
Perspectives in Drug Discovery and Design, **20:** 171–190, 2000.
KLUWER/ESCOM

Virtual screening with solvation and ligand-induced complementarity

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.

Summary. We present our database-screening tool SLIDE, which is capable of screening large data sets of organic compounds for potential ligands to a given binding site of a target protein. Its main feature is the modeling of induced complementarity by making adjustments in the protein side chains and ligand upon binding. Mean-field theory is used to balance the conformational changes in both molecules in order to generate a shape-complementary interface. Solvation is considered by prediction of water molecules likely to be conserved from the crystal structure of the ligand-free protein, and allowing them to mediate ligand interactions, if possible, or including a desolvation penalty when they are displaced by ligand atoms that do not replace the lost hydrogen bonds. A data set of over 175 000 organic molecules was screened for potential ligands to the progesterone receptor, dihydrofolate reductase, and a DNA-repair enzyme. In all cases the screening time was less than a day on a Pentium II processor, and known ligands as well as highly complementary new potential ligands were found.

Key words: bound water, dihydrofolate reductase, DNA repair enzymes, docking, drug design, flexibility, molecular recognition, progesterone receptor

Introduction

Screening a large database of organic compounds for potential ligands to a protein is often seen as a simple extension of the docking problem, which is the prediction of the favorable binding mode for a single ligand. When doing ligand screening by docking, as in our screening tool SLIDE, the docking problem must be solved for each ligand candidate in the database. But, because hundreds of thousands of ligands are screened, the time that a screening tool can spend for each compound must be far less than several minutes, which is the typical runtime for fast docking tools that model full ligand flexibility [1–3]. When spending only one minute per compound, the screening time for a database of 100 000 molecules is about 10 weeks. In order to

* To whom correspondence should be addressed. E-mail: kuhn@agua.bch.msu.edu.

reduce the runtime to a realistic time frame, say, a day, it is first important to efficiently rule out infeasible candidates, then to focus on the few promising molecules in the database.

The limitation of the time for conformational search in a screening tool also affects its scoring function, which is used to rate the complementarity of protein and ligand in a given conformation. Rather than estimating the binding affinity, a computational intensive and still imprecise science, the goal of the scoring function in a screening tool is to give an appropriate relative ranking for the potential ligands, with known or new 'real' ligands obtaining top ranks. Ideally, such a scoring function should be robust, with real ligands obtaining high scores irrespective of their exact binding mode. Another important and often neglected aspect of docking and database screening is the induced shape complementarity of the protein upon ligand binding. Many cases are known in which the binding site undergoes significant conformational changes when binding different ligands [4–6]. When assessing the quality of docking tools, typically known ligands are redocked into fixed binding sites that are tailored to bind that very ligand, since they are taken from the corresponding crystallographic complex. Although this is likely to bias the selectivity for the known ligand and its score relative to other candidates, the effect might be minor for lead optimization, where similar ligands are docked to compare their binding modes and relative affinities. However, when screening compounds from a database for lead discovery, bias towards known ligands should be avoided in the search. Our approach is to screen using the ligand-free conformation of the target protein, when available, and to model induced complementarity for the protein side chains as well as ligand when screening and docking a large variety of potential ligands.

In this article, we describe applications of our screening tool SLIDE [7], which is able to reduce large compound databases of more than 175 000 organic molecules to a ranked list of approximately 100 docked potential ligands within an hour to two days, depending on the binding-site characteristics and degree of flexibility in the screened molecules. In addition to ligand flexibility, SLIDE models full flexibility of ligands and protein side chains when docking potential ligands, and uses *Consolv* [8] to predict water-mediated interactions with the ligand.

Background

The majority of results reported for database screening are based on the application of tools that were designed to predict the favorable binding mode and the binding affinity for a single ligand. Several docking tools are available, and they are widely used for structure-based ligand design [2,3,9–19].

Docking tools can be classified by the method they use to represent the binding site, by the technique for sampling ligand conformations, and by the way they construct the docked ligand. All docking tools fast enough to screen a large data set of molecules are based on so-called descriptor-matching approaches [20], which means that they represent the binding site by a template of points, onto which ligand atoms are mapped during the search. The template points can describe the shape of the binding site [9,21], or favorable chemical interaction centers above the protein surface, where hydrogen-bond donors or acceptors, metal ions, or hydrophobic groups of the ligands can be placed [1–3,11,22]. During the search for the optimal binding mode of a ligand, different conformers for the molecule are generated, which can be done randomly, e.g., by using a genetic algorithm [16,17,19,23], or by systematically sampling discrete torsional angles for the rotatable bonds of the ligand [1–3]. The faster docking tools construct the ligand incrementally in the binding site [2,3], or dock fragments of the ligand independently and chemically link them later, if the combination is feasible [1,12–14,24]. All docking tools that have been used for database screening employ a binding-site template for guiding the search and incremental construction of the ligand in the binding site [1,25–30].

While all recent docking tools consider full ligand flexibility, induced complementarity of the protein upon ligand binding is not modeled, at least not in the faster docking tools. In approaches that model protein flexibility, this is often limited, e.g., by only rotating terminal hydrogens to optimize intermolecular hydrogen bonding [16,17], or by using rotamer libraries for the side-chain conformations [31,32]. Other approaches model explicitly defined side-chain flexibility [33], hinge bending [34], or they dock ligands against an ensemble of protein structures [35]. Molecular dynamics simulations [36–38] may yield the most realistic models of protein and ligand flexibility, but the resulting runtime for docking a single ligand is likely to be in the range of hours.

A drawback of many docking tools is that they neglect the effect of binding-site solvation and the potential for water-mediated interactions between protein and ligand [8,39–42]. While it is certainly possible to consider bound water molecules as part of the rigid protein in most docking tools, recently three sophisticated approaches have been reported, which either predict conserved binding-site waters [8], compute potential water positions prior to docking [43], or solvate the ligand molecule [29].

Recent docking and screening tools can identify potential ligands from up to 150 000 compounds within a few days, when considering full ligand flexibility [1,25–28,44]. Even in the absence of modeling inducible complementarity and solvation, there have been successful project reports in structure-based

lead discovery or design, including the identification of new inhibitors for thymidylate synthase [25], *P. carinii* dihydrofolate reductase (DHFR) [27], *P. falciparum* DHFR [45], trypanothione reductase [46], and human thrombin [47]. Our goal with the new screening tool SLIDE is to incorporate a balanced model of protein and ligand flexibility as well as a knowledge-based model of solvation that is fast enough to be used for screening and docking hundreds of thousands of compounds.

Methods: The screening tool SLIDE

SLIDE (for 'Screening for Ligands by Induced-fit Docking') can screen databases of 3D structures of over 100 000 small organic molecules, typically within hours to a day, on an ordinary desktop workstation. It has also been used for screening 185 000 peptides, which are more flexible, within a few days [7,48]. SLIDE uses multi-level hashing, mean-field theory, and an empirically tuned scoring function to efficiently recognize infeasible compounds, dock the most promising ligand candidates, and produce a ranked list of some 100 potential ligands for a given protein target.

Representing the binding site

The binding site of the protein is described by a template of favorable interaction points above its surface, onto which ligand atoms are mapped during the search. A template includes four different types of points:

- **Hydrogen-bond donor point.** During screening, SLIDE can place a hydrogen-bond donor of the ligand onto this point, which is determined to be within favorable hydrogen-bonding distance of a protein hydrogen-bond acceptor.
- **Hydrogen-bond acceptor point.** Each acceptor point is within favorable hydrogen-bonding distance of a protein hydrogen-bond donor.
- **Hydrogen-bond donor/acceptor point.** This is within hydrogen-bonding distance of both a hydrogen-bond acceptor and a donor of the protein, so either a ligand hydrogen-bond donor or acceptor can be placed here, or a group that can accept and donate at the same time (e.g., hydroxyl oxygen).
- **Hydrophobic interaction center.** These points are placed above a hydrophobic surface patch of the protein, and are matched by the centers of the most hydrophobic ligand groups, hydrocarbon rings.

The template can be automatically generated based on the ligand-free structure of the protein, which reduces bias towards known ligands. For automatic template generation, the binding site is filled with random points that are 2.5 to 5.0 Å from a protein atom. To determine favorable hydrogen-bonding positions, each of these points is checked for donors or acceptors in the protein within a distance of 2.5 to 3.5 Å; for protein hydrogen-bond donors, the angle between the donor, the donated hydrogen, and the probe point is also taken into account, and must be larger than 120°. Hydrophobic points are located between 3.5 and 5.0 Å from the nearest protein atom. For these points, the average hydrophilicity of all protein atoms within 5.0 Å is below 0.1, indicating a hydrophobic site (based on the values provided in Reference 49). All points of the same type are then clustered using complete-linkage clustering [50] to yield a computationally tractable number of template points (typically up to 200). Similarly, interaction points in each potential ligand are defined as those that can act as hydrogen bond acceptors, donors, acceptors and/or donors (e.g., hydroxyl oxygen atoms), or hydrophobic centers. The latter are defined as the centers of hydrocarbon rings with 6 or fewer carbon atoms (e.g., cyclohexane or benzene rings). Hydrogen-bond donors or acceptors in the ligand candidates are identified for oxygen, nitrogen, sulfur, and halogen atoms based on the molecular orbital type, valency, and presence of hydrogen atoms in SYBYL mol2 format files prepared for each molecule in the ligand database. The interaction points in ligand candidates are mapped onto points in the binding site template having the same chemistry.

Alternatively, the template can be defined based on interaction patterns observed in complexes with known ligands, biasing the search towards ligands with similar interaction patterns, similar to pharmacophore-based screens. For either the ‘unbiased’, automatically generated templates, or templates designed based on known ligand binding, special key interaction points that must be matched by the ligand can also be included. This is useful to ensure that a certain part of the binding site is covered, or that a docked ligand makes particular interactions. Beyond the template, which governs the selection of complementary ligands, the binding site of the protein is represented by a shell of surface residues and water molecules likely to mediate protein-ligand interactions.

During the ligand search, all triangles of hydrogen-bond and hydrophobic 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 a molecule into the binding site. A multi-level hashing approach is used to directly access all template triangles with feasible chemistry and geometry for a given set of three interaction centers in the ligand. 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 (indexing) tables. The indices in these hash tables are based on the chemistry (H-bond donor/acceptor or hydrophobic) of the three triangle points, on the perimeter of the triangle, and then on the longest and the shortest side for each of the indexed template triangles. 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. For feasible matches between each ligand triangle and template triangle, the geometrically best mapping is computed, which is then used to transform the ligand triangle onto the corresponding template points by applying a least-squares fit superposition. When including key points in the template, only those triangles that include at least one of these key interaction centers are indexed in the hash tables.

Docking the anchor fragment

The matched ligand interaction centers define the anchor fragment, which is the part of the molecule containing the three interaction centers. To maintain the distances between these matched points, all flexible bonds within this anchor fragment are rigidified. All chemically and geometrically feasible anchor fragments are then exhaustively tested in each ligand candidate for their ability to match triangles within the protein template. Collisions of the anchor fragment with protein main-chain atoms are resolved by iterative translations of the fragment as a rigid body. For this, a global translation vector is used to shift the anchor fragment the minimal amount necessary to resolve all collisions [5]. If all main-chain collisions can be resolved, the remaining atoms of the ligand are added to the anchor fragment in the conformation found for the molecule in the database. These atoms outside the anchor fragment are considered flexible, such that all single bonds in these parts can be rotated later, to resolve collisions with protein atoms. At this point we retain only those ligand dockings with at least 50% of their carbon atoms buried against the protein in order to keep only those dockings with good shape complementarity and minimal exposure of hydrophobic atoms to solvent; our analysis of 89 known protein-ligand complexes [51] showed they all met this criterion [7].

Modeling induced complementarity

Induced fit between the two molecules is modeled by resolving any collisions of their flexible parts by directed rotations of single bonds in either the ligand or side chains of the protein. This follows the paradigm that in most cases the two molecules will move as little as possible in order to be shape-complementary. There are typically several rotations that will re-

solve an intermolecular collision, and an approach based on mean-field theory [32,52,53] is used to decide which rotations to use to improve the shape complementarity.

For each pairwise intermolecular collision, the bonds in each molecule that can resolve the collision are identified. 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 two values provide the basis for a force measuring the cost of the 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 several cycles of the mean-field optimization, these probabilities are updated and converge to higher values for those rotations that represent a globally optimal choice. When applying these rotations, a maximal number of collisions is resolved with minimal conformational changes in both molecules, without bias to one or the other; details of the mathematics of this procedure are provided in Reference 7.

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 its correlations with other rotations in the system. 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. We do 10 cycles of the optimization, then the probabilities have typically converged to define a near-optimal set of rotations. All feasible rotations are applied in the order provided by the computed probabilities. Since it is likely that not all rotations can be resolved and that new collisions might have emerged, the mean-field optimization process is iterated up to 10 times. Intramolecular collisions are also tolerated, since it is assumed that they will be resolved in a future iteration. The result of the mean-field optimization process is either the exclusion of a molecule as infeasible, if collisions cannot be resolved, or a shape-complementary docking of the two molecules.

Considering binding-site solvation

In order to not bias the search towards known ligands, we typically use the binding site from a ligand-free crystal structure of the target protein for screening. Water molecules are often observed in these crystal structures, and SLIDE can consider tightly bound waters when docking potential ligands. The current approach is to either translate a water molecule, if it overlaps with a ligand atom after docking the ligand into the binding site, or to displace it. A bound water molecule is only displaced if its collisions cannot be resolved by iterative translations, which are computed by summing the translation vectors that resolve each collision between the water molecule and a protein or ligand

atom. SLIDE considers a penalty term for each displaced water when scoring a complex, and only displacements by non-polar ligand atoms are penalized.

To select which protein-bound water molecules to include in the screening and docking, we use a knowledge-based approach to determine those waters likely to be conserved upon ligand binding and to fix a penalty for their displacement. The tool *Consolv* [8], a k-nearest-neighbor classifier, is used to predict which binding-site waters will be conserved and which will be displaced upon ligand binding. *Consolv*'s prediction is based on several features that describe the favorability of the local environment of a water molecule, and its knowledge base is a set of 5542 water molecules taken from 30 independently solved protein structures. Prior to screening, we remove all waters that are predicted to be displaced and for the remaining waters we use *Consolv*'s prediction confidence to scale the penalties for their displacement. To compute the penalty, we count the number of hydrogen bonds that are lost by displacing this water and scale this number by *Consolv*'s prediction confidence (between 50 and 100%).

Scoring a potential ligand

Whenever a collision-free complex is generated, a score is assigned to the ligand based on the number of intermolecular hydrogen bonds and the hydrophobic complementarity of its interface with the protein. If not provided in the protein or ligand structure, the position of the shared hydrogen in each intermolecular hydrogen bond is computed. This position is well-defined for all but the terminal hydrogens in lysine and hydroxyl side chains; for these cases we choose the optimal hydrogen position subject to bonding constraints. All hydrogen bonds with a donor–acceptor distance up to 3.5 Å 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 counted. Intra-protein hydrogen bonds that were broken due to the rotation of a protein side chain, or hydrogen bonds to waters that were displaced upon ligand docking, lower the overall hydrogen-bond count by the number of lost hydrogen bonds. Note that this does not penalize the displacement of a water molecule by a polar ligand atom that preserves the hydrogen bond to the protein. The number of hydrogen bonds lost by displacing water molecules is weighted by *Consolv*'s prediction confidence of their displacement. The final intermolecular hydrogen-bond score between protein P and ligand L, reflecting loss in intra-protein and water-mediated hydrogen bonds, is HBONDS(P,L).

For computing the hydrophobic complementarity value, atomic hydrophilicity values were taken from a statistical survey of hydration of the different atom types in 56 protein structures [49] (hydrophobicity values for protein

atoms were taken from Table II and values for ligand atoms from Table III in Reference 49). The contribution of a single ligand atom is based on the comparison of its hydrophobicity value with the average hydrophobicity of the surrounding protein surface atoms [7]. Given the hydrophobicity $h(a)$ of an atom a, with $h(a) \in [0...635]$ calculated as the average number of hydrations per 1000 occurrences of that atom type (Table II in Reference 49), a value of 0 represents a maximally hydrophobic atom, 635 is maximally hydrophilic, and 317 is intermediate. The hydrophobic complementarity of the contact surface between protein P and ligand L is computed as:

$$\text{HPHOB(P,L)} = \sum_{l_i \in L, \#P_i > 0} \frac{\text{avg}\{h'(l_i), \overline{h}(P_i)\}}{\max\{\text{abs}\,(\,h'(l_i) - \overline{h}(P_i)\,), 32\}}$$

where

$$h'(l_i) = \max\{317 - h(l_i), 0\}$$

considers only the hydrophobic contribution of ligand atoms l_i, since values larger than 317 refer to hydrophilic atoms. The hydrophobicity $\overline{h}(P_i)$ of the protein neighborhood P_i for a single ligand atom l_i is defined as the average hydrophobic contribution of all protein atoms p_j within a distance of 4.0 Å of the ligand atom l_i:

$$\overline{h}(P_i) = \max\{\,(\,317 - \frac{1}{\#P_i} \cdot \sum_{p_j \in P_i} h(p_j)\,)\,, 0\,\}$$

The denominator in each term of the sum describing the hydrophobic score, HPHOB(P,L), is always greater than or equal to 32, which is 10% of the maximum score for a single ligand atom. This ensures that the overall HPHOB(P,L) score is not dominated by a few contacts with very small differences between protein and ligand hydrophobicity.

The scoring function SCORE(P,L) for a collision-free complex is a linear combination of the hydrophobic and hydrogen-bond terms:

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

The relative contribution of these terms was tuned for best fit to the experimentally determined affinities of 89 protein-ligand complexes [51], giving the weight of 1.3:1.0 for the hydrogen-bond term relative to the hydrophobic term.

Results

SLIDE was previously used to screen for potential ligands to a bacterial aspartic protease, the human estrogen receptor, glutathione transferase, and HIV-1 protease [7,48]. Here, we screen for potential ligands to human uracil-DNA glycosylase (coordinates of a complex with 6-amino-uracil provided by C. Mol and J.A. Tainer, The Scripps Research Institute), to the ligand-binding domain of the human progesterone receptor (PDB entry 1a28), and to *E. coli* dihydrofolate reductase (PDB entry 1ra9). The modeling of inducible complementarity and the control of molecular diversity in the set of potential ligands found by SLIDE has been described elsewhere [7], and here we include knowledge-based solvation.

We screened two different databases for the three target proteins:

- A subset of 70 113 compounds taken from the Cambridge Crystallographic Database System (CSD, http://www.ccdc.cam.ac.uk). These are all organic compounds with less than 100 atoms and at least three interaction centers that can be mapped onto template points.
- 105 517 compounds from the NCI database (http://dtp.nci.nih.gov), which were taken from the conformers for the open set of this database as they were prepared by the group of J. Gasteiger using CORINA [54].

We used different approaches for designing the binding-site templates. The smallest template, consisting only of six points, was generated for the progesterone receptor. The interaction points were generated based on the centers of four carbon rings and two ketone oxygen atoms in the progesterone bound to the receptor in PDB entry 1a28, resulting in a template consisting of four hydrophobic and two acceptor points. One water molecule, which interacts with the bound progesterone in this structure was included in the binding site during screening. A search with such a small template is like a pharmacophore-based search, which restricts the set of potential ligands that SLIDE finds to compounds more or less similar to the known ligand, since that ligand and each potential new ligand share at least three interaction centers due to the triangle matching during docking. With this small template for the progesterone receptor, the total screening time for the more than 175 000 compounds was about nine minutes on an Intel Pentium II/450 processor running Solaris 2.7. Figure 1 shows a typical example of the kind of ligand SLIDE found for this screen with the small template. The ligand is 16α,17α-cyclopentenoprogesterone (CSD entry BUBRUJ), which is the known ligand progesterone with a cyclopentene substituent added to its D ring. It obtained a score of 39.9, which ranks it 52th out of the 175 630 screened compounds. The highest-

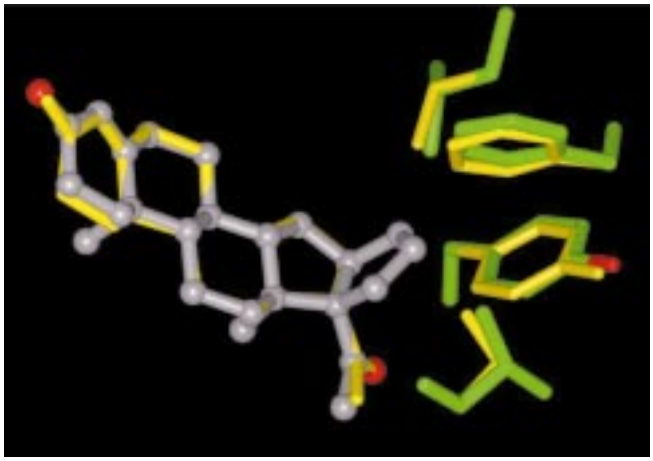

Figure 1. A cyclopenteno-progesterone (grey) from the CSD (entry BUBPUJ) was identified as a potential ligand and docked by SLIDE into the ligand-binding site of the human progesterone receptor (PDB 1a28). The template was based on six interaction centers of the progesterone from the crystal structure, which is shown in yellow tubes and is overlaid virtually exactly by SLIDE's ligand. To fit the additional cyclopentene substituent, four side chains in the receptor underwent minor conformational changes; their native conformation is shown in yellow, and SLIDE's conformation for these side chains is colored by atom type (green: carbon; red: oxygen). Note that the hydrophobic cyclopentene is in contact with hydrophobic groups in the receptor.

ranked progesterone from the CSD received a score of 37.4. A total number of 154 potential ligands were docked into the binding site and obtained a score higher than 35, which is a reasonable cutoff for ligands similar in size and chemistry to the known ligand. Like the progesterone in the crystal structure in PDB 1a28, this ligand makes one water-mediated hydrogen bond. To fit the highly rigid cyclopentene-progesterone into the binding site, adjustments in protein side chains were necessary. The figure shows four side chains in their native conformation together with the final, rotated conformation proposed by SLIDE. Note that only minor rotations were necessary to accommodate the cyclopentene, which demonstrates favorable hydrophobic complementarity with the neighboring side chains in the progesterone receptor.

In the screening for ligands for the human uracil-DNA glycosylase [55,56], the binding site was taken from a crystal structure of a complex with 6-amino-uracil bound deep in the active-site cleft. Twelve water molecules from this structure were predicted as being conserved by *Consolv* and included in the binding site during screening. A known inhibitor for this DNA-repair enzyme is a 84-residue protein that mimics DNA but binds irreversibly to the glycosylase [57]. We used the positions of five H-bond donors and acceptors in the bound 6-amino-uracil as key points (out of which at least one must be

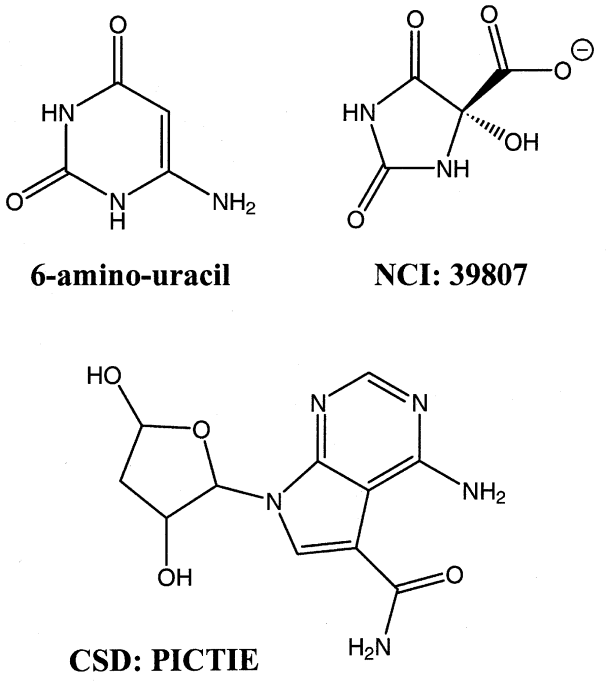

Figure 2. The structure of 6-amino-uracil, the ligand present in the crystal structure of the human uracil-DNA glycosylase used for screening, is shown together with two highly-ranked molecules suggested by SLIDE as potential ligands for this enzyme.

matched by any potential ligand) in a template consisting otherwise of 150 automatically generated interaction points. The cumulative screening time for both databases was slightly over 17 h. Figure 2 shows the structure of 6-amino-uracil and two of SLIDE's ligands, one with obvious resemblance to the known ligand. Figures 3 and 4 show these ligands in SLIDE's binding modes together with key side chains and waters that interact with them. The ligand in Figure 3, CSD entry PICTIE, obtained a score of 32.8 and rank 55 with three water-mediated interactions to the protein, and the ligand in Figure 4, NCI entry 39807 (CAS 6313-89-9), obtained a score of 27.2, which ranked it 384th in the set of 683 potential ligands that were docked by SLIDE and scored higher than 25.0 out of the data set of over 175 000 compounds that were screened. The latter ligand binds similarly to 6-amino-uracil, but shows better complementarity due to additional water-mediated hydrogen bonds to the protein.

In the screening runs against *E. coli* dihydrofolate reductase (PDB entry 1ra9), again a hybrid template design was used. To ensure that all ligands docked by SLIDE interact with the side chains binding pyrimidine in the

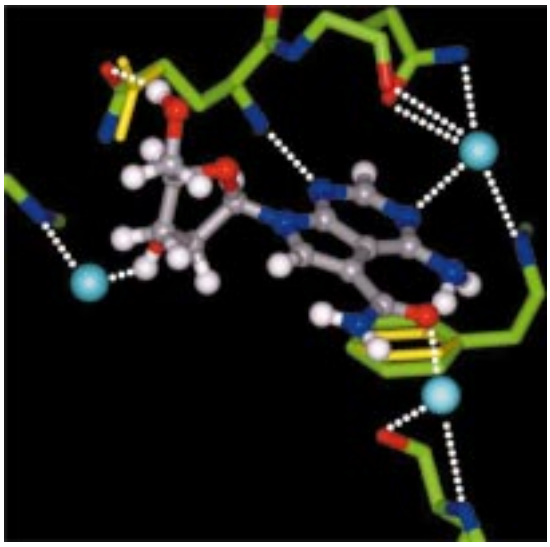

Figure 3. This figure shows 3′-deoxysangivamycin (CSD entry PICTIE), which was docked by SLIDE as a potential ligand into the active site of human uracil-DNA glycosylase, a DNA-repair enzyme. Key side chains and binding-site waters that interact with the ligand are shown, and feasible hydrogen bonds are indicated by dotted lines. Two side chains of the enzyme, a phenylalanine and a glutamine, were rotated by SLIDE to accommodate the ligand and are also shown in their original conformation (yellow).

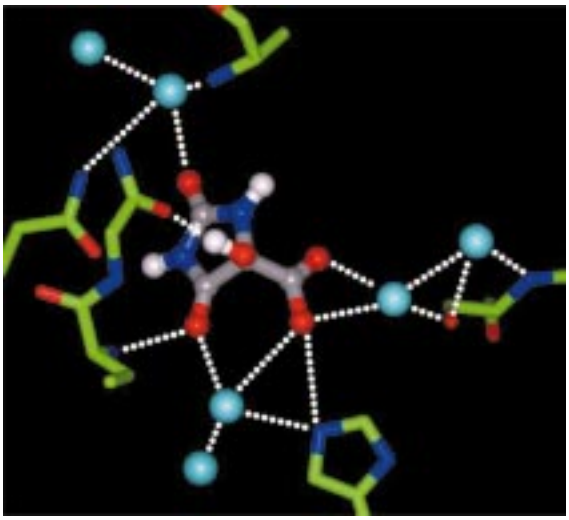

Figure 4. A ligand from the NCI database (entry 39807), docked by SLIDE into the active site of human uracil-DNA glycosylase. It mimics the binding of 6-amino-uracil, the ligand bound in the structure that was used for screening. This ligand shows better complementarity than the original one, due to the additional carboxylate group that interacts with two conserved waters and a histidine side chain.

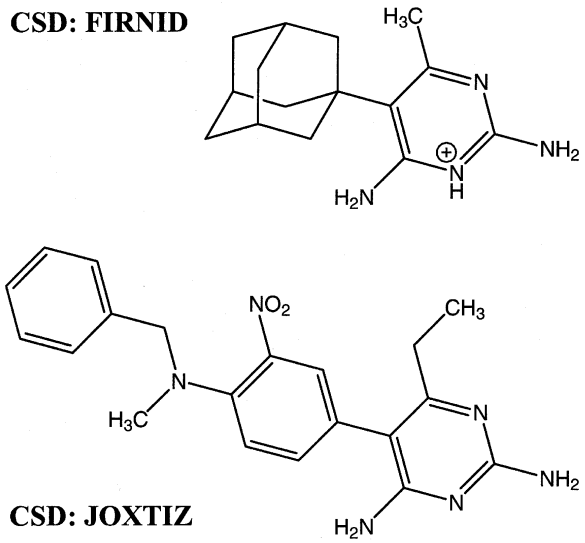

Figure 5. Two CSD compounds that were selected and docked into the active site of dihydrofolate reductase by SLIDE and obtained high scores. Both are known DHFR inhibitors.

known ligands methotrexate and dihydrofolate, two runs of the automatic template generator were done: one to specify 23 key points located in the pyrimidine region of the binding site, and another to fill the remaining part of the binding site with 64 additional template points. Four binding-site waters from the crystal structure of the ligand-free DHFR were predicted to mediate interactions and included during screening. The screening time for the 175 000 compounds was about 14 h. In the set of potential ligands identified by SLIDE were at least two known DHFR inhibitors (Figure 5), and their key interactions are shown in Figures 6 (CSD entry JOXTIZ) and 7 (CSD entry FIRNID). SLIDE's scores for these ligands were 51.1 and 51.9, which ranked them 205th and 141th, respectively. Both ligands place a pyrimidine group in the same site, and the adamantyl-pyrimidine (FIRNID, Figure 7) binds deeper in the corresponding cleft. In the docking of CSD ligand JOXTIZ (Figure 6) a second water molecule fills the non ligand-occupied space. This water was displaced by an amino group in the docking of CSD ligand FIRNID (Figure 7), which replaces the hydrogen bonds to the other water and to the aspartic acid side chain of DHFR, so that this displacement was not penalized in SLIDE's score.

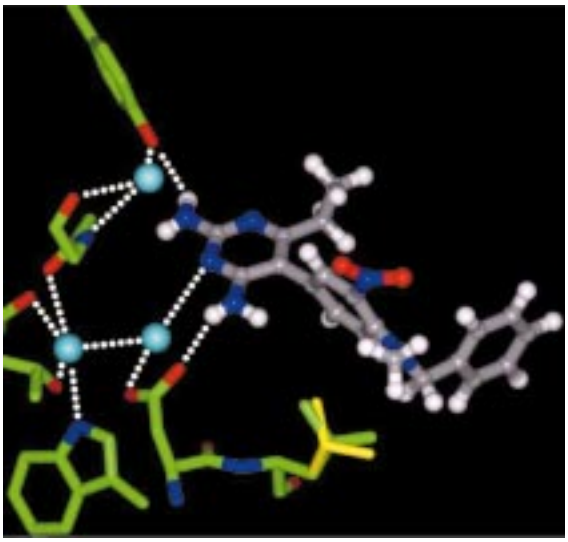

Figure 6. Methylbenzoprim (CSD entry JOXTIZ), a known potent DHFR inhibitor, docked by SLIDE into the active site of *E. coli* dihydrofolate reductase. The ligand was selected by SLIDE out of 175 000 compounds in the screening database. Its pyrimidine group binds in the same cavity as the pyrimidine of the natural ligand, dihydrofolate, which was aided by positioning key template points in that area. The deeper part of this cavity is occupied by two bound water molecules, which were observed in the ligand-free protein structure that was used for screening (PDB 1ra9) and predicted by *Consolv* as being conserved upon ligand binding. One side chain, a leucine, was rotated by SLIDE upon ligand docking, and its original conformation is shown in yellow.

Discussion

SLIDE is an efficient database screening tool, which searches data sets of structures of more than 175 000 organic compounds within minutes, when using a small template, as we did in the case of the progesterone receptor screen, or within several hours, as shown for uracil-DNA glycosylase and dihydrofolate reductase, where we used a more general binding-site template. It accomplishes this by using an efficient multi-level hashing scheme to directly access triplets of feasible interaction points in the binding-site template, onto which triplets of ligand interaction centers are mapped. On one hand this is a straightforward way to compute a transformation of the ligand into the binding site, so that the ligand already makes three favorable interactions, and on the other hand it is also an efficient way to rule out infeasible compounds: all compounds that lack a set of three favorable interactions are discarded before attempting docking into the binding site. For the progesterone receptor with the very specific 6-point template, more than 163 000 compounds, i.e., 93% of the screening databases, never needed to be docked into the binding site

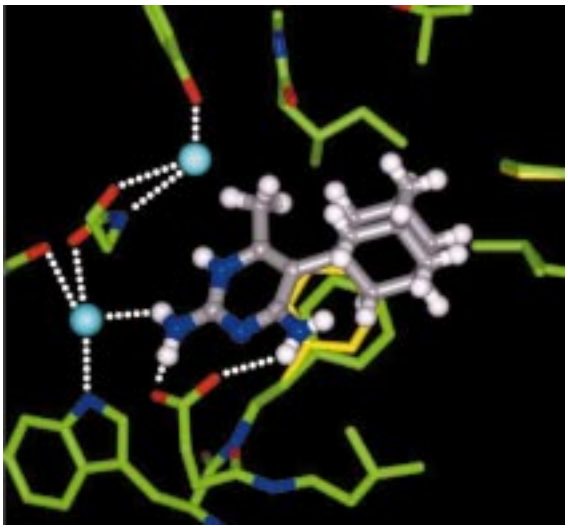

Figure 7. Another known inhibitor for DHFR found by SLIDE in the CSD: 2,4-diamino-5-(1-adamantyl)-6-methylpyrimidine (CSD entry FIRNID), which binds with higher affinity to DHFR than methotrexate [65]. Again, the pyrimidine ring is located in the targeted area of the binding site and makes one water-mediated interaction. The other water molecule located in that area (shown in Figure 6) was displaced by a polar amino group, resulting in no desolvation penalty. Note the hydrophobic complementarity of the side chains in contact with the adamantan (yellow indicates their initial conformations).

for this reason. For more general templates, like the 155-point template for uracil-DNA glycosylase, docking and conformational search were performed for more than 70 000 compounds (40% of the database).

Early in the development of SLIDE, we tried to reduce the complexity of the conformational search for the protein by using a rotamer library for the side chains, which had been done in docking approaches [31,32]. However, in the majority of cases all rotamers cause new collisions, and in a recent study it was shown that side chains close to ligand-binding sites tend to adopt non-rotameric conformations [58]. In most cases, including the examples described above, only minor rotations in both ligand and protein are necessary to generate a shape-complementary interface. These rotations are computed exactly by SLIDE, avoiding costly sampling of rotational angles.

The conformational search is the most computationally complex step of screening with SLIDE. Our model of flexibility is more realistic than that in docking or screening tools that only consider ligand flexibility, since ligand and protein flexibility are treated equally, and the mean-field optimization selects those rotations for resolving collisions that cause the minimal overall distortion for the complex. Full conformational search is not done for the ligand, but rather its database conformation is used as a starting conformation

for docking. Since the structures for potential ligands are taken from crystal structures (CSD) or rule-derived models (NCI), they begin in a low-energy conformation. To deal with cases where the binding conformation of a ligand is very different from the database conformation, the database can be enriched by a series of low-energy conformers for screening [28,59].

Although our scoring function was empirically tuned based on published affinities for PDB complexes, we do not try to predict precise binding affinities in SLIDE. Several empirically derived scoring functions can be found in the literature [51,60–65]. Scoring functions sensitive to small conformation changes may not be appropriate for a screening tool like SLIDE, which cannot perform a conformational search for 100 000 or more ligand candidates. A sensitive scoring function is more appropriate in a fine-docking tool, which must predict differences in binding affinities for very similar conformations of a complex during the search. The scoring function in SLIDE is designed instead to rank the set of all potential ligands based on their complementarity. All examples described above were ranked within the top potential ligands for each target protein. SLIDE includes a web-based interface that enables the user to easily browse through the potential ligands and visualize SLIDE's docking.

The inclusion of binding-site solvation is in accordance with our models of induced fit and scoring. The positions of water molecules in the binding site from the crystal structure of the target protein are analyzed, and those predicted as conserved by *Consolv* are kept. In contrast to a method that precomputes several favorable water positions prior to docking, then picks the best positions to fill gaps between the molecules [43], SLIDE starts with 'real' water molecules and shifts them when they collide with ligand atoms. As in the conformational search, the idea is to start with a reasonable configuration and make only minimal changes, as necessary. If the collision of a water molecule cannot be resolved, the water is displaced and a desolvation penalty term is only applied when a lost hydrogen bond is not replaced by a corresponding protein-ligand interaction.

While SLIDE's docking procedure must be very quick, rather than comprehensive, in order to screen a large number of molecules, its inclusion of protein flexibility and solvation gives SLIDE advantages over other docking procedures. Because of the fast screening time, SLIDE can be used to search very large compound databases for the discovery of novel lead structures, and due to distinguishing a rigid anchor fragment for each screened molecule attached to flexible side chains, it will be straightforward to extend SLIDE for combinatorial screening.

Acknowledgements

We thank Cliff Mol and John Tainer from The Scripps Research Institute for making the coordinates of the human uracil-DNA glycosylase complex available to us. The development of SLIDE was sponsored by the Deutsche Forschungsgemeinschaft (grant Schn 576/1-1 to V.S.), the National Science Foundation (grant DBI-9600831 to L.A.K.), and the American Heart Association (grant 994009IN to L.A.K.).

References

1. Welch, W., Ruppert, J. and Jain, A.N., Chem. Biol., 3 (1996) 449.
2. Rarey, M., Wefing, S. and Lengauer, T., J. Comput.-Aided Mol. Design, 10 (1996) 41.
3. Rarey, M., Kramer, B., Lengauer, T. and Klebe, G., J. Mol. Biol., 261 (1996) 470.
4. Jorgensen, W.L., Science, 254 (1991) 954.
5. Schnecke, V., Swanson, C.A., Getzoff, E.D., Tainer, J.A. and Kuhn, L.A., Proteins Struct. Funct. Genet., 33 (1998) 74.
6. Betts. M.J. and Sternberg, J.E., Protein Eng., 12 (1999) 271.
7. Schnecke, V. and Kuhn, L., Proteins Struct. Funct. Genet., 2000 (manuscript submitted).
8. Raymer, M.L., Sanschagrin, P.C., Punch, W.F., Venkataraman, S., Goodman, E.D. and Kuhn, L.A., J. Mol. Biol., 265 (1997) 445.
9. Kuntz, I.D., Blaney, J.M., Oatley, S.J., Langridge, R. and Ferrin, T.E., J. Mol. Biol., 161 (1982) 269.
10. Meng, E.C., Shoichet, B.K. and Kuntz, I.D., J. Comput. Chem., 13 (1992) 505.
11. Shoichet, B.K. and Kuntz, I.D., Protein Eng., 6 (1993) 723.
12. Böhm, H.-J., J. Comput.-Aided Mol. Design, 6 (1992) 61.
13. Böhm, H.-J., J. Comput.-Aided Mol. Design, 6 (1992) 593.
14. Böhm, H.-J., J. Comput.-Aided Mol. Design, 10 (1996) 265.
15. Rarey, M., Kramer, B. and Lengauer, T., J. Comput.-Aided Mol. Design, 11 (1997) 369.
16. Jones, G., Willett, P. and Glen, R.C., J. Mol. Biol., 245 (1995) 43.
17. Jones, G., Willett, P., Glen, R.C., Leach, A.R. and Taylor, R., J. Mol. Biol., 267 (1997) 727.
18. Morris, G.M., Goodsell, D.S., Huey, R. and Olson, A.J., J. Comput.-Aided Mol. Design, 10 (1996) 293.
19. Morris, G.M., Goodsell, D.S., Halliday, R.S., Huey, R.S., Hart, W.E., Belew, R.K. and Olson, A.J., J. Comput. Chem., 19(14) (1998) 1639.
20. Kuntz, I.D., Meng, E.C. and Shoichet, B.K., Acc. Chem. Res., 27 (1994) 117.
21. Fischer, D., Lin, S. L., Wolfson, H. L. and Nussinov, R., J. Mol. Biol., 248 (1995) 459.
22. Ruppert, J., Welch, W. and Jain, A.N., Protein Sci., 6 (1997) 524.
23. Oshiro, C.M., Kuntz, I.D. and Scott Dixon, J., J. Comput.-Aided Mol. Design, 9 (1995) 113.
24. Eisen, M.B., Wiley, D.C., Karplus, M. and Hubbard, R.E., Proteins Struct. Funct. Genet., 19 (1994) 199.
25. Shoichet, B.K., Stroud, R.M., Santi, D.V., Kuntz, I.D. and Perry, K.M., Science, 259 (1993) 1445.
26. Böhm, H.-J., J. Comput.-Aided Mol. Design, 8 (1994) 623.

27. Gschwend, D.A., Sirawaraporn, W., Santi, D.V. and Kuntz, I.D., Proteins Struct. Funct. Genet., 29 (1997) 59.
28. Lorber, D.M. and Shoichet, B.K., Protein Sci., 7 (1998) 938.
29. Shoichet, B.K., Leach, A.R. and Kuntz, I.D., Proteins Struct. Funct. Genet., 34 (1999) 4.
30. Makino, S., Ewing, T.J.A. and Kuntz, I.D., J. Comput.-Aided Mol. Design, 13 (1999) 513.
31. Leach, A.R., J. Mol. Biol., 235 (1994) 345.
32. Jackson, R.M., Gabb, H.A. and Sternberg, M.J.E., J. Mol. Biol., 276 (1998) 265.
33. Totrov, M. and Abagyan, R., Proteins Struct. Funct. Genet., Supplement 1 (1997) 215.
34. Sandak, B., Wolfson, H.J. and Nussinov, R., Proteins Struct. Funct. Genet., 32 (1998) 159.
35. Knegtel, R.M.A., Kuntz, I.D. and Oshiro, C.M., J. Mol. Biol., 266 (1997) 424.
36. Wassserman, Z.R. and Hodge, C.N., Proteins Struct. Funct. Genet., 24 (1996) 227.
37. Apostolakis, J., Plückthun, A. and Caflisch, A., J. Comput. Chem., 19 (1998) 21.
38. Mangoni, M., Roccatano, D. and Di Nola, A., Proteins Struct. Funct. Genet., 35 (1999) 153.
39. Poornima, C.S. and Dean, P.M., J. Comput.-Aided Mol. Design, 9 (1995) 500.
40. Poornima, C.S. and Dean, P.M., J. Comput.-Aided Mol. Design, 9 (1995) 513.
41. Ladbury, J.E., Chem. Biol., 3 (1996) 973.
42. Sanschagrin, P.C. and Kuhn, L.A., Protein Sci., 7 (1998) 2054.
43. Rarey, M., Kramer, B. and Lengauer, T., Proteins Struct. Funct. Genet., 34 (1999) 17.
44. Lawrence, M.C. and Davis, P.C., Proteins Struct. Funct. Genet., 12 (1992) 31.
45. Toyoda, T., Brobey, R.K.B., Sano, G., Horii, T., Tomioka, N. and Itai, A., Biochem. Biophys. Res. Commun., 235 (1997) 515.
46. Horvath, D., J. Med. Chem., 40 (1997) 2412.
47. Burkhard, P., Taylor, P. and Walkinshaw, M.D., J. Mol. Biol., 277 (1998) 449.
48. Schnecke, V. and Kuhn, L.A., In Procs. ISMB 99, 7th Int. Conf. on Intelligent Systems for Molecular Biology, AAAI Press, Menlo Park, CA, 1999, pp. 242–251.
49. Kuhn, L.A., Swanson, C.A., Pique, M.E., Tainer, J.A. and Getzoff, E.D., Proteins Struct. Funct. Genet., 23 (1995) 536.
50. Duda, R.O. and Hart, P.E., Pattern Classification and Scene Analysis, John Wiley & Sons, New York, NY, 1973.
51. Eldridge, M.D., Murray, C.W., Auton, T.R., Paolini, G.V. and Mee, R.P., J. Comput.-Aided Mol. Design, 11 (1997) 425.
52. Koehl, P. and Delarue, M., J. Mol. Biol., 239 (1994) 249.
53. Koehl, P. and Delarue, M., Curr. Opin. Struct. Biol., 6 (1996) 222.
54. Sadowski, J., Gasteiger, J. and Klebe, G., J. Chem. Inf. Comput. Sci., 34 (1994) 1000.
55. Mol, C.D., Arvai, A.S., Slupphaug, G., Kavli, B., Alseth, I., Krokan, H.E. and Tainer, J.A., Cell, 80 (1995) 869.
56. Parikh, S.S., Mol, C.D., Slupphaug, G., Bharati, S., Krokan, H.E. and Tainer, J.A., EMBO J., 17 (1998) 5214.
57. Mol, C.D., Arvai, A.S., Sanderson, R.J., Slupphaug, G., Kavli, B., Krokan, H.E., Mosbaugh, D.W. and Tainer, J.A., Cell, 82 (1995) 701.
58. Heringa, J. and Argos, P., Proteins Struct. Funct. Genet., 37 (1999) 44.
59. Knegtel, R.M.A., Bayada, D.M., Engh, R.A., von der Saal, W., van Geerestein, V.J. and Grootenhuis, P.D.J., J. Comput.-Aided Mol. Design, 13 (1999) 167.
60. Böhm, H.-J., J. Comput.-Aided Mol. Design, 8 (1994) 243.
61. Böhm, H.-J., J. Comput.-Aided Mol. Design, 12 (1998) 309.
62. Jain, A.N., J. Comput.-Aided Mol. Design, 10 (1996) 427.

63. Head, R.D., Smythe, M.L., Opera, T.I., Waller, C.L., Green, S.M. and Marshall, G.R., J. Am. Chem. Soc., 118 (1996) 3959.
64. Murray, C.W., Auton, T.R. and Eldridge, M.D., J. Comput.-Aided Mol. Design, 12 (1998) 503.
65. Mügge, I. and Martin, Y.C., J. Med. Chem., 42(5) (1999) 791.
66. Cody, V., Sutton, P.A. and Welsh, W.J., J. Am. Chem. Soc., 109 (1987) 4053.