also used for docking to investigate the influence of the presence of the insertion loop in thrombin, which is absent in trypsin.

Ligand conformations were either those as in the original crystal structures or conjugate-gradient energy minimized for 200 steps without applying electrostatics. For minimization the Batchmin script supplied with Sybyl 6.3 was used to regularize their geometries. In addition, rule-based ligand conformations were generated with CORINA 2.0 which in our experience performs comparably to CONCORD [31]. All flexible dockings were performed with the energy-minimized inhibitors since these are expected to more accurately represent the standard geometries used to construct 3D molecular databases.

Atom types, protons and charges were added manually to ligand structures using Sybyl since automatic processing with Sybyl and DOCK yielded unprotonated 4-amino pyridines. For DOCK, Mol2 files with Gasteiger–Marsili charges [32] were generated while for FlexX partial charges were only set for charged hetero atoms. All calculations were performed on an SGI Power Challenge equipped with R10000 processors running at 200 MHz.

Diverse conformers of all inhibitors were generated by the CONF program which was developed in-house. CONF generates diverse conformers of a given compound in three main steps. The first step consists of finding the rotatable bonds in a molecule stored in Sybyl Mol2 or MDL SD format and defining how important rotating around these bonds is for generating diverse conformers. The second step is the actual generation of the conformers and the third step consists of the selection of a diverse subset.

CONF defines a rotatable bond as a bond that is not in a ring, is a single bond and is not terminal. We also exclude bonds linked to a terminal symmetrical group such as the C-X bond in C-XY_n, where XY_n can be, for instance, CO_2 or SO_3. Weighting of the rotatable bonds is done as follows. Consider a rotatable bond connecting two atoms labeled 1 and 2. The length l_1 (respectively l_2) of the longest of the shortest paths [33] starting from 1 (respectively 2) not going through the central bond is computed. Then the weight assigned to the central bond is the smaller of l_1 and l_2. This is illustrated in Figure 2. For bond A connecting atoms 8 and 9, the length of the longest path starting from atom 8 without passing through bond A is 3, i.e. paths 8-3-2-1, 8-3-4-5 and 8-7-6-5. For atom 9, the path is 9-10-11-12-13 which is of length 4. Thus, the weight given to rotatable bond A is 3. This weight is an estimation of the relative influence of rotating around this bond on the entire ligand conformation. It is assumed that when rotating around a bond, the larger part of the molecule is not rotating. If the bond is terminal, then rotating around it will have no effect, whereas rotating around a bond situated more centrally in a molecule could produce dramatic changes. This can be seen in Figure 2, where rotation around bond B is intuitively less ‘important’ than rotation around bond A.

Two parameters (F and MaxConf) define the depth of the conformational search. F is a tunable scaling factor used to adjust the number of rotations per bond and MaxConf is the maximal number of conformers requested by the user. The number of rotations (including the initial random rotation) around a given rotatable bond i of weight $W(i)$ is equal to $\mathrm{nbRot}(i) = F \times W(i) + 1$. Thus, the total number of conformers (TNC) generated is

$$\mathrm{TNC} = \sum_{\text{all rotatable bonds}} (\mathrm{nbRot}(i))$$

The value of F is reduced as long as the value of TNC is larger than MaxConf. For a bond i, the sampling angle will be $2\pi/\mathrm{nbRot}(i)$ rad. All the cosines and sines necessary to perform the rotations are then precomputed and stored. An initial random rotation is applied around each rotatable bond and all rotations are made

Table 1. Thrombin, trypsin and inhibitor crystal structures used for flexible docking. Protein sources, inhibitor codes, resolutions and PDB codes are listed. A 'UNP' PDB code indicates that this structure was taken from the Organon-Boehringer Mannheim unpublished database of thrombin/inhibitor complex structures

No.	Protein	Inhibitor	Resolution (Å)	PDB code
1	Bovine β-trypsin	3-TAPAP	1.9	1PPH
2	Bovine β-trypsin	AMH	1.8	1TNG
3	Human α-thrombin	Argatroban	3.0	1DWC
4	Bovine β-trypsin	Benzamidine	2.2	1DWB
5	Bovine α-thrombin	BM14.1203	2.9	UNP
6	Bovine α-thrombin	BM14.1224	2.2	UNP
7	Bovine α-thrombin	BM14.1238	2.7	UNP
8	Bovine α-thrombin	BM14.1241	2.5	UNP
9	Bovine α-thrombin	BM14.1243	2.5	UNP
10	Bovine α-thrombin	BM14.1244	2.4	UNP
11	Bovine α-thrombin	BM14.1248	2.5	1UVT
12	Bovine α-thrombin	BM18.0537	2.9	UNP
13	Bovine α-thrombin	BM18.0540	3.2	UNP
14	Bovine α-thrombin	BM51.0986	2.8	UNP
15	Bovine α-thrombin	BM51.1011	2.8	1UVS
16	Bovine α-thrombin	BM51.1022	2.5	UNP
17	Bovine α-thrombin	BM51.1023	2.8	UNP
18	Bovine α-thrombin	BM51.1031	2.5	UNP
19	Human α-thrombin	BM51.1037	2.4	UNP
20	Bovine α-thrombin	BM51.1045	3.0	UNP
21	Bovine α-thrombin	BM51.1047	2.5	UNP
22	Bovine α-thrombin	BM51.1059	2.9	UNP
23	Bovine α-thrombin	BM51.1081	2.5	UNP
24	Bovine α-thrombin	BM51.1110	2.6	UNP
25	Human α-thrombin	BMS186282	2.6	1BMM
26	Human α-thrombin	BMS186090	2.8	1BMN
27	Human α-thrombin	DAPA	2.3	UNP
28	Bovine β-trypsin	FBA	1.8	1TNH
29	Human α-thrombin	NAPAP	3.0	1DWD
30	Bovine β-trypsin	PPA	1.8	1TNK
31	Bovine β-trypsin	PTA	1.8	1TNL
32	Bovine ϵ-thrombin	4-TAPAP	2.5	1ETT

for each bond in a depth-first search fashion. Each time a final leaf of the depth-first search tree is reached, the conformer is checked for van der Waals clashes and written to a file if it passes this check. The whole process of conformer generation took approximately 1 h of CPU time and generated in all around three and a half million conformers for the 32 inhibitors (see Table 2). After elimination of conformers with van der Waals clashes, 614 353 conformers remained.

To select the most diverse conformers, we have used a selection technique that has the advantage of being relatively fast. First, a set of 38 3D descriptors is generated for each conformer (see the Appendix and Reference 34). The generation of the descriptors for the 32 sets of conformers (~600 000) took less than 8 h of CPU time on an SGI R10000. For each set of conformers, a Kohonen network [35] was trained using the descriptors. The size of this network was $25 \times 40 = 1000$ neurons. Once trained, the conformers closest to the neuron's weights, for non-empty neurons, were selected. This yielded a total of 1000 selected conformers or less per compound and a total

Table 2. Overview of ligand parameters and generated conformers for the construction of a flexibase of 32 thrombin inhibitors

Name	Number of rotatable bonds	Max. number of conformers	Final number of conformers	Number of selected conformers
3-TAPAP	7	126000	20496	998
AMH	1	21	21	21
Argatroban	9	153600	6071	912
Benzamidine	0	1	1	1
BM14.1203	8	90000	38078	998
BM14.1224	8	69120	35266	1000
BM14.1238	8	96000	44127	1000
BM14.1241	8	69120	40026	997
BM14.1243	8	69120	11391	942
BM14.1244	8	129600	32922	995
BM14.1248	8	69120	21890	986
BM18.0537	8	69120	37442	983
BM18.0540	8	69120	18550	982
BM51.0986	12	93312	38750	996
BM51.1011	12	82944	13459	876
BM51.1022	7	53760	23488	992
BM51.1023	15	746496	14172	978
BM51.1031	8	69120	36725	999
BM51.1037	8	69120	10446	950
BM51.1045	7	53760	24647	933
BM51.1047	8	51840	19402	961
BM51.1059	8	129600	12221	943
BM51.1081	11	259200	20155	988
BM51.1110	12	20736	1114	482
BMS186282	13	46656	743	416
BMS189090	10	172800	13019	979
DAPA	10	288000	39692	1000
FBA	1	21	21	21
NAPAP	10	124416	22438	992
PPA	3	315	271	196
PTA	1	31	31	31
4-TAPAP	7	129600	17278	975
Total	252	3401669	614353	25523

of 25 523 conformers for the entire set of inhibitors (see Table 2). The selection process did not take more than a few minutes on an SGI R10000.

For the selection of diverse conformers for our flexibase, most of the 3D descriptors available from an in-house program for the generation of molecular descriptors were used. In the future, a more reasonable number of descriptors should be chosen and the quality of these descriptors assessed for conformer diversity analysis. In addition, the selection of approximately 1000 diverse conformers for most of the ligands is probably too ambitious and further work is necessary to reduce this number. Both adaptations are expected to further reduce the amount of time required for the generation of diverse flexibases.

For DOCK 4.0, protein coordinates and charges were generated in Sybyl Mol2 format and processed with CHEMGRID [19] to yield energy and chemical scoring grids using a 4r distance-dependent dielectric (see References 36 and 37) with a 10 Å cut-off, a

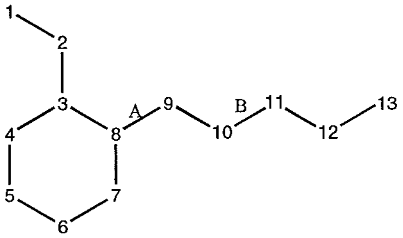

Figure 2. Example of the determination of weights for the sampling of torsion angles around rotatable bonds, as used in the generation of diverse conformers by CONF. Atoms in the molecule are indicated by numbers and two rotatable bonds have been indicated with A and B. After identifying the longest of the shortest paths starting from atoms 8 and 9, not going through A, the weight for rotating around bond A is taken as the smallest of the lengths of both paths. For bonds A and B this analysis yields weights of 3 and 2, respectively. The lower weight for bond B agrees with the assumption that rotation around bond B, while keeping the larger part of the molecule fixed, is expected to affect the overall conformation of the molecule less than rotation around A.

grid spacing of 0.3 Å and a united atom model. For FlexX, protein structures were translated to FlexX receptor description files, except for 1DWC for which an example file was already provided with the package.

Rmsd values with respect to the experimentally determined binding modes were calculated by summing over identical non-hydrogen atoms. This increases the rmsd for cases where symmetrically substituted rings or functional groups are flipped with respect to the native binding mode. The effect will be diluted, however, when such groups are part of larger, non-symmetric inhibitors that are not perfectly superimposed on the conformation observed in the crystal. For benzamidine, for instance, the rmsd between the native orientation and one that is rotated by 180° around its symmetry axis is 1.96 Å.

Flexible docking with DOCK 4.0

The majority of the parameters set for database docking with DOCK 4.0 were taken from the example files provided with the package. Using DOCK 4.0 in database mode implies that intra-molecular non-bonded energies are included in the total score used to rank different conformations of one particular ligand but not for the mutual ranking of all ligands in a database.

All rigid body dockings were performed with DOCK 4.0 since FlexX 1.5 does not contain functionality to this purpose. Sampling of atom-sphere matches in DOCK 4.0 was performed using uniform sampling of 500 conformations as described by Ewing and Kuntz [14]. For the rigid docking of a large database of random conformers, this number was reduced to 50. In the case of flexible docking, the 10 best scoring partial solutions proceeded to the next round of ligand reconstruction (which will be referred to as standard sampling). For increased sampling the last parameter was set to 250. A maximum of three bumps between ligand and receptor was accepted. Flexible docking using standard sampling and a maximum of two bumps yielded similar results for chemical scoring, but decreased the number of docking solutions within 0–4 Å rmsd with four when energy scoring was applied (results not shown). The minimum anchor size for automatic anchor fragment selection was set to four atoms in order to include the guanidino group of arginine derivatives.

A total of 10–11 spheres were generated with SPHGEN [38] and placed only in the P1 binding pocket of thrombin, in order to focus sampling of the anchor placement at the specificity pocket. Similar docking results were obtained when spheres filling the entire binding site were used (results not shown), albeit at the cost of a small increase in the CPU time required. In all cases, dihedral and rigid body minimization was applied with a maximum of 100 cycles and convergence at 0.1 score units. Because the simplex minimizer in DOCK 4.0 yielded very different solutions depending on the random number seed used to initialize it, the DOCK source code was modified such that the same random number generator seed was used for each ligand in our database. Using the unmodified DOCK 4.0 code on input databases with the inhibitors sorted in different order yielded average deviations in the order of 2 kcal/mol in score and 1.4 Å in rmsd (results not shown).

For chemical scoring, the definitions of polar atoms (hydroxyls and fluorine), hydrogen bond donors, hydrogen bond acceptors, hydrophobic atoms and all remaining Sybyl atom types (named ‘null’) were used as provided with DOCK 4.0 [39]. Standard weights for the attractive portion of the van der Waals interaction are listed in Table 3 and were used for all dockings unless stated differently. When all weights in Table 3 are set to 1, the standard DOCK force field energy is regained.

Flexible docking with FlexX 1.5

For flexible docking with FlexX 1.5, default settings were used as provided with the package. The example receptor definition file for the 1DWC thrombin structure was used as a template to generate the 1ETS receptor definition file. Normal sampling was done

Table 3. Default scaling factors for the attractive part of the van der Waals interaction between different chemical types as used in chemical scoring of DOCK 4.0

Chemical type	Null	Hydrophobic	Donor	Acceptor	Polar
Null	0.5				
Hydrophobic	0.5	1			
Donor	0.5	0	0		
Acceptor	0.5	0	1	0	
Polar	0.5	0	1	1	1

with 100 solutions per ligand and 400 solutions per partial solution. For increased sampling both values were increased to 250 and 1000, respectively. Docking was done using the 'AUTODOCK' command which automatically docks a flexible ligand to a receptor site and is suitable for batch processing of a database of molecules. This command implies automatic selection of suitable anchor fragments [15]. Torsional minimization was applied in all cases and the best scoring solutions were saved as Sybyl Mol2 files.

Results

Rigid body docking of thrombin inhibitors

Rigid body docking of the bound conformations of our set of thrombin inhibitors provides a point of reference for the evaluation of flexible docking. Besides the bound ligand conformations, a gliding scale of ligand flexibility was introduced by also docking single, rule-based ligand conformations generated with CORINA 2.0 and a database of multiple, diverse conformers (i.e. a flexibase [40]) all docked as rigid bodies [41]. With approximately 1000 conformations per molecule, the size of our flexibase is expected to be close to the maximally feasible in practical applications. An overview of the numbers of ligand conformations considered is listed in Table 2. It is expected that the use of random conformers is less efficient than protocols such as the incremental construction algorithm, which include knowledge of the receptor in generating ligand conformers. Rigid docking of such a large database requires, despite the use of reduced orientational sampling, approximately 10 times more CPU time than flexible docking of the same set of inhibitors.

The resulting frequency distributions of the rmsd values of the rigidly docked bound, and CORINA-generated conformers are depicted in Figures 3A and B. For DOCK 4.0, the results are presented obtained with force field scoring as well as chemical scoring. As expected, docking of the rigid crystal structure conformations of thrombin inhibitors yields correct solutions in most instances. Docking solutions with rmsd values higher than 2 Å are mostly the smaller inhibitors (benzamidine, PPA, PTA). The ligand conformations generated by CORINA are usually extended and yield, as expected, only a few correct dockings for the smallest inhibitors. Energy scoring produces slightly more low rmsd dockings, probably due to the more accurate representation of attractive van der Waals interactions compared to chemical scoring.

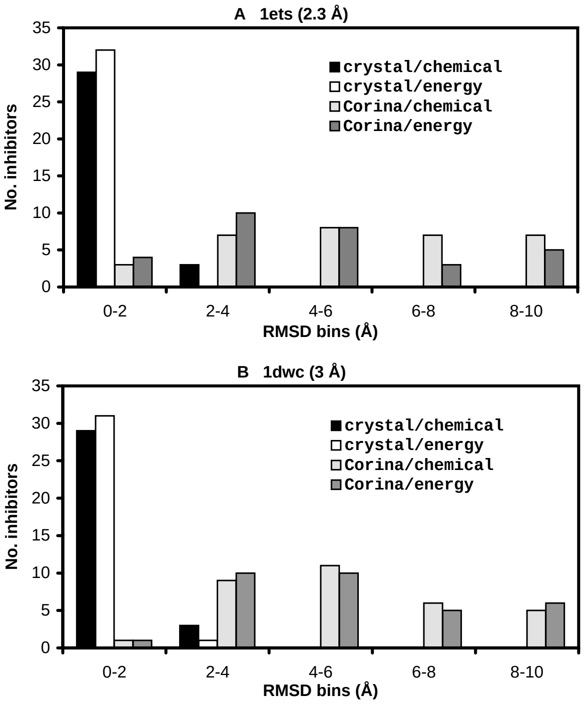

Figure 3. Frequency distributions of the rmsd values with respect to the bound conformations of 32 rigidly docked thrombin inhibitors. Results are shown for rigid body dockings with DOCK 4.0 of the bound inhibitor conformations and conformers generated by CORINA both scored with chemical and energy scoring. All compounds were docked against two thrombin crystal structures determined at 2.3 and 3 Å resolution, respectively, and indicated by their PDB codes 1ETS (A) and 1DWC (B).

Docking of a flexibase of diverse conformers is compared to that using the CORINA-generated ligand structures in Figures 4A and B. The use of multiple conformations clearly yields an improvement over using only a single conformation. However, even though a relatively large number of conformers (~1000) are docked per inhibitor, only 5–7 of the highest ranking

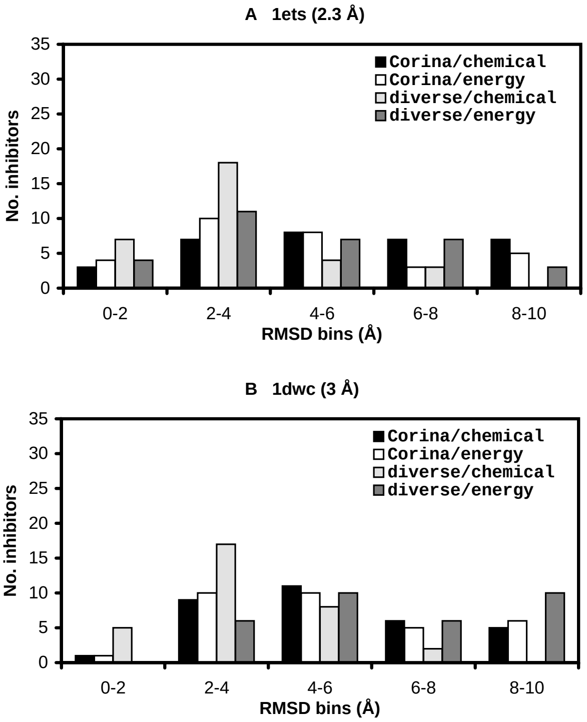

Figure 4. Frequency distributions of the rmsd values with respect to the bound conformations of 32 rigidly docked thrombin inhibitors. Results are shown for rigid body dockings with DOCK 4.0 of single conformers generated by CORINA and a flexibase containing 25 523 conformations of 32 inhibitors both ranked with chemical and energy scoring. All compounds were docked against two thrombin crystal structures determined at 2.3 and 3 Å resolution, respectively, and indicated by their PDB codes 1ETS (A) and 1DWC (B).

conformers are docked within 2 Å of the experimentally determined conformation. Although the flexibase contained several conformers within 2 Å of each bound conformation, small deviations from the true, bound conformations can still cause severe van der Waals clashes when they are placed in the active site. Interestingly, chemical scoring selects more correct conformations from the flexibase than energy scoring, suggesting that chemical scoring is preferred when ligand flexibility is introduced in molecular docking.

Flexible docking with standard sampling

In Figures 5A and B the frequency distributions of the rmsd values of 32 flexibly docked thrombin inhibitors with respect to their crystal structure conformation are depicted. For DOCK 4.0, the results are presented obtained with force field scoring as well as chemical scoring. It is clear from Figure 5 that only a fraction of the inhibitors can be docked with rmsd values smaller than 4 Å and even less with rmsd values below 2 Å (10–35%). Chemical scoring yields again more solutions with rmsd values below 4 Å than force field scoring and the Böhm scoring used by FlexX. Interestingly, the number of docked inhibitors with rmsd values below 2 Å obtained with FlexX seems comparable to that obtained with DOCK 4.0 using force field scoring or docking of rigid CORINA conformers, even though the empirical scoring function of FlexX is expected to offer a more accurate description of the binding free energy. In this respect, it is noteworthy that chemical scoring as implemented in DOCK 4.0 again yields better results than DOCK force field scoring and FlexX, even though it is a fairly crude, intuitive scoring method.

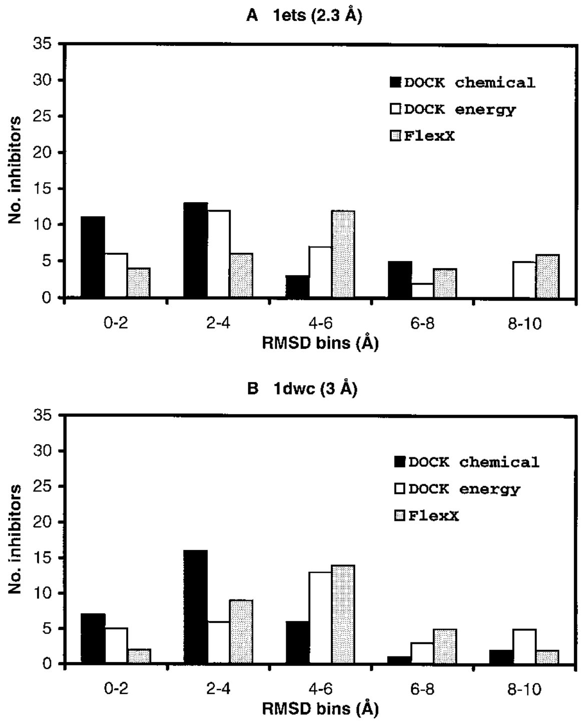

Figure 5. Frequency distributions of the rmsd values with respect to the bound conformations of 32 flexibly docked thrombin inhibitors. Results are shown for flexible dockings with DOCK 4.0 using chemical and energy scoring and FlexX 1.5. All compounds were docked against two thrombin crystal structures determined at 2 and 3 Å resolution, respectively, and indicated by their PDB codes 1ETS (A) and 1DWC (B).

Analysis of the van der Waals and electrostatic contributions to the total DOCK scores showed no apparent correlations with rmsd. The van der Waals contribution to the chemical score is reduced to an average of −10 ± 5 kcal/mol with respect to its contribution to the energy score (−35 ± 10 kcal/mol)

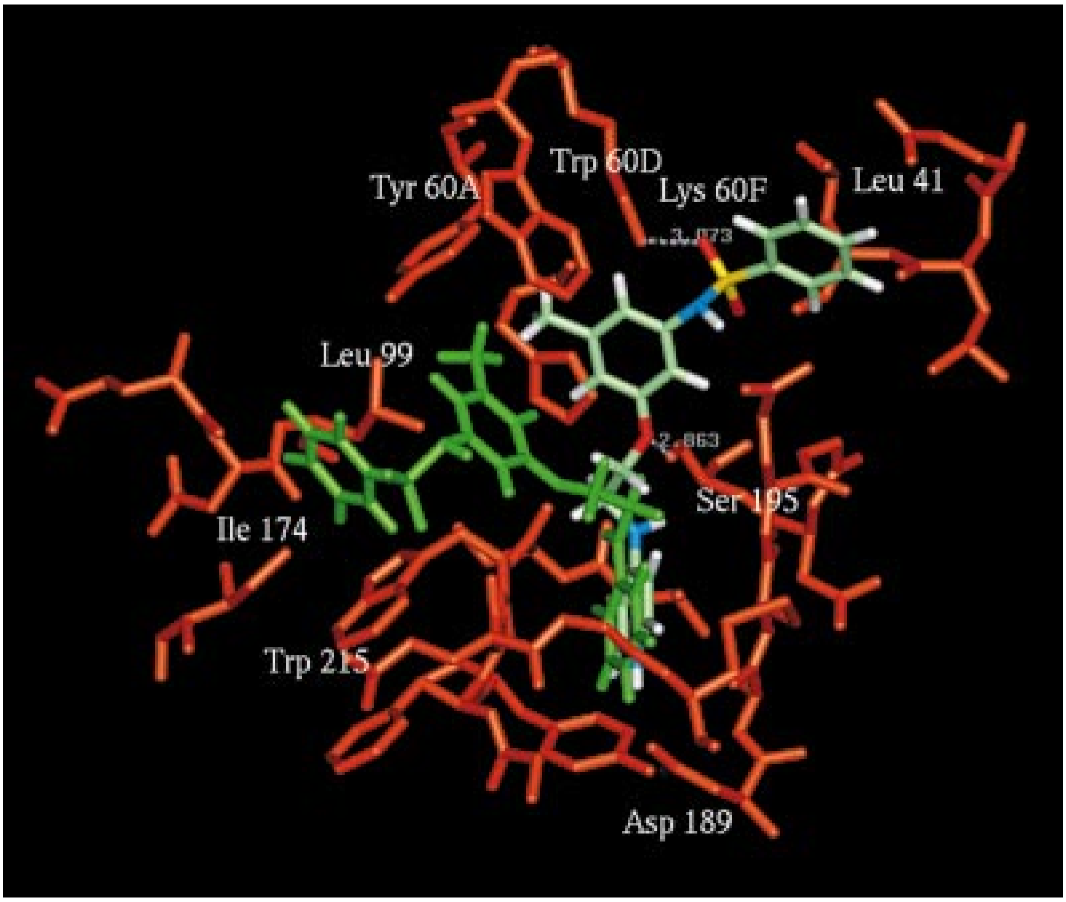

Figure 6. Example of an incorrectly docked thrombin inhibitor (BM14.1248; rmsd 9.1 Å) placed in the active site by DOCK 4.0 using chemical scoring. The experimentally determined bound conformation is shown in green. The incorrect docking solution shows reasonable interaction with the enzyme, including a hydrogen bond of Lys^{60F} with the sulfonamide moiety of the inhibitor and favorable hydrophobic interactions between Leu^{41} and the terminal phenyl group.

while the electrostatic contributions are almost identical (-12 ± 2 kcal/mol). Judging from the performance of chemical scoring in flexible docking, discarding or penalizing van der Waals interactions between non-complementary functional groups apparently models the thermodynamic costs of bringing, for instance, polar surfaces of the inhibitor in contact with hydrophobic surfaces in the active site. Using unmodified force field scoring, such unfavorable interactions would always be rewarded with favorable van der Waals interactions. For the limited number of compounds that are docked within 2 Å rmsd of their known binding modes, correlation coefficients in the order of 0.8 were found when their DOCK scores were correlated with measured K_i's (results not shown). Such correlations are similar to those reported elsewhere for force-field-based scoring of thrombin inhibitors (see References 36 and 37).

In order to investigate whether the charged nature of the inhibitors affects the docking outcomes, flexible dockings were performed with DOCK 4.0 using contact scoring and chemical/energy scoring without electrostatics. In our hands, contact scoring did not produce realistic dockings, with the inhibitors penetrating deep into the thrombin P1 pocket while making numerous close contacts. As expected, docking without electrostatics produces more high rmsd dockings, although still 11 and 3 out of 32 inhibitors were docked within 2 Å rmsd using chemical and energy scoring, respectively. Apparently, the scaling of non-complementary van der Waals interactions as used with chemical scoring still produces better dockings than plain energy scoring when electrostatics are omitted.

Visual inspection of some of the incorrectly docked inhibitors showed that in most instances the anchor fragment is placed correctly, but that the remaining part of the inhibitor structure was incorrect. Many inhibitors were not directed towards the P2 and P3 binding pockets of thrombin but rather towards the oxyanion hole or had the complementary functional groups of the P2 and P3 pockets interchanged. Especially when using DOCK's chemical scoring or FlexX, the incorrect dockings still made reasonable hydrogen bonds and hydrophobic interactions and did not display highly unlikely ligand conformations as illustrated in Figure 6. In other words, without previous knowledge of the binding mode of thrombin inhibitors it would have been difficult to select the correct conformations by visual inspection or calculated score alone.

Interestingly, the incorrect docking of BM14.1248 (Figure 6) is most likely caused by the conformation of the peptide plane defined by Ser^{214} and Trp^{215} in the 1ETS and 1DWC thrombin structures. This peptide linkage was observed to be rotated in the complex of BM14.1248 with thrombin [25], allowing for the formation of a hydrogen bond between the 4-amino group of the inhibitor and the Ser^{214} carbonyl oxygen. Since this interaction cannot be established with the 1DWC and 1ETS thrombin conformations, the incremental construction algorithm alternatively directed the 4-amino group towards the Ser^{195} hydroxyl oxygen. This conformation is apparently more favorable in the context of the 1ETS and 1DWC structures than the experimentally observed orientation of the 4-amino group, and the remaining part of the inhibitor can therefore no longer be reconstructed correctly from the 4-amino group onwards. The role of local receptor conformation in this example was confirmed by flexibly docking BM14.1248 against the 1UVT thrombin structure, which yielded a solution within 0.9 Å rmsd of the native conformation. This example illustrates the sensitivity of the incremental construction algorithm to small, localized differences in receptor conformation. Such differences can strongly affect the scoring of partial docking solutions and thereby bias or misguide the overall placement of the ligand.

Subtle differences are observed between dockings against the high-resolution crystal structure of thrombin (1ETS) and those against 1DWC. As for rigid body docking, the 1ETS structure yields the best overall results for DOCK 4.0 using chemical scoring which may be due to a less restrictive conformation of the D-loop containing residues Tyr^{60A} and Trp^{60D} in this structure. However, docking of all inhibitors to the 1PPH trypsin structure, which lacks the D-loop, yielded results similar to those obtained with thrombin and DOCK chemical scoring by docking eight inhibitors with rmsd values below 2 Å (results not shown).

Table 4. Ranking of 32 flexibly docked thrombin inhibitors with rmsds values, values smaller than or equal to 2 Å among the 10 best scorers. Docking was performed with standard sampling. A >10 rank implies that no docking solutions with an rmsd below 2 Å were among the 10 highest ranking dockings

Inhibitor	FlexX	DOCK (chemical)	DOCK (energy)
3-TAPAP	6	>10	>10
AMH	1	8	7
Argatroban	>10	1	>10
Benzamidine	1	1	1
BM14.1203	>10	>10	>10
BM14.1224	>10	3	>10
BM14.1238	>10	7	>10
BM14.1241	>10	>10	1
BM14.1243	>10	2	3
BM14.1244	>10	1	1
BM14.1248	>10	1	3
BM18.0537	>10	>10	>10
BM18.0540	>10	7	>10
BM51.0986	>10	1	>10
BM51.1011	>10	>10	>10
BM51.1022	5	1	3
BM51.1023	>10	>10	>10
BM51.1031	>10	3	>10
BM51.1037	>10	1	6
BM51.1045	>10	1	>10
BM51.1047	>10	1	4
BM51.1059	>10	4	>10
BM51.1081	>10	8	1
BM51.1110	>10	1	4
BMS186282	>10	>10	>10
BMS189090	>10	>10	>10
DAPA	>10	4	>10
FBA	2	1	3
NAPAP	1	3	4
PPA	>10	10	>10
PTA	1	2	8
4-TAPAP	>10	>10	>10

In order to assess if conformations close to the experimentally determined binding mode are among the highest scorers, the 10 highest scoring docking solutions for each ligand were examined (see Table 4). For

FlexX, 7 out of 32 inhibitors had a conformer with less than 2 Å rmsd among the 10 best dockings. For DOCK 4.0, this number was 23 and 14 for chemical and energy scoring, respectively. This suggests that even when conformations other than the highest ranking one are considered, the true binding mode may still not be identified. Inhibitors that are docked correctly by all three methods are mostly small and characterized by a relatively small number of non-conjugated rotatable bonds (for instance, NAPAP). PPA has one more rotatable bond compared to FBA and presents a much more difficult case for flexible docking. A similar reasoning appears to apply to argatroban, DAPA, BMS186282 and BMS189090, which contain flexible arginine side chains. Besides size and flexibility, differences in sampling, scoring and torsion angle libraries among the two methods may also play a role, but their influence is less easily distilled from the relative rankings.

Flexible docking with increased conformational sampling

In an attempt to increase the number of correctly docked solutions, we increased the sampling of both DOCK 4.0 and FlexX in terms of the number of partial solutions that are kept for further extension and the total number of complete solutions stored (see the Methods section). It should be noted that such an increase in sampling also increases the time required to perform a docking to over 20 min per compound with DOCK 4.0 and over 5 min with FlexX, rendering application to large databases impractical. Although DOCK samples less conformations than FlexX in our increased sampling runs (250 compared to 1000 final structures), the run times are much longer. Figures 7A and B show the results obtained with increased sampling with DOCK 4.0 while applying standard chemical scoring and the results obtained with FlexX. For FlexX an increase in sampling does not appear to result in a significantly larger number of low rmsd solutions. Increased sampling with DOCK 4.0 yields more low rmsd solutions (38–47% below 2 Å) when chemical scoring is applied. Interestingly, the results obtained with energy scoring appear to deteriorate with increased sampling, suggesting the existence of minima in the energy function different from the native conformation or still incomplete conformational sampling. Indeed, if the rmsd values of individual inhibitors docked with standard and increased sampling are compared, it is seen that for some compounds the rmsd increases by 1–2 Å with sampling. This shows that even at increased sampling convergence is not always reached. For 7–10 out of 32 inhibitors, the DOCK score of the flexibly docked inhibitor is more favorable than that of the docked crystal structure conformation even though these dockings have rmsd values of 3–5 Å. This indicates that the minima of both chemical and energy scoring in DOCK 4.0 do not always correspond to the true bound conformation.

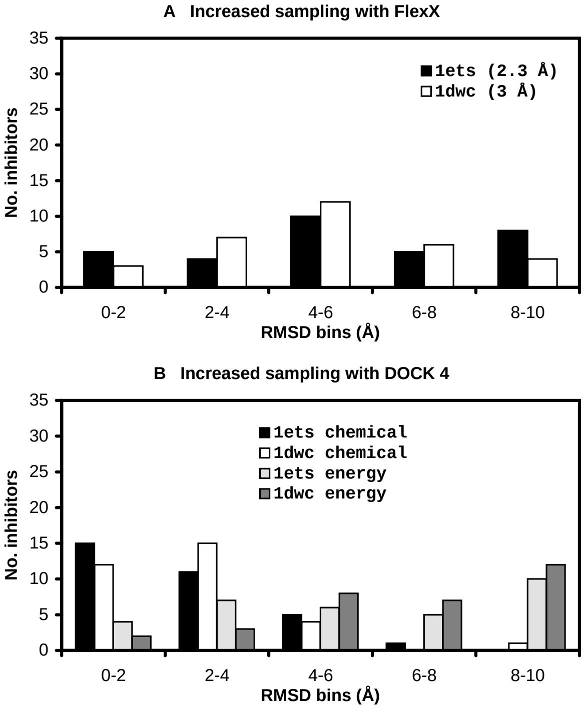

Figure 7. Effect of increased conformational sampling on the frequency distributions of ligand rmsd values with respect to the bound conformations of 32 flexibly docked thrombin inhibitors. Results are shown for flexible dockings with FlexX 1.5 (A) and DOCK 4.0 (B) using chemical and energy scoring. All compounds were docked against two thrombin crystal structures determined at 2.3 and 3 Å resolution, respectively, and indicated by their PDB codes 1ETS and 1DWC.

In conclusion, for both receptor conformations it is seen that flexible docking with DOCK 4.0 and rigid body docking of diverse conformers using chemical scoring yield the most low rmsd solutions. Flexible docking produces slightly more solutions with rmsd values below 2 Å and, especially when using energy scoring, appears to perform better on the higher resolution thrombin structure 1ETS.

Discussion

Although the incremental construction algorithm is currently the fastest available method to dock a fully flexible molecule in the context of a receptor binding site, the accuracy with which ligands can be docked is still limited. Less than 40% of the 32 flexible ligands are docked within 2 Å of their known binding modes. Increased conformational sampling only improved the results for DOCK 4.0 using chemical scoring. More rigorous sampling seems undesirable in terms of the CPU time required to perform a database docking. In this respect, an average time of 1–3 min spent per ligand under standard sampling conditions is already too long for searching, for instance, a database consisting of more than 100 000 compounds such as the available chemicals directory (MDL, San Leandro, CA, U.S.A.). An accurate flexible docking tool operating at such a speed would nevertheless still be valuable for docking small (combinatorial) libraries of compounds in drug-design projects. For such applications, however, the quality of dockings generated with DOCK 4.0 and FlexX 1.5 still appears to be too unreliable. It should be noted that the CPU time required for increased sampling with DOCK is on average 4–5 times longer than for FlexX, even though less conformations are sampled. It is unclear though whether this is due to a different implementation, more thorough minimization by DOCK, faster scoring by FlexX or other factors.

Flexible docking with DOCK 4.0 does, however, perform significantly better than rigid docking of single conformations generated by CORINA and comparable to docking of multiple conformers from a flexibase. Slightly more low rmsd solutions are found with flexible docking with DOCK compared to our flexibase results. Given the limitations of using thrombin as a single test-case, however, this difference may not be significant. Rigid body docking of a single conformation for each molecule is still often used for searching large (>100 000 compounds) 3D databases. On the basis of our results, it appears likely that the majority of larger, putative inhibitors in a single conformer database will not be docked correctly and may therefore not be identified as inhibitors due to the lack of ligand flexibility. The diverse, multiple conformer database generated by CONF provides a good, although more time-consuming, intermediate solution for introducing ligand flexibility. It has the additional advantages of not requiring the presence of anchor fragments in every ligand and being less sensitive to the scoring function during ligand placement. However, since the CPU time required for docking a flexibase scales linearly with the number of conformers that is examined, docking of larger flexibases still remains unfeasible.

Possible explanations for the less accurate dockings performed by FlexX compared to DOCK 4.0 concern the observations that the experimentally determined binding modes do not always have the most favorable Böhm score and that automatic selection of multiple anchor fragments leads to more and different solutions than those obtained using manual selection of the ‘true’ anchor fragment. Both factors have been discussed previously by Rarey et al. [15] for the docking of argatroban to 1DWC where the known binding mode scored 10 kJ/mol less favorable compared to the conformations suggested by FlexX. We were able to reproduce these results for argatroban using automatic anchor detection (rmsd 8 Å) and manual selection of the guanidino group as the anchor fragment (rmsd 2.9 Å), suggesting that our protocol is similar to that used by Rarey et al. Apparently, the choice of anchor fragments has important consequences for the final composition of the ensemble of conformations generated by FlexX, and the Böhm scoring function is not always able to differentiate between correct and incorrect binding modes within such an ensemble. In this respect, it is noted that the relatively simple chemical scoring scheme of DOCK 4.0 performs better than both the DOCK interaction energy and Böhm’s scoring methods in docking our selection of thrombin inhibitors. Perhaps more effective scoring functions for flexible docking can be derived in the context of ligand flexibility rather than by fitting binding data to static complexes. Such an approach would yield some indication of the capability of such a scoring function to select the correct bound conformation from an ensemble of reasonable ligand conformations. Finally, the use of discrete torsion angle libraries may also prevent some inhibitor binding modes from being generated.

In our dockings, FlexX sampled the entire binding site while for DOCK 4.0 sampling was focused on the P1 specificity pocket by only placing spheres in that region. The results obtained with DOCK sphere sets describing the entire active site did not deviate much from those obtained with focused sampling, suggesting that focused sampling on the P1 pocket does not explain the differences in performance between DOCK and FlexX.

Although the incremental construction algorithm represents an important step forward towards fast docking of flexible ligands, some aspects of it may still require improvement. Since the algorithm constructs ligands one fragment at a time, it is highly sensitive to decisions taken at each point of extension. A locally favorable solution may have unfavorable consequences for completion of the ligand further on in the building process. One could perceive algorithms which after placement of the anchor fragment attempt to place the rest of the ligand in a single step. Alternatively, back-tracking might allow the docking program to return to earlier partial solutions when problems are encountered with reconstruction of a ligand [12]. Finally, although the use of rigid anchor fragments is favorable both from a computational point of view and on the basis of entropy considerations, flexible docking of ligands lacking such fragments is likely to remain problematic.

Conclusions

Two different implementations of the incremental construction algorithm for flexible docking, DOCK 4.0 and FlexX 1.5, have been compared in their ability to flexibly dock 32 thrombin inhibitors. Neither method is capable of modeling all inhibitors correctly as only 10–35% of them are docked within 2 Å of their known binding modes. Increased conformational sampling did improve the number of low rmsd solutions for DOCK 4.0 with chemical scoring, but not for FlexX 1.5 and DOCK 4.0 with energy scoring. In general, chemical scoring as implemented in DOCK 4.0 performs better in docking our database of thrombin inhibitors than interaction energy scoring by DOCK and the Böhm scoring function implemented in FlexX. Nevertheless, flexible docking with DOCK 4.0 does provide a more accurate alternative to rigid body docking of single ligand conformations. The use of a diverse flexibase such as generated with CONF provides an intermediate approach between using single conformer databases and fully flexible docking at the cost of longer processing times.

Subtle differences in the quality of the docking results are observed depending on the receptor conformation. Due to the sensitivity of the incremental construction algorithm to local receptor structure in scoring partial solutions, differences in the final ligand conformations are to be expected when using different receptor structures. Therefore, if different receptor structures are available they may require individual evaluation as to their suitability for docking or the entire ensemble could be used [42].

Increased conformational sampling in DOCK 4.0 yielded correct dockings for about 47% of the inhibitors. However, the best dockings obtained with limited sampling are not always a subset of those obtained with increased sampling. Apparently, it is not always possible to reach convergence within a reasonable amount of time. In addition, increased sampling removes the advantage of speed originally offered by the incremental construction algorithm.

Chemical scoring, which is an intuitively modified version of interaction energy scoring in DOCK 4.0, yields higher quality dockings than molecular mechanics energy scoring in DOCK 4.0 or the Böhm scoring function used by FlexX 1.5 for our set of thrombin inhibitors. Apparently, scoring functions that are more readily interpretable from a physical point of view do not automatically offer better performance in molecular docking. In this respect, optimization of scoring functions for molecular docking applications should perhaps be done in the context of ligand flexibility rather than by fitting static protein–ligand complexes.

Acknowledgements

We thank Dr. M. Rarey of the German National Research Center for Information Technology at Sankt Augustin (Germany) for making an evaluation version of FlexX v. 1.5 available to us, and A.G. Skillman and T. Ewing of the Department of Pharmaceutical Chemistry at the University of California at San Francisco for helpful discussions and assistance with DOCK 4.0.

References

1. Kuntz, I.D., Science, 257 (1992) 107.
2. Kuntz, I.D., Meng, E.C. and Shoichet, B.K., Acc. Chem. Res., 27 (1994) 117.
3. Charifson, P.S. and Kuntz, I.D., In Charifson, P.S. (Ed.) Practical Application of Computer-Aided Drug Design, Marcel Dekker, New York, NY, 1997, pp. 1–38.
4. Grootenhuis, P.D.J., Knegtel, R.M.A., Heikoop, J.C. and van Boeckel, C.A.A., In van der Goot (Ed.) Trends in Drug Research II, Elsevier, Amsterdam, 1998, pp. 7–14.
5. Lorber, D.M. and Shoichet, B.K., Protein Sci., 7 (1998) 938.
6. Rosenfeld, R., Vajda, S. and DeLisi, Annu Rev. Biophys. Biomol. Struct., 24 (1995) 677.
7. Lambert, M.H., In Charifson, P.S. (Ed.) Practical Application of Computer-Aided Drug Design, Marcel Dekker, New York, NY, 1997, pp. 243–303.

8. Jones, G., Willett, P., Glen, R.C., Leach, A.R. and Taylor, R., J. Mol. Biol., 267 (1997) 727.
9. Goodsell, D.S., Morris, G.M. and Olson, A.J., J. Mol. Recog., 9 (1996) 1.
10. Leach, A.R. and Kuntz, I.D., J. Comput. Chem., 13 (1992) 730.
11. Rarey, M., Kramer, B., Lengauer, T. and Klebe, G., J. Mol. Biol., 261 (1996) 470.
12. Makino, S. and Kuntz, I.D., J. Comput. Chem., 18 (1997) 1812.
13. Welch, W., Ruppert, J. and Jain, A.J., Chem. Biol., 3 (1996) 449.
14. Ewing, T.J.A. and Kuntz, I.D., J. Comput. Chem., 18 (1997) 1175.
15. Rarey, M., Kramer, B. and Lengauer, T., J. Comput.-Aided Mol. Design, 10 (1997) 369.
16. Kramer, B., Rarey, M. and Lengauer, T., Proteins Struct. Funct. Genet. (Suppl. 1) (1997) 221.
17. Böhm, H.-J., J. Comput.-Aided Mol. Design, 8 (1994) 243.
18. Weiner, S.J., Kollman, P.A., Nguyen, D.T. and Case, D.A., J. Comput. Chem., 7 (1986) 230.
19. Meng, E.C., Shoichet, B.K. and Kuntz, I.D., J. Comput. Chem., 13 (1992) 505.
20. Böhm, H.-J. and Klebe, G., Angew. Chem., Int. Ed. Engl., 35 (1996) 2588.
21. Knegtel, R.M.A. and Grootenhuis, P.D.J., In Kubinyi, H., Folkers, G. and Martin, Y. (Eds.) 3D QSAR in Drug Design. Recent Advances, Kluwer, Dordrecht, 1998, pp. 99–114.
22. Dixon, J.S., Proteins Struct. Funct. Genet. (Suppl. 1) (1997) 198.
23. Grootenhuis, P.D.J. and Karplus, M., J. Comput.-Aided Mol. Design, 10 (1996) 1.
24. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.T.J., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M., J. Mol. Biol., 112 (1977) 535.
25. Engh, R.A., Brandstetter, H., Sucher, G., Eichinger, A., Baumann, U., Bode, W., Huber, R., Poll, T., Rudolph, R. and von der Saal, W., Structure, 4 (1996) 1353.
26. Von der Saal, W., Kucznierz, R., Leinert, H. and Engh, R.A., Bioorg. Med. Chem. Lett., 7 (1997) 1283.
27. Brandstetter, H., Turk, D., Hoeffken, W., Grosse, D., Sturzebecher, J., Martin, P.D., Edwards, B.F.P. and Bode, W., J. Mol. Biol., 226 (1992) 1085.
28. Banner, D.W. and Hadvary, P., J. Biol. Chem., 266 (1991) 20085.
29. Bode, W. and Turk, J., Eur. J. Biochem., 193 (1990) 175.
30. Sadowski, J., Gasteiger, J. and Klebe, G., J. Chem. Inf. Comput. Sci., 34 (1994) 1000.
31. Rusinko, A., Sheridan, R.P., Nilakantan, R., Haraki, K.S., Bauman, N. and Venkataraghavan, R., J. Chem. Inf. Comput. Sci., 29 (1989) 251.
32. Gasteiger, J. and Marsili, M., Tetrahedron Lett., 36 (1980) 3219.
33. McHugh, J., Algorithmic Graph Theory, Prentice-Hall, 1990, pp. 90–114.
34. Bayada et al., J. Chem. Inf. Comput. Sci., submitted.
35. Bauknecht, H., Zell, A., Bayer, H., Levi, P., Wagener, M., Sadowski, J. and Gasteiger, J., J. Chem. Inf. Comput. Sci., 36 (1996) 1205.
36. Grootenhuis, P.D.J. and van Galen, P.J.M., Acta Crystallogr., D51 (1995) 560.
37. Bursi, R.B. and Grootenhuis, P.D.J., J. Comput.-Aided Mol. Design (1999) in press.
38. Kuntz, I.D., Blaney, J.M., Oatley, S.J., Langridge, R. and Ferrin, T.E., J. Mol. Biol., 161 (1982) 269.
39. Ewing, T. (Ed.) DOCK Version 4.0 Manual, Regents of the University of California, San Francisco, CA, 1997.
40. Kearsley, S.K., Underwood, D.J., Sheridan, R.P. and Miller, M.D., J. Comput.-Aided Mol. Design, 8 (1994) 565.
41. Miller, M.D., Kearsley, S.K., Underwood, D.J. and Sheridan, R.P., J. Comput.-Aided Mol. Design, 8 (1994) 153.
42. Knegtel, R.M.A., Kuntz, I.D. and Oshiro, C.M., J. Mol. Biol., 266 (1997) 424.
43. Todeschini, R., Lasagni, M. and Marengo, E., J. Chemometrics, 8 (1994) 263.
44. Todeschini, R., Gramatica, P., Provenzani, R. and Marengo, E., Chemometr. Intell. Lab. Syst., 27 (1995) 221.
45. Todeschini, R. and Gramatica, P., Quant. Struct.-Act. Relat., 16 (1997) 113.
46. Broto, P., Moreau, G. and Vandycke, C., Eur. J. Med. Chem., 19 (1984) 66.
47. Broto, P., Moreau, G. and Vandycke, C., Eur. J. Med. Chem., 19 (1984) 79.

Appendix

Descriptors used to select a diverse set of conformers from a library of random conformers

Distance bins	$dB_1 \Rightarrow d_1 = 2$ Å
$\text{volume}_x =$ spherical volume of ray d_x-	$dB_2 \Rightarrow d_2 = 4$ Å
spherical volume of ray d_{x-1}	$dB_3 \Rightarrow d_3 = 6$ Å
$\text{distBin}_x =$ (number of atoms in volume_x centered at molecule's centroid)/volume_x	$dB_4 \rightarrow d_4 = 9$ Å
Note: $d_0 = 0$	
WHIMS (no weight)	λ 1
[43–45]	$\lambda 2$
	$\upsilon 1$
	$\upsilon 2$
	$\gamma 1$
	$\gamma 2$
	$\kappa 1$
	$\kappa 2$
WHIMS	$V\lambda 1$
(van der Waals weight)	$V\lambda$
[43–45]	$V\upsilon 1$
	$V\upsilon 2$
	$V\gamma 1$
	$V\gamma 2$
	$V\kappa 1$
	$V\kappa 2$
Acceptor – acceptor distance bins	AA1 $\Rightarrow d_1 = 4$ Å
AAx = number of acceptor atoms at	AA2 $\Rightarrow d_2 = 6$ Å
a distance from another acceptor	AA3 $\Rightarrow d_3 = 9$ Å
atom $> d_{x-1}$ and $\leq d_x$	AA4 $\Rightarrow d_4 = 12$ Å
Note: $d_0 = 2$ Å	
Acceptor–donor distance bins	AD1
See above	AD2
	AD3
	AD4
Donor–donor distance bins	DD1
See above	DD2
	DD3
	DD4
Van der Waals surface area	surface
Autocorrelation of partial charges	AUTO1 $\Rightarrow d_1 = 4$ Å
[46,47]	AUTO2 $\Rightarrow d_2 = 6$ Å
	AUTO3 $\Rightarrow d_3 = 9$ Å
	AUTO4 $\Rightarrow d_4 = 12$ Å
	AUTO5 $\rightarrow d_5 = 15$ Å