Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology,
T. Lengauer, R. Schneider, P. Bork, D. Brutlag, J. Glasgow, H.-W. Mewes, R. Zimmer (eds.)
AAAI Press, Menlo Park, CA, 1999, pp. 242–251

Database Screening for HIV Protease Ligands: The Influence of Binding-Site Conformation and Representation on Ligand Selectivity

Volker Schnecke and **Leslie A. Kuhn**
Protein Structural Analysis and Design Laboratory
Department of Biochemistry, Michigan State University
East Lansing, MI 48824-1319, U.S.A.
Phone: (517) 355-3455, Fax: (517) 353-9334
E-mail: volker@sol.bch.msu.edu, kuhn@agua.bch.msu.edu
WWW: http://www.bch.msu.edu/labs/kuhn

Abstract

Screening for potential ligands and docking them into the binding sites of proteins is one of the main tasks in computer-aided drug design. Despite the progress in computational power, it remains infeasible to model all the factors involved in molecular recognition, especially when screening databases of more than 100,000 compounds. While ligand flexibility is considered in most approaches, the model of the binding site is rather simplistic, with neither solvation nor induced complementary usually taken into consideration. We present results for screening different databases for HIV-1 protease ligands with our tool SLIDE, and investigate the extent to which binding-site conformation, solvation, and template representation generate bias. The results suggest a strategy for selecting the optimal binding-site conformation, for cases in which more than one independent structure is available, and selecting a representation of that binding site that yields reproducible results and the identification of known ligands.

Introduction

The screening of compound databases for ligands that bind to a target protein is the computational counterpart to high-throughput screening and one of the early steps in computer-aided drug design. Screening can be done by similarity search, i.e., looking for ligands that resemble known ligands in structure and activity. When the binding site of the target protein is known, another way to identify potential ligands is by docking the screened compounds into the binding site. Although this search is more focused, because additional information is included in the model, there are several points that influence the outcome of the search process, like

- the conformation of the binding site,
- the solvation of the binding site, and
- the template used to represent favorable interaction centers within the binding site.

These points are also relevant for approaches to the docking problem, i.e., the computation of a favorable binding mode for a single ligand, but most docking methods require a computation time that makes them infeasible for screening a set of, say, 100,000 compounds. Reasonable screening times should be in the order of a few days for a large database. Spending only one minute per compound, which is still a lower bound for the time used by the fastest docking tools (Welch, Ruppert, & Jain 1996; Rarey *et al.* 1996; Knegtel, Kuntz, & Oshiro 1997), would allow the screening of only 10,000 compounds within a week. In fact, since it is more effective to screen as much as diverse a set of molecules as possible, then tune them using medicinal chemistry strategies, the goal is to develop computational screening methods for millions of compounds.

New docking tools are typically verified on a large variety of test cases, with up to 200 protein-ligand complexes redocked to test the robustness of the method (Jones *et al.* 1997; Baxter *et al.* 1998; Rarey, Kramer, & Lengauer 1999). Compared to this, screening tools are often tested only for a small number of targets, but databases of more than 100,000 compounds are typically screened (Welch, Ruppert, & Jain 1996; Lorber & Shoichet 1998; Shoichet, Leach, & Kuntz 1999). In this article, we focus on a different point: we have only one target, HIV-1 protease, but structures with diverse conformations and several binding-site representations to test, and we screen three different databases for potential ligands. We are trying to assess how binding-site conformation and template configuration influence the outcome of a database search, and use this to help define a more realistic model for simulating molecular recognition during screening and docking.

Background

The fastest of the existing docking tools, some of which have been already used for database screening, represent the binding site of the protein by a template of points, onto which ligand atoms or interaction centers are matched during the search for favorable binding modes. In the docking tool DOCK, the template typically consists of up to 100 spheres, which generate a negative image of the binding site (Kuntz *et al.* 1982; Shoichet & Kuntz 1993). During the search, subsets

of ligand atoms are matched to these spheres, based on internal distances. Ligand flexibility can be considered either by incremental construction of a ligand in the binding site or by docking different ligand conformations separately during the optimization. However, when screening databases, DOCK typically keeps the ligands rigid, or docks a set of rigid conformers for each compound (Lorber & Shoichet 1998; Shoichet, Leach, & Kuntz 1999).

Other approaches specify a set of interaction points, defining favorable positions for placing polar ligand atoms or hydrophobic (nonpolar) centers, e.g., aromatic rings. Such a template can be generated automatically, e.g., by placing probe points on the solvent-accessible surface of the binding site (Ruppert, Welch, & Jain 1997), or interactively by superimposing known complexes to identify favorable interaction points based on observed ligand binding modes (Schnecke *et al.* 1998).

FlexX uses a template of 400 to 800 points to define positions for favorable interactions of hydrogen-bond donors and acceptors, metal ions, aromatic rings, and methyl groups, when docking drug-sized molecules (Rarey *et al.* 1996). The ligand is fragmented, incrementally constructed in the binding site, and matched to template points based on geometric hashing techniques. Bond-torsional flexibility is modeled discretely, and a tree-search algorithm is used to keep the most promising partially constructed ligand conformations during the search. Although FlexX is a fast docking tool, no applications to screening have been reported.

Hammerhead uses up to 300 hydrogen-bond donor/acceptor and van der Waals interaction points to define a template, and the ligand is incrementally constructed (Welch, Ruppert, & Jain 1996). A fragment is docked based on matching ligand atoms to template points with compatible internal distances, similar to the matching algorithm used in DOCK. Hammerhead was used to screen 80,000 fully flexible ligands within a few days on typical hardware.

GOLD uses a template based on hydrogen-bond donors and acceptors of the protein and applies a genetic algorithm to sample over all possible combinations of intermolecular hydrogen bonds and ligand conformations (Jones *et al.* 1997). It has been shown to reproduce known complexes for a large variety of cases. However, especially due to the use of a non-deterministic optimization technique, the computation time for docking a single ligand is much higher than for the incremental approaches, which makes GOLD infeasible for screening large sets of molecules.

The results of fast docking tools to predict ligand binding modes *a priori*, i.e., for cases, where no structures with the particular ligand or similar ligands are available, are less reliable (Dixon 1997). The main reason for this may be the lack of modeling the induced fit of the binding site upon ligand docking. Today, most docking tools model full ligand flexibility, but at least in the faster approaches the binding site is kept rigid. Limited protein side-chain flexibility is exploited by GOLD, which considers rotational flexibility of hydrogens (Jones *et al.* 1997), other approaches use rotamer libraries (Leach 1994; Jackson, Gabb, & Sternberg 1998), are based on molecular dynamics simulations (Wassserman & Hodge 1996; Apostolakis, Plückthun, & Caflisch 1998), or dock ligands into aligned ensembles of different structures of the target protein (Knegtel, Kuntz, & Oshiro 1997). Approaches to use explicit protein flexibility (Totrov & Abagyan 1997) or domain movements (Sandak, Wolfson, & Nussinov 1998) in docking simulations have also been reported. However, binding-site flexibility during screening has only been implemented in Slide, and its precursor Specitope, both developed in our laboratory (Schnecke *et al.* 1998; Schnecke & Kuhn 1999a).

Another point that influences the accuracy of docking simulations is the solvation of the binding site (Ladbury 1996). Water bound in the ligand site is known to be a critical determinant of ligand specificity for HIV-1 protease (Lam *et al.* 1994), cholera toxin (Merritt *et al.* 1994), and other proteins, and is a ubiquitous component in molecular recognition (Raymer *et al.* 1997). For the docking tool FlexX a technique called the particle concept has been recently proposed, which adds water molecules at favorable positions during the incremental construction of the ligand in the binding site (Rarey, Kramer, & Lengauer 1999). It was tested for 200 protein-ligand complexes and the accuracy of the predicted binding modes increased for several cases, including HIV protease. A different approach has been introduced for database screening using DOCK: here, the molecules in the database are solvated, which improves the ranking for known ligands and filters out molecules with inappropriate charge states and sizes in comparison to screening without solvation (Shoichet, Leach, & Kuntz 1999).

Methods

In this section we give an overview of our screening tool, Slide, which is described in detail elsewhere (Schnecke & Kuhn 1999a). It takes a more general approach to binding-site representation, solvation, ligand types, and induced flexibility than its precursor Specitope, which we used to screen a peptidyl database for ligands matching a hydrogen-bond interaction template for a protein binding site (Schnecke *et al.* 1998; Schnecke & Kuhn 1999b). Slide uses a larger, more general template consisting of hydrogen-bond and hydrophobic interaction points. Critical solvent is included based on *Consolv* predictions (Raymer *et al.* 1997) or by bound water sites conserved in a series of ligand-bound structures. Ligand dockings are based on mapping dynamically chosen anchor fragments onto triangles of template points, and full ligand flexibility and protein side-chain flexibility are applied to model induced complementarity (Figure 1).

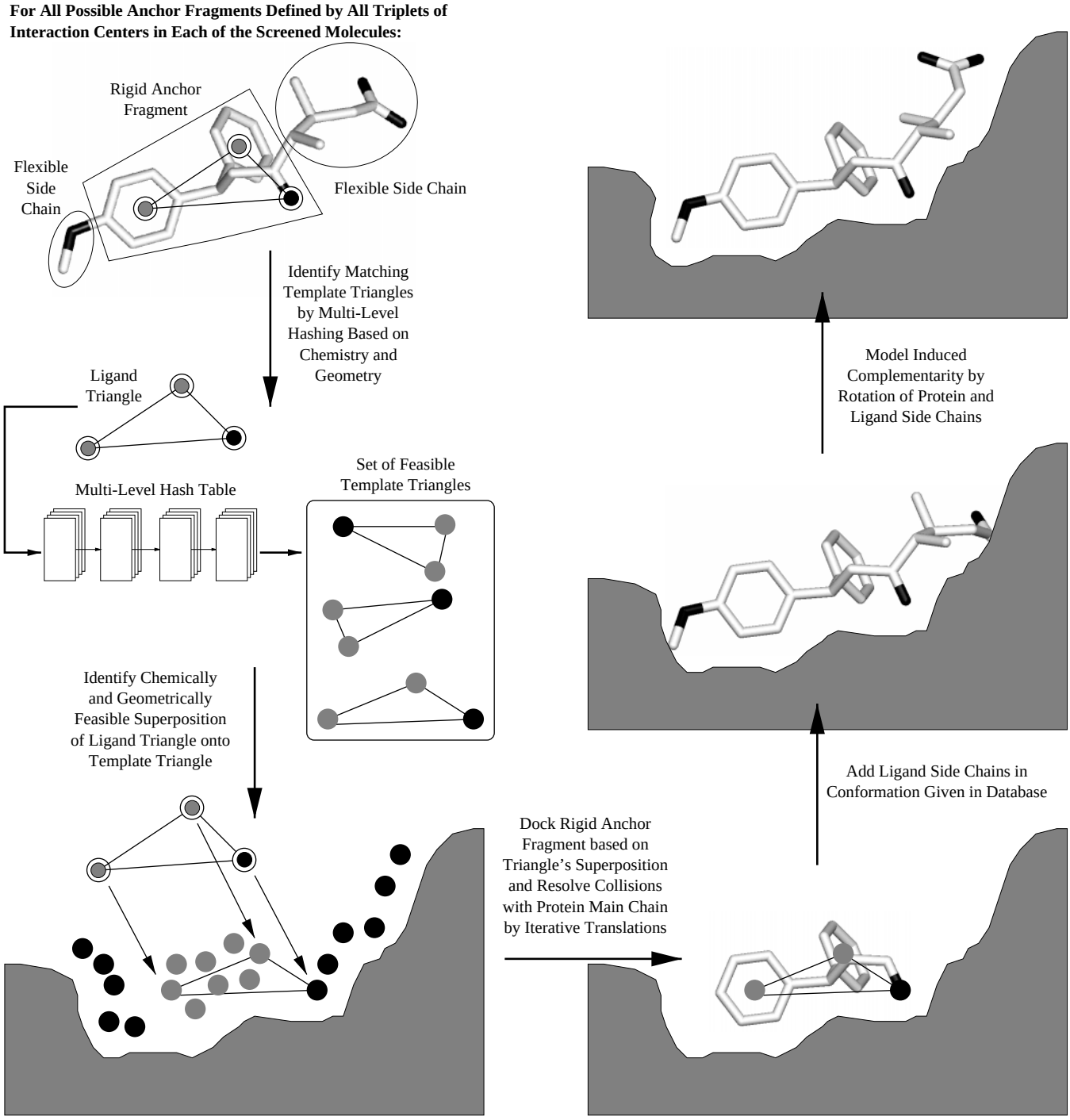

Figure 1: SLIDE's docking of potential ligands into the binding site is based on mapping triplets of ligand interaction centers (H-bond donors, acceptors, or hydrophobic ring centers) onto triangles of template points located above the protein surface. Feasible template triangles for each possible triplet in a screened molecule are directly accessed via a multi-level hash table, and the corresponding mapping is used to dock the rigid anchor fragment of the ligand. Single bonds in the flexible parts of both molecules are rotated to generate a shape-complementary interface, before the complex is scored by the number of intermolecular hydrogen bonds and hydrophobic complementarity of the contact surfaces. In all steps the ligand triplets or dockings that do not meet a particular threshold are discarded.

The screening tool SLIDE

SLIDE (for 'Screening Ligands by Induced-fit Docking Efficiently') is a tool that is able to screen a database of 3D structures of over 67,000 small organic molecules within hours or more than 185,000 peptides, which are more flexible, within a couple of days on an ordinary desktop workstation. The binding site of the target protein is described by a template of favorable interaction points above the surface, onto which ligand atoms are mapped during the search. The two interaction types are hydrogen-bond points, which are matched by ligand H-bond donors or acceptors, and hydrophobic points, which are matched by the centers of ligand carbon rings. A typical template has between 10 and 150 points, and depending on its size and shape, the search can be biased towards ligands with similar interaction patterns, or provide a large variety of potential ligands. Special interaction points that must be matched by the ligand can also be included, which is useful to ensure that a certain part of the binding pocket is covered.

Multi-level hashing. During the search, all triangles of interaction points in the screened molecules are mapped exhaustively onto triangles of template points with compatible geometry and chemistry, and such a mapping serves as a basis for docking the molecule into the binding site. A multi-level hashing approach is used to directly access all template triangles with feasible chemistry and geometry for each set of three ligand interaction centers. Before the search, all possible template triangles are generated from the set of binding-site template points and are indexed via four levels of hash tables.

The first hash table is based on the chemistry of the interaction points of the triangles. There are four types of points (hydrophobic, H-bond donor, H-bond acceptor, and H-bond donor/acceptor), which provides 20 indices when taking all possible three-element combinations out of these four types. The index in the second hash table is based on the perimeter of the triangles, the third on the length of their longest side, and the fourth on the length of the shortest side. By using these four properties for a given triplet of interaction centers in a ligand candidate, all template triangles with compatible geometry and chemistry can be directly and very efficiently accessed. The best feasible one-to-one mapping of the ligand triangle onto each of the indexed triangles is computed, which is then used to transform the ligand interaction centers onto the corresponding template points by applying a least-squares fit superposition.

Docking of the anchor fragment. The matched ligand interaction centers define the anchor fragment, and all chemically and geometrically feasible anchor fragments are tested for each ligand candidate. All flexible bonds within this part of the ligand are rigidified. The remaining parts of the ligand are kept flexible (Figure 1, top), such that all single bonds in these parts can be rotated later, if necessary to resolve collisions with protein atoms. Collisions of the anchor fragment with protein main-chain atoms are resolved by iterative translations of the fragment as a rigid body (Schnecke *et al.* 1998). If all main-chain collisions can be resolved, the side-chains are added to the anchor fragment, in the conformation found for the ligand in the database.

Modeling of induced complementarity. Induced fit for the interface between the two molecules is modeled by resolving any collisions of their flexible parts by directed rotations of single bonds either in the ligand or in side chains of the protein. There are typically several applicable rotations to resolve an intermolecular collision, and an approach based on mean-field theory (Koehl & Delarue 1994) is used to decide which rotations will improve the shape complementarity in the current conformation.

For each pairwise intermolecular collision, those bonds are identified that can be rotated to resolve it without causing an intramolecular collision. They are stored in a system together with the corresponding minimum rotation angle and the number of non-hydrogen atoms that will be displaced by the rotation. These values provide the basis for a force, which represents the cost of a rotation. A probability is assigned to each rotation, and all rotations that can be used to resolve one particular collision are initialized with equal probabilities. During the mean-field based optimization, these probabilities converge to assign higher values to those rotations that represent a near-optimal choice to resolve a maximal number of collisions with minimal conformational changes in both molecules, without bias to one or the other.

In each cycle of the mean-field optimization process, a mean force is computed for each rotation in the system, which is based on the force associated with this rotation and all correlations with other rotations in the system. Two rotations correlate, when they are not independent of each other, hence, only one of them should be applied at the end of an iteration of the mean-field optimization process. There are both positive and negative correlations, which decrease or increase the mean force, respectively, and their contributions are weighted by the probabilities of the corresponding rotations. It is especially beneficial when a rotation of a single bond in the system can resolve more than one collision. An example for a negative correlation is a rotation that would displace another bond that is included in the system, so that all corresponding computations regarding that bond would no longer be valid.

The probabilities for all rotations in the system are updated at the end of each cycle, taking into account the mean forces of alternative rotations for the same collision. After up to ten cycles, the probabilities have con-

verged to an approximate optimal set of values, which assigns the highest probabilities to those rotations that solve most of the collisions with the least overall cost. These rotations are applied if they do not cause any intramolecular collisions. If there are still overlaps, up to ten iterations of the mean-field optimization are done to generate shape-complementary conformations of the two molecules. This comprehensive approach to simultaneously modeling protein and ligand flexibility provides a more realistic representation of induced complementarity than has achieved for other screening approaches, which focus on ligand flexibility.

Scoring of protein-ligand complexes. Whenever a collision-free complex is generated, a score is assigned to the ligand based on the number of intermolecular hydrogen bonds, HBONDS(P, L), and the hydrophobic complementarity, HPHOB(P, L), of the interface (Schnecke *et al.* 1998). If not provided in the protein or ligand structure, the position of the shared hydrogen in each intermolecular hydrogen bond is computed, and all hydrogen bonds with a donor–acceptor distance between 2.7 and 3.5Å and a donor-hydrogen-acceptor angle larger than 120° contribute equally to the score. If water molecules are included in the interface, all water-mediated hydrogen bonds are also considered.

For computing the hydrophobic complementarity, HPHOB(P, L), hydrophilicity values from a study on the relative hydration of protein atoms are used (Kuhn *et al.* 1995). The hydrophobicity value of each ligand atom is compared to the average hydrophobicity of all protein atoms within 4.0Å. If the environment is compatible to the ligand atom, a positive contribution is added to the overall hydrophobic complementarity (Schnecke *et al.* 1998).

The score for a protein-ligand complex is the weighted sum of these terms:

$$\mathrm{SCORE(P,L)} = \mathrm{A} \cdot \mathrm{HBONDS(P,L)} + \mathrm{B} \cdot \mathrm{HPHOB(P,L)}$$

The relative contributions A and B were tuned to reflect experimentally determined affinities of 89 protein-ligand complexes (Eldridge *et al.* 1997).

HIV protease conformation and binding-site representation

Three different structures of HIV-1 protease available in the Brookhaven Protein Databank (PDB) (Abola *et al.* 1987) and the HIV protease database[1] were used for the experiments:

- PDB entry 1dif (HIVdb 16nci) is a complex with a inhibitor, which contains a difluoroketone motif, at 1.7Å resolution (R-value 19.8%) (Silva *et al.* 1996).
- PDB entry 1htg (HIVdb 3glx) is a complex with a penicillin-derived inhibitor at 2.0Å resolution (R-value 19.0%) (Jhoti *et al.* 1994).
- PDB entry 1hhp (HIVdb 1pip) is a ligand-free structure with 2.7Å resolution (R-value 19.0%) (Spinelli *et al.* 1991).

These structures were chosen because of their relatively high resolution, their diverse ligands (Figure 2), and their identical amino-acid sequence.

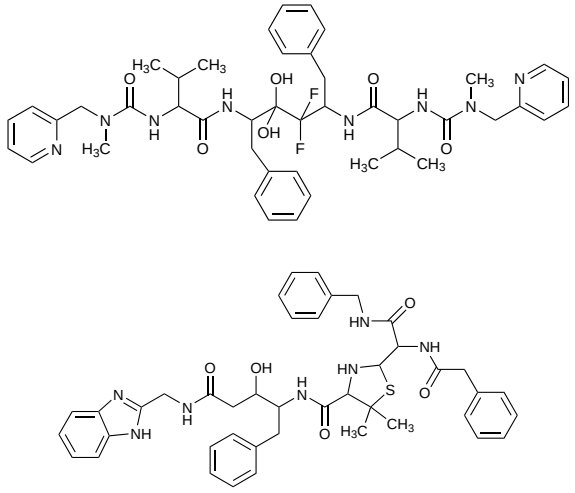

Figure 2: The ligands from PDB complexes 1dif, **A79285** (top), and 1htg, **GR137615** (bottom)

Different approaches to generate the binding-site templates were used. For the two target structures taken from complexes (PDB 1dif and 1htg), the binding site was kept in the conformation optimized to bind the known ligand, i.e., no energy minimization was performed after removing the ligand. Two templates were generated based on crystallographically observed binding modes of HIV-1 protease ligands. The first template for each case (template **K**, for 'known ligand') is based on positions of H-bond donors and acceptors and centers of aromatic rings in the corresponding known ligand, which yielded 16 points for 1dif and 32 for 1htg. The template in the latter case is larger, because two ligand-binding modes were observed in the crystal structure.

The second template (**A**, for 'average') is equal for both structures and is based on 15 binding modes of ligands in 13 relatively high-resolution complexes (PDB 1dif, 1hiv, 1hpv, 1hpx, 1hsg, 1htg, 1hvc, 1hvi, 1hvj, 1hvk, 1hvl, 5hvp, and 9hvp). The coordinates were taken from the aligned structures as provided in the HIV protease database. Close interaction points were clustered using complete-linkage clustering, and the final template consists of 92 points.

The third template (**UB**, for 'unbiased') was automatically generated by filling the binding site with random points and probing the protein neighborhood of each point for hydrogen-bond donors or acceptors or hydrophobic surface atoms. Points of equal type were then clustered to produce a template of 125, 124, and

[1] `http://www-fbsc.ncifcrf.gov/HIVdb`

92 points for 1dif, 1htg, and 1hhp, respectively.

Binding-site solvation

The inclusion of binding-site solvation for the HIV protease in this study is very simple: we consider one water molecule which occurs at the same position in the complex structures (HOH 2 in PDB 1dif and HOH 308 in PDB 1htg), and it is considered an essential part of the protein. In addition to reducing the volume of the binding site, which might rule out some ligands that only fit into the unsolvated binding site or require conformational changes for their docking, this water molecule can mediate interactions and thus influence the ranking of the potential ligands. The templates **KW** and **AW** are the templates **K** and **A**, respectively, with the water molecule added to the binding site.

The databases

Three databases are used in the experiments that are described below. The first database is a subset of the Cambridge Crystallographic Database System (CSD). This subset includes 67,573 crystal structures of small organic compounds with fewer than 100 heavy atoms and at least three interaction centers. The second database consists of 3D structures of 185,235 tetrapeptides. These have been generated by taking all overlapping fragments from 762 dissimilar (< 25% identity) protein chains in the PDB (PDB-select list (Hobohm & Sander 1994), August 1998) and assigning hydrogen atoms. The third database contains the ligands of 34 HIV-1 protease complexes from the PDB. Their conformations were taken from the complex, and hydrogen atoms were added.

Results

The difference in the binding-site conformation of the ligand-bound structures PDB 1dif and 1htg, caused by the diverse ligands, is shown in Table 1. The focus in this section is on the influence of the binding-site conformation on the potential ligands that were identified by SLIDE when screening with equal, or binding-site specific templates. Ideally, the search should be robust, i.e., similar potential ligands should be identified, independent of binding-site conformation and template.

CSD Screening

The focus here is on the influence of the binding-site conformation and representation on the set of potential ligands found in the CSD screening database. For the highly specific templates **K** and **KW** there is almost no overlap of the sets for binding sites 1dif and 1htg (Table 3), as expected, because the new ligands should be biased towards the known ligand for each case. When using the many-ligand template **A**, 18.5% of the potential ligands are found for both binding sites. The unbiased templates **UB**, although they are different for 1dif and 1htg and extremely binding-site specific, give 65.6% of the CSD ligands in common. When comparing these results with the ligand set for the unbound target structure 1hhp (**UB**), 25.4% of these ligands match those found with the unbiased template for 1dif, and 27.6% for 1htg.

	all atoms	**main chain**
# Atoms	154	88
RMSD	0.555Å	0.390Å
Average	0.423Å	0.327Å
Minimum	0.012Å	0.012Å
Maximum	2.792Å	1.062Å

Table 1: The difference of atom positions in the binding sites of PDB 1dif and 1htg. Listed are the root-mean-square deviation (RMSD), and the average, minimal, and maximal displacement between corresponding atoms in the superimposed structures. The values are based on 22 residues that have at least one atom within 3.5 Å of a ligand atom.

In Table 2, the focus is on the potential CSD ligands that SLIDE identified for different templates with the same binding-site conformation. The sets for the same target have only a small percentage of ligands in common. For example, 15.0% of all ligands that were found for binding site 1dif when screening with template **K** were also found when screening with template **A**. Although the known ligand for this structure was included in the set of the 15 HIV protease inhibitors used to generate templates **A** and **AW**, template **K** is not simply a subset of template **A**. Each of the 15 ligands provided between 10 and 32 template points, and nearby points were clustered, so that the final template consisted of average favorable interaction points for the known ligands.

PDB 1dif				
	KW	**A**	**AW**	**UB**
K	**58.8%**	15.0%	3.8%	0.0%
KW		11.8%	5.9%	0.0%
A			**98.3%**	5.7%
AW				3.5%

PDB 1htg				
	KW	**A**	**AW**	**UB**
K	**100.0%**	**31.0%**	4.8%	**23.8%**
KW		13.0%	8.7%	**26.1%**
A			**94.6%**	10.6%
AW				10.8%

Table 2: The percentage of potential CSD ligands in common for different template representations for the two binding sites, 1dif and 1htg.

The binding-site conformation in structure 1htg shows a lower sensitivity to different templates, e.g.,

CSD Ligands		PDB 1dif					1hhp
		K	**KW**	**A**	**AW**	**UB**	**UB**
PDB 1htg	**K**	0.0%	2.3%	11.9%	0.0%	9.5%	9.5%
	KW	0.0%	0.0%	0.0%	0.0%	8.7%	13.0%
	A	0.0%	2.0%	18.5%	7.8%	**21.3%**	9.9%
	AW	0.0%	2.0%	12.9%	7.5%	7.5%	7.5%
	UB	2.5%	5.9%	9.1%	9.6%	**65.6%**	**27.6%**
PDB 1hhp	**UB**	1.3%	2.0%	5.6%	6.1%	**25.4%**	N/A

K/KW – template based on binding mode of one known ligand without/with water
A/AW – template based on average binding modes of 15 ligands without/with water
UB – unbiased, automatically generated template for one protein structure

Table 3: The set of potential ligands identified from the screened CSD compounds varies with binding-site conformation and template configuration. This table shows the percentage of ligands found in common between the structures and templates for each possible combination.

the overlaps of the ligand sets for the unbiased template **UB** with all the others varies between 10.6% and 26.1%, compared to 0.0% to 5.7% for binding site 1dif (Table 2). As one might expect, the addition of the single water molecule yields only a small change in the overall composition of the ligands found in both cases. However, due to water-mediated interactions and the resulting conformational changes in the ligands, there can be significant variation in the relative ranking of different ligands, which is not reflected in these tables.

Peptide screening

With over 185,000 tetrapeptides and the non-uniform frequency of the twenty amino acids in proteins there is some redundancy in the peptide database, which provides different conformers for some four-residue sequences. Thus, in the following, we consider two peptides as equal if they have the same amino-acid sequence.

Overall, the results obtained when screening the peptide database are similar to those from the CSD screens in terms of sensitivity to conformation and template design. What is not apparent in the tables is that for the binding site 1htg with the average templates (**A** and **AW**), about twice as many potential ligands (259) were found in the peptide database than in the CSD subset (122). For binding site 1dif the effect is opposite. The average number of potential ligands for the other templates are 49 (CSD) and 83 (peptides) for **K**/**KW** and 254 (CSD) and 95 (peptides) for **UB**. There are fundamental differences between the two databases: all peptides are of roughly similar size, very polar and similar to the known HIV-protease inhibitors, although the known ligands are larger, consisting of five to six residues. The CSD compounds are less flexible, smaller, and generally more hydrophobic. The sensitivity to the template (Table 5) is higher when screening peptides compared to the CSD results (Table 2); the fewer ligands found in common for peptide screening with different templates relative to organic compounds may be explained by the greater conformational flexibility of peptides and the difficulty of assessing optimal conformations in the time allowed for screening.

PDB 1dif				
	KW	**A**	**AW**	**UB**
K	**90.9%**	5.6%	0.0%	11.0%
KW		9.1%	0.0%	9.1%
A			**88.0%**	4.9%
AW				4.0%

PDB 1htg				
	KW	**A**	**AW**	**UB**
K	**71.0%**	16.1%	9.7%	0.0%
KW		15.2%	9.1%	0.0%
A			**95.8%**	7.3%
AW				5.4%

Table 5: The percentage of peptides in common for different template representations.

HIV protease ligands

When screening the HIV ligand database with templates **K** and **KW**, the known ligands were identified for 1dif and 1htg. In addition, for binding site 1dif, seven other HIV protease ligands were correctly identified with this simple template (Table 6).

	K	**KW**	**A**	**AW**	**UB**
PDB 1htg	1	1	8	8	0
PDB 1dif	8	8	17	17	2
PDB 1hhp	N/A	N/A	N/A	N/A	1

Table 6: The number of ligands that were identified by SLIDE out of the set of 34 known HIV protease inhibitors for different templates and binding sites.

Screening with templates **A** or **AW**, in all cases the known ligand was docked correctly (RMSD: 0.268Å for

Peptidyl Ligands		PDB 1dif					1hhp
		K	**KW**	**A**	**AW**	**UB**	**UB**
PDB 1htg	**K**	0.0%	0.0%	3.2%	0.0%	3.2%	0.0%
	KW	0.0%	0.0%	3.0%	0.0%	0.0%	0.0%
	A	11.1%	9.1%	**20.6%**	**24.0%**	7.6%	3.6%
	AW	11.1%	**18.2%**	14.8%	**18.0%**	5.8%	2.4%
	UB	0.0%	0.0%	3.2%	4.0%	9.8%	3.6%
PDB 1hhp	**UB**	0.0%	0.0%	1.2%	0.0%	1.2%	N/A

Table 4: The influence of the binding-site conformation when screening the peptide database.

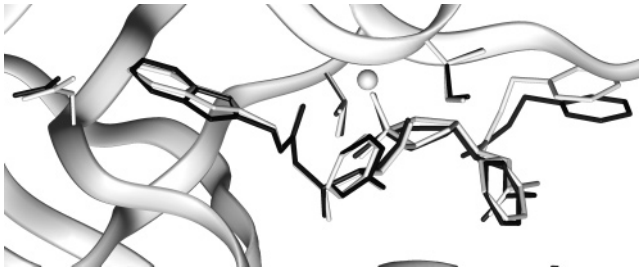

Figure 3: The ligand from PDB 1htg docked by SLIDE correctly into the binding site of PDB 1dif when screening with the average template incorporating the conserved water site (**AW**); the sphere above the ligands is conserved water HOH 2. SLIDE's binding mode is shown in black, and the superimposed ligand from complex 1htg is shown in white (RMSD 0.518Å). Three key protein side chains, which were rotated upon docking, are shown in their original conformation (white) and in their final conformation (black).

1dif, 0.451Å for 1htg). With this configuration, either with or without including the conserved water molecule, 17 of the 34 known ligands were identified for binding site 1dif, in comparison to 8 ligands for binding site 1htg using the same template (Table 6). A possible reason for this is that the binding site in structure 1dif is more open, since its ligand is larger (Figure 2; 1dif ligand: volume 614.0Å^3, molecular weight 831; 1htg ligand: volume 553.8Å^3, weight 762). Figure 3 shows the ligand from 1htg as it was correctly docked by SLIDE into the binding site of 1dif using template **AW**. With the unbiased template **UB**, none of the known ligands was identified for the binding site 1htg; for the ligand-free structure, 1hhp, one known ligand was identified, and for 1dif two were identified.

Discussion

The results presented above show that the outcome of the screening process is quite dependent on binding-site conformation and template configuration. It might seem sobering that for most cases half or more of the 34 known inhibitors were not identified by SLIDE, especially since their conformations in the database were already the favorable ones for interacting with HIV proteases. However, in the context of screening, the time spent for conformational search, when generating complementary conformations of the two molecules, is necessarily very limited. One factor that determines the screening time is the template configuration. For the CSD ligands, the screening time varied between 20 minutes (i.e., 0.018 seconds per molecule on a Sun Ultra 1/140) for the 16-point template **K** (1dif) versus 10.5 hours (0.58 seconds per molecule) for the average template **A** (1dif), up to 18.5 hours (1.02 seconds per molecule) for the unbiased template **UB** for binding site 1htg.

When screening the peptides, there were more ligand interaction centers, and thus many more ways of docking the anchor fragment, and in addition to this, the peptides were much more flexible than the CSD compounds. Thus, the vast majority of peptides went through the conformational search, such that about one second was spent on average per molecule for the smaller templates, and up to 4.8 seconds for the large, unbiased templates. For the most successful screen of HIV protease ligands (1dif with template **A**), about 17.6 seconds were spent for each molecule. Considering the size and especially the high degree of conformational freedom for these ligands, even the fastest docking tools will take several minutes when redocking them to a given rigid binding site in the favorable conformation. Because of their more thorough conformational search, these tools might be able to dock ligands into binding sites optimized for other ligands, but the final binding modes are likely to be imperfect, when induced flexibility of the protein is not modeled.

Another point that clearly influences the screening results is the configuration of the binding-site template. With a template based on known binding modes, it is likely that the known ligands are identified by the screening process, on the other hand, the search can be biased towards ligands with similar interaction patterns, especially when using only a few template points (Figure 4). Screening with the unbiased template minimized the effect of binding-site conformation for the CSD screens, thus was the most robust screening run, but failed to dock all but two of the known ligands.

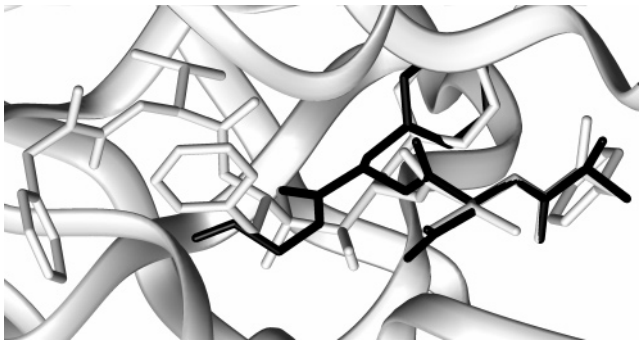

Figure 4: An example for a four-residue peptide (sequence ADFG, black tubes) docked by SLIDE into the binding site of PDB 1dif using template **K**, which was based on the known ligand, which is shown in white tubes. Note the nice resemblance of this peptide to that part of the known ligand, onto which it is mapped.

The approach described in this paper can be used as a general method to decide which structure out of a set of comparable structures would be the best target when screening for new lead compounds, and which template representation is the best. When identifying new ligands, the search should be robust, i.e., influenced as little as possible by the template representation. When comparing the results for the binding sites 1dif and 1htg, the binding site 1htg gives more consistent results, in terms of showing greater overlap in the sets of potential ligands for different template representations, especially for the CSD screens. This robustness is a reason for preferring this binding site over the other. However, when screening the 34 known HIV protease inhibitors, the binding site 1dif, independent of the template representation, always resulted in more known ligands being identified than for binding site 1htg. In fact, when screening with the unbiased template **UB** for that binding site, this was the only case where none of the known ligands was identified.

Conclusions

Three databases were screened for potential ligands to three different conformations of HIV-1 protease, represented by five different binding-site templates. Using an average template based on 15 known ligands of HIV protease, our screening algorithm, SLIDE, was able to identify half of the known inhibitors and dock them accurately while modeling the necessary induced complementarity of protein and ligand. However, our main goal was to develop a method for optimally choosing a protein conformation and template design for use in screening and docking, independent of the particular docking strategy chosen. Our results show that the choice of the structure to use as a screening target, when several independent structures are available, depends on whether reproducibility (ability to identify the same ligands, regardless of the template representation) or identification of known ligands is more important. Reproducibility can be analyzed by comparing the number of ligands in common for the different binding-site conformations, using unbiased templates or templates based on the average positions in known ligands. Similarly, the different binding-site conformations can be used to screen known ligands to identify which conformation identifies the most. In terms of optimal template representation, the unbiased template representation tended to find the most CSD ligands in common, given different protein conformations. The average template, based on averaged positions of polar and hydrophobic centers in the known ligands, resulted in at least moderate commonality of ligands for the different binding-site conformations and always identified the most known ligands.

Acknowledgements

The authors gatefully acknowledge support for this project by the Deutsche Forschungsgemeinschaft (DFG postdoctoral fellowship 576/1-1 to V.S.) and the National Science Foundation (NSF grant BIR9600831 to L.A.K.).

References

Abola, E. E.; Bernstein, F. C.; Bryant, S. H.; Koetzle, T. F.; and Weng, J. 1987. Protein Data Bank. In Allen, F. H.; Bergerhoff, G.; and Sievers, R., eds., *Crystallographic Databases – Information Content, Software Systems, Scientific Applications*. Bonn/Cambridge/Chester: Data Commission of the International Union of Crystallography. 107–132.

Apostolakis, J.; Plückthun, A.; and Caflisch, A. 1998. Docking small ligands in flexible binding sites. *Journal of Computational Chemistry* 19(1):21–37.

Baxter, C. A.; Murray, C. W.; Clark, D. E.; Westhead, D. R.; and Eldridge, M. D. 1998. Flexible docking using tabu search and an empirical estimate of binding affinity. *Proteins: Structure, Function, and Genetics* 33:367–382.

Dixon, J. S. 1997. Evaluation of the CASP2 docking section. *Proteins: Structure, Function, and Genetics* Supplement 1:198–204.

Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; and Mee, R. P. 1997. Empirical scoring functions: I. the development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *Journal of Computer-Aided Molecular Design* 11:425–445.

Hobohm, U., and Sander, C. 1994. Enlarged representative set of protein structures. *Protein Science* 3(3):522–524.

Jackson, R. M.; Gabb, H. A.; and Sternberg, M. J. E. 1998. Rapid refinement of protein interfaces incorpo-

rating solvation: Application to the docking problem. *Journal of Molecular Biology* 276:265–285.

Jhoti, H.; Singh, O. M.; Weir, M. P.; Cooke, R.; Murray-Rust, P.; and Wonacott, A. 1994. X-ray crystallographic studies of a series of penicillin-derived asymmetric inhibitors of HIV-1 protease. *Biochemistry* 33(28):8417–8427.

Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; and Taylor, R. 1997. Development and validation of a genetic algorithm for flexible docking. *Journal of Molecular Biology* 267:727–748.

Knegtel, R. M. A.; Kuntz, I. D.; and Oshiro, C. M. 1997. Molecular docking to ensembles of protein structures. *Journal of Molecular Biology* 266:424–440.

Koehl, P., and Delarue, M. 1994. Application of a self-consistent mean field theory to predct protein sidechains conformation and estimate their conformational entropy. *Journal of Molecular Biology* 239:249–275.

Kuhn, L. A.; Swanson, C. A.; Pique, M. E.; Tainer, J. A.; and Getzoff, E. D. 1995. Atomic and residue hydrophilicity in the context of folded protein structures. *Proteins: Structure, Function, and Genetics* 23:536–547.

Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; and Ferrin, T. E. 1982. A geometric approach to macromolecule-ligand interactions. *Journal of Molecular Biology* 161:269–288.

Ladbury, J. E. 1996. Just add water! the effect of water on the specificity of protein-ligand binding sites and its potential application to drug design. *Chemistry & Biology* 3:973–980.

Lam, P. Y. S.; Jadhav, P. K.; Eyermann, C. J.; Hodge, C. N.; Ru, Y.; Bacheler, L. T.; Meek, J. L.; Otto, M. J.; Rayner, M. M.; Wong, Y. N.; Chang, C.-H.; Weber, P. C.; Jackson, D. A.; Sharpe, T. R.; and Erickson-Viitanen, S. 1994. Rational design of potent, bioavailable, nonpeptide cyclic ureas as HIV protease inhibitors. *Science* 263:380–384.

Leach, A. R. 1994. Ligand docking to proteins with discrete side-chain flexibility. *Journal of Molecular Biology* 235:345–356.

Lorber, D. M., and Shoichet, B. K. 1998. Flexible ligand docking using conformational ensembles. *Protein Science* 7(4):938–950.

Merritt, E. A.; Sarfaty, S.; van den Akker, F.; L'Hoir, C.; Martial, J. A.; and Hol, W. G. 1994. Crystal structure of cholera toxin B-pentamer bound to receptor GM1 pentasaccharide. *Protein Science* 3(2):166–175.

Rarey, M.; Kramer, B.; Lengauer, T.; and Klebe, G. 1996. A fast flexible docking method using an incremental construction algorithm. *Journal of Molecular Biology* 261:470–489.

Rarey, M.; Kramer, B.; and Lengauer, T. 1999. The particle concept: Placing discrete water molecules during protein-ligand docking predictions. *Proteins: Structure, Function, and Genetics* 34:17–28.

Raymer, M. L.; Sanschagrin, P. C.; Punch, W. F.; Venkataraman, S.; Goodman, E. D.; and Kuhn, L. A. 1997. Predicting conserved water-mediated and polar ligand interactions in proteins using a k-nearest-neighbors genetic algorithm. *Journal of Molecular Biology* 265:445–464.

Ruppert, J.; Welch, W.; and Jain, A. N. 1997. Automatic identification and representation of protein binding sites for molecular docking. *Protein Science* 6:524–533.

Sandak, B.; Wolfson, H. J.; and Nussinov, R. 1998. Flexible docking allowing induced fit in proteins: Insights from an open to closed conformational isomers. *Proteins: Structure, Function, and Genetics* 32:159–174.

Schnecke, V., and Kuhn, L. 1999a. Modeling induced fit and controlling molecular diversity during database screening for ligands. *Proteins: Structure, Function, and Genetics* in review.

Schnecke, V., and Kuhn, L. A. 1999b. Flexibly screening for molecules interacting with proteins. In Thorpe, M. F., and Duxbury, P. M., eds., *Rigidity Theory and Applications*. Plenum Press. 385–400.

Schnecke, V.; Swanson, C. A.; Getzoff, E. D.; Tainer, J. A.; and Kuhn, L. A. 1998. Screening a peptidyl database for potential ligands to proteins with side-chain flexibility. *Proteins: Structure, Function, and Genetics* 33:74–87.

Shoichet, B. K., and Kuntz, I. D. 1993. Matching chemistry and shape in molecular docking. *Protein Engineering* 6(7):723–732.

Shoichet, B. K.; Leach, A. R.; and Kuntz, I. D. 1999. Ligand solvation in molecular docking. *Proteins: Structure, Function, and Genetics* 34:4–16.

Silva, A. M.; Cachau, R. E.; Sham, H. L.; and Erickson, J. W. 1996. Inhibition and catalytic mechanism of HIV-1 aspartic protease. *Journal of Molecular Biology* 255(2):321–346.

Spinelli, S.; Liu, Q. Z.; Alzari, P. M.; Hirel, P. M.; and Poljak, R. J. 1991. The three-dimensional structure of the aspartyl protease from the HIV-1 isolate BRU. *Biochimie* 73(11):1391–1396.

Totrov, M., and Abagyan, R. 1997. Flexible protein-ligand docking by global energy optimization in internal coordinates. *Proteins: Structure, Function, and Genetics* Supplement 1:215–220.

Wassserman, Z. R., and Hodge, C. N. 1996. Fitting an inhibitor into the active site of thermolysin: A molecular dynamics case study. *Proteins: Structure, Function, and Genetics* 24:227–237.

Welch, W.; Ruppert, J.; and Jain, A. N. 1996. Hammerhead: Fast, fully automated docking of flexible ligands to protein binding sites. *Chemistry & Biology* 3:449–462.