Journal of Computer-Aided Molecular Design, **16:** 883–902, 2002.
KLUWER/ESCOM

Distilling the essential features of a protein surface for improving protein-ligand docking, scoring, and virtual screening

Maria I. Zavodszky[1,*], Paul C. Sanschagrin[1,*,†], Rajesh S. Korde[1,3] & Leslie A. Kuhn[1,2]
[1]*Protein Structural Analysis and Design Laboratory, Department of Biochemistry and Molecular Biology,* [2]*Center for Biological Modeling, and* [3]*Department of Computer Science and Engineering, Michigan State University, East Lansing, MI 48824, USA*

MS received 6 Sept 2002; accepted in final form 16 Dec 2002

Key words: drug design, glutathione S-transferase, molecular recognition, protein-ligand interactions, SLIDE, template-based design, thrombin, virtual screening, 3-D pharmacophore

Summary

For the successful identification and docking of new ligands to a protein target by virtual screening, the essential features of the protein and ligand surfaces must be captured and distilled in an efficient representation. Since the running time for docking increases exponentially with the number of points representing the protein and each ligand candidate, it is important to place these points where the best interactions can be made between the protein and the ligand. This definition of favorable points of interaction can also guide protein structure-based ligand design, which typically focuses on which chemical groups provide the most energetically favorable contacts. In this paper, we present an alternative method of protein template and ligand interaction point design that identifies the most favorable points for making hydrophobic and hydrogen–bond interactions by using a knowledge base. The knowledge-based protein and ligand representations have been incorporated in version 2.0 of SLIDE and resulted in dockings closer to the crystal structure orientations when screening a set of 57 known thrombin and glutathione S–transferase (GST) ligands against the *apo* structures of these proteins. There was also improved scoring enrichment of the dockings, meaning better differentiation between the chemically diverse known ligands and a ∼15,000-molecule dataset of randomly-chosen small organic molecules. This approach for identifying the most important points of interaction between proteins and their ligands can equally well be used in other docking and design techniques. While much recent effort has focused on improving scoring functions for protein-ligand docking, our results indicate that improving the representation of the chemistry of proteins and their ligands is another avenue that can lead to significant improvements in the identification, docking, and scoring of ligands.

Abbreviations: PDB – Protein Data Bank, CSD – Cambridge Structural Database, GST – glutathione S-transferase, HIV – human immunodeficiency virus, RMSD – root mean square deviation.

Introduction

Protein function centers on the specific recognition and binding of other molecules. Identifying the structural and chemical interactions that are most important in protein-ligand binding can help us understand how proteins screen the ambient molecules to select their molecular partners. This knowledge can also be applied to develop new, protein–specific ligands for disease therapy and help elucidate the roles of the increas-

*These authors contributed equally to this work.
†Current address: Institute of Pharmaceutical Chemistry, University of Marburg, D–35032 Marburg, Germany.
Corresponding author: Leslie A. Kuhn, Protein Structural Analysis and Design Laboratory, Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI 48824; Phone: 517–355–3455; Fax: 517–353–9334; E–mail: kuhn@agua.bch.msu.edu.; web site: http://www.bch.msu.edu/labs/kuhn

ing number of proteins with known 3-dimensional structure but unknown function. Our ligand docking and screening software, **SLIDE** (**S**creening for **L**igands by **I**nduced-fit **D**ocking, **E**fficiently) [1–3] models flexible protein-ligand interactions based on steric complementarity combined with hydrogen bonding and hydrophobic interactions. SLIDE, as used here, can also provide a test bed for modeling the importance of different factors in molecular recognition.

SLIDE efficiently eliminates infeasible ligand candidates using geometric indexing and distance geometry filtering on discrete representations of the protein and ligand candidates. Approximately 100,000 small molecules can be screened and docked in a day on a typical desktop workstation. Because SLIDE models protein side-chain and ligand flexibility, it can identify and correctly dock diverse, known ligands into the ligand-free conformation of the binding site for a variety of proteins (e.g., subtilisin, cyclodextrin, glycosyltransferase, uracil DNA glycosylase, rhizopuspepsin, HIV protease, estrogen receptor, and Asn tRNA synthetase) [1–4, Schnecke and Kuhn, Proteins, in review]. Scoring of the docked protein-ligand complex by SLIDE is based on the number of hydrogen bonds and the hydrophobic complementarity between the ligand and its protein environment. The main steps involved in screening with SLIDE are shown in Figure 1, and the algorithm has been described in detail elsewhere [1].

The evolution of SLIDE and its protein and ligand representations

Specitope was the precursor of SLIDE and focused on peptide-protein docking [4]. The binding site was represented as a steric shell of protein surface atoms, plus a few points (typically, six or fewer) where atoms from the peptide (which could represent the binding epitope of a larger protein) could be placed and make good hydrogen-bond interactions with the protein. The peptide structural database used for screening was a set of ~155,000 successive, overlapping peptides from a non-redundant (<25% identity) version of the Protein Data Bank (PDB) [5]. To be docked by Specitope, a peptide needed to match at least a subset I of the J template points. Usually peptides 3–5 residues in length were screened, and a subset of their interaction points was required to match all points of the 4 or 5-point protein template in order to dock the peptide. The placement of these template points was typically based on known ligand interactions, and therefore represented a pharmacophore model of desired ligand features for binding to the protein.

While hydrogen-bond interactions were emphasized initially because they provide orientational specificity to the protein-ligand interaction, the role of hydrophobic contacts in the enthalpy of binding was also recognized and implemented in the first version of SLIDE. SLIDE [1–3] can screen and dock organic small molecules (including peptides) using either a pharmacophore model, specifying a few points of interaction, or an unbiased representation of the entire ligand-binding site. For the unbiased representation, the binding site is filled with a large number of points (20,000–60,000). From each point, the type of protein interaction that can be made is then determined, and the points are labeled accordingly as hydrophobic, hydrogen-bond donor, acceptor, or donor/acceptor (donor and acceptor, due to either interacting with a donor and acceptor group in the protein, e.g., hydroxyl group, or being able to interact with separate acceptor and donor groups). Points at which a ligand cannot make significant hydrophobic or hydrogen-bond interactions are discarded from the set, while those with similar chemistry labels are clustered in 3D to reduce the number of template points to a manageable number of 100–150 points. Each ligand candidate is similarly represented as a set of hydrogen-bond donor or acceptor atoms and hydrophobic centers (defined in the initial version of SLIDE as the centroids of rings containing only carbon and hydrogen atoms). SLIDE can then test each ligand candidate for docking to the protein, by assessing whether one of its triplets of interaction points matches one or more of the triplets of protein template points in terms of donor/acceptor/hydrophobic labels and triangle shape. All ligand and template triplets are exhaustively tested for matching, and every such match that leads to a docking orientation that is evaluated by scoring. Protein-bound water molecules can also be included in SLIDE docking, and are retained or displaced according to their probability of being conserved upon ligand binding (as determined, for example, by using the Consolv software [6]). These interfacial water molecules contribute to the hydrogen-bond score between the docked ligand and protein, with the second term in the scoring function measuring hydrophobic complementarity between the two molecules [1].

The ~100-point unbiased template of the ligand binding site using hydrogen-bonding and hydrophobic interaction points has allowed SLIDE to work well on a number of protein systems [1–4, Schnecke and

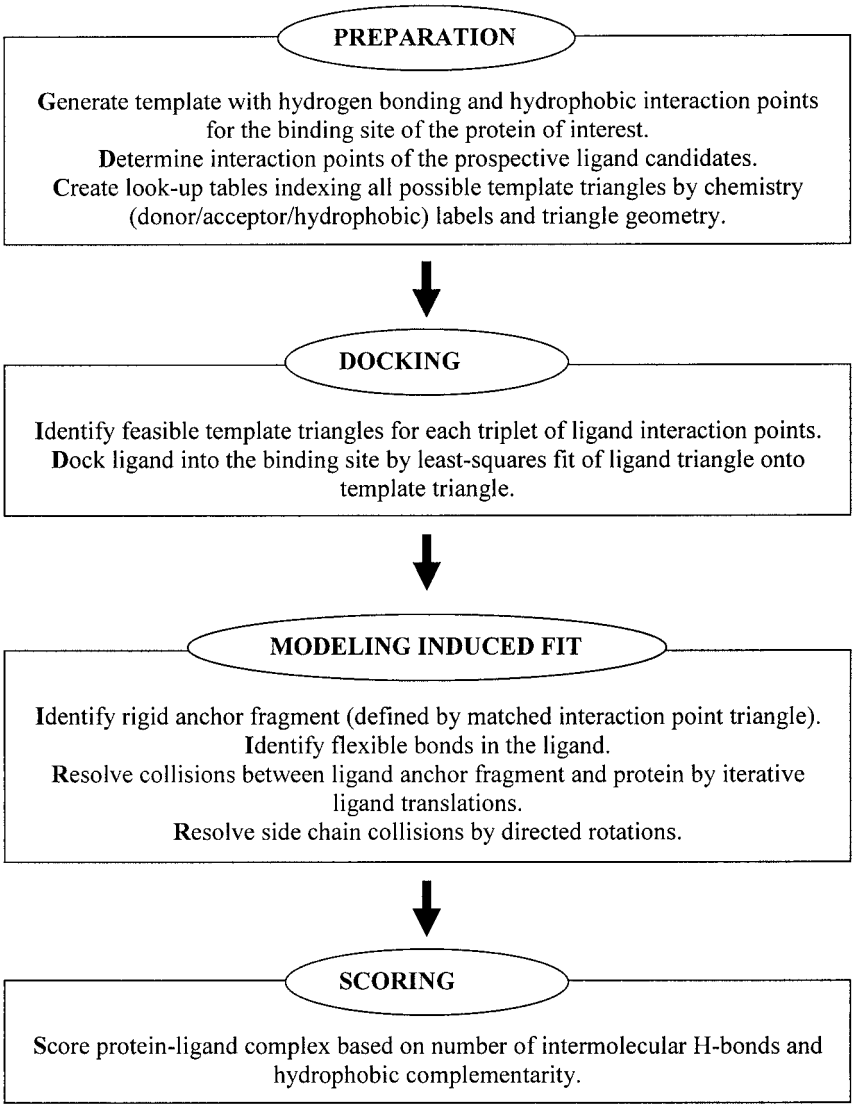


Figure 1. An overview of the SLIDE screening and docking algorithm. See Schnecke & Kuhn [1] for more details.

Kuhn, Proteins, in review]. However, it was found that when a known ligand failed to dock using SLIDE 1.0, it was typically because the matching between triplets of protein template points and ligand interaction points was not close enough. This was due to a combination of slight shifting of hydrogen-bonding points relative to positions where they could match the ligand, and to incomplete sampling of hydrophobic surface in the protein. Simply increasing the tolerance of matching led to much longer run times, due to finding many more (not necessarily better) dockings. In the SLIDE paradigm, protein and ligand flexibility are equally balanced, and the molecules flex as little as needed in order to form a complex free of van der Waals overlaps. This assumption of minimal side-chain motion upon complex formation appears to be true for many proteins, and is being directly addressed in ongoing work.

Improving the success rate of docking known ligands to a protein structure that does not already have correct side-chain conformations for that ligand (e.g., an '*apo*' structure of the protein, solved in the ligand-free state) was the motivation for the present work, which is aimed at defining protein templates that capture optimal points for interacting with the protein. Knowledge bases of hydrogen-bonding geometry

around protein groups [7, 8] allow us to focus now on optimal (rather than just feasible) positions for hydrogen bonding. Significantly hydrophobic positions at the protein surface can also be distinguished from the background level of solvent-exposed carbon atoms, based on the local enhancement of hydrophobic atoms. Similarly, the interaction points on ligand candidates can be sampled to have similar density and chemistry to the hydrophobic and hydrogen-bonding assignments in the protein template. While this work has been driven by the aim to improve the modeling of protein recognition through docking in SLIDE, this representation of key interacting groups in proteins and ligand candidates is also expected to be useful for other docking methods, and to provide a focus on optimal interactions to make in structure-based protein and ligand design.

Other approaches for discrete representation of protein binding sites

Reduced representations of protein binding sites have been developed by other groups for use in modeling protein recognition. Typically, the protein's binding site is discretized to a set of 100 or fewer interaction points to enable fast comparison between the protein and each ligand. Many of these methods use reduced representations to aid in matching the protein and ligand surfaces. The initial, computationally complex search of the 6 degrees of rotational and translational freedom of the ligand relative to the protein is reduced to a problem of matching a set of N points on the ligand to the best-matching subset of N points from M points on the protein. N and M typically must be small due to the factorial complexity of the number of ways of matching N points to a larger set of M points. In the case of SLIDE, 3-point subsets of N interaction points on the ligand are tested for matching to all 3-point subsets of a set of typically 100–150 template points representing the protein.

In the case of DOCK [9], the earliest protein-ligand docking technique, the binding site is filled with spheres, whose centers serve as possible ligand atom positions. Chemical properties or other characteristics can be associated with the spheres, and a sphere with a particular characteristic can only be matched with a ligand atom of complementary character [10]. Jones et al. [11] identify solvent-accessible hydrogen-bond donor and acceptor atoms within the active site of the protein and associate virtual points with each hydrogen and lone pair of these atoms, enabling the genetic algorithm employed by GOLD [12] to transform the ligand into the binding site by minimizing the least-square distance between protein virtual points and similarly defined ligand virtual points. Ruppert et al. [13] coat the protein's binding site surface with probes of three types, hydrophobic, acceptor and donor, which could potentially interact with the protein. These probes can serve as potential alignment points for ligand atoms and are scored to represent the probe's affinity for the protein. High affinity probe-clusters identify sticky spots, or regions of strongest potential binding. This method can also be used to find binding pockets on the surface of a protein. FlexX [14] uses a multi-layered representation of the binding site adopted from its predecessor LUDI [47]: interaction types are arranged on three levels depending on their directionality, with H-bonds being the most directional at level three and hydrophobic interactions the least directional at level one. Each group capable of forming an interaction is characterized by an interaction center and a surface, the latter being approximated by a finite number of points. Ligand interaction centers are superimposed over these points and aligned, giving preference to higher-level interaction points over lower-level ones. In an approach related to that of SLIDE, Fischer et al. [48, 49] describe the surfaces of the protein and ligand by a set of critical points and their normals, then apply geometric indexing to dock the ligands into the protein by matching the critical points and vectors.

Grid-based representations are also used to map favorable points of interaction with proteins. In preparation for docking with AutoDock [15], the protein binding site is placed in a grid. The protein-ligand pair-wise interaction energies are precalculated at each grid point for each possible ligand atom type and are stored in a look-up table for use during the docking simulation. The Grid technique developed by Boobbyer et al. [16] calculates for each grid point an empirical energy designed to represent the interaction energy of a chemical probe group, such as a carbonyl oxygen or an amine nitrogen atom, around the target molecule. This function is used to determine the sites where ligands may bind to the target, such as a protein.

Finally, knowledge bases of the frequency of pair-wise atomic or functional group interactions deduced from the crystallographic protein structures in the PDB [17] and small organic molecule structures in the Cambridge Structural Database (CSD) [18] can be used to map favorable sites for ligand interactions with proteins. Relibase [19], a database system of protein-

ligand interactions from the PDB, has been used to derive atomic potentials between protein and ligand atom groups for use in DrugScore [20]. DrugScore can then calculate 'hotspots' for interactions with different ligand atom types, which are displayed as contour maps within the binding site [21]. Similarly, the SuperStar software [22], based on pair-wise interaction frequencies in the CSD database, can calculate hotspots for the binding of 16 probe atom types to proteins. A recent paper analyzes how the interaction maps developed from PDB versus CSD data complement each other [23]. Another knowledge-based approach was taken by Moreno and Leon [24] to describe the binding site for DOCK: templates of attached points or contact points are constructed for each amino acid type, representing the geometry of the interactions observed in the different protein-ligand complexes from the PDB.

In this paper, we show how a knowledge-based approach for describing favorable interaction sites on proteins and ligands can improve the performance of SLIDE when a database of known ligands combined with a random selection of CSD compounds is screened against two protein targets, thrombin and glutathione S-transferase.

Methods

Template generation to represent binding sites in proteins

Two methods to generate a template for the binding site of interest were initially implemented in SLIDE: small, biased, pharmacophore-like templates, and unbiased, grid-based approaches. The biased template is based on known ligand binding modes and consists of coordinates of ligand atoms making hydrogen bonds or engaging in hydrophobic interactions with the protein of interest, as seen in crystal structures of protein-ligand complexes. This pharmacophore-like representation of binding determinants is biased towards known ligands, and is especially appropriate when the aim is to identify other molecules that make similar interactions. When the goal instead is to identify new classes of ligands or help define the ligand specificity for protein structures with unknown functions, an unbiased, thorough representation of the potential ligand-binding site is preferable. Therefore, SLIDE also has an option to automatically generate an unbiased template based on a ligand-free structure of the protein. To generate an unbiased template in version 1 of SLIDE, the binding site was filled with a large number of points, initially located on a fine grid with a spacing of 0.3–0.7 Å (Figure 2A; ref. [1] and Schnecke & Kuhn (2002) *Proteins*, in review). Initial experiments with random placement of the points showed significant under-representation of some areas in the binding site, so the grid-based approach was adopted instead. Only points located 2.5 to 5.0 Å from the nearest protein atom were kept. Each point was then checked to determine if it could serve as a hydrogen bond donor, acceptor, or form a hydrophobic interaction with the protein, and was either labeled as such, or eliminated from the set. All points of the same class were then clustered using complete linkage clustering to reduce the number of template points to 150 or fewer.

Because grid placement of hydrophobic and hydrogen-bond points is not always optimal with respect to protein interactions, here we describe the development of a knowledge-based approach to placing points in an unbiased template. Geometrically favored subsites for ligand hydrogen-bonding atoms are assigned based on the distance and angle to protein hydrogen-bonding partners (Figure 2B). After identifying the protein atoms capable of hydrogen bonding, a number of template points are placed at and around the optimal hydrogen bonding position for each of these atoms, using the geometries shown in Figure 3. The template points belonging to one hydrogen-bonding protein atom are separated by $\sim$1 Å and are placed at a distance of 2.9 Å (for Asp, Glu, Lys, Thr and Tyr side chains) or at 3.0 Å (for all the other side chains and backbone oxygen and nitrogen) from the protein donor or acceptor atom. The parameters for optimal hydrogen bonding geometry were taken from the literature [7, 8]. The points are labeled as donors, acceptors or donor/acceptors, depending on the role an atom at this position would have in hydrogen bonding to the protein. A donor template point, for example, is located near an acceptor protein atom, such as a backbone carbonyl oxygen, and represents a favorable placement for a ligand atom acting as an H-bond donor. A donor/acceptor point is defined in two cases: when a ligand atom at that point could make favorable hydrogen bonds with separate hydrogen-bond donor and acceptor atoms in the protein, or when it could interact with a group that both donates and accepts hydrogen bonds (e.g., –OH in the side chains of Ser, Thr, or Tyr). Template points that overlap with those belonging to neighboring atoms (template points separated by $<$1 Å) are clustered and relabeled, and

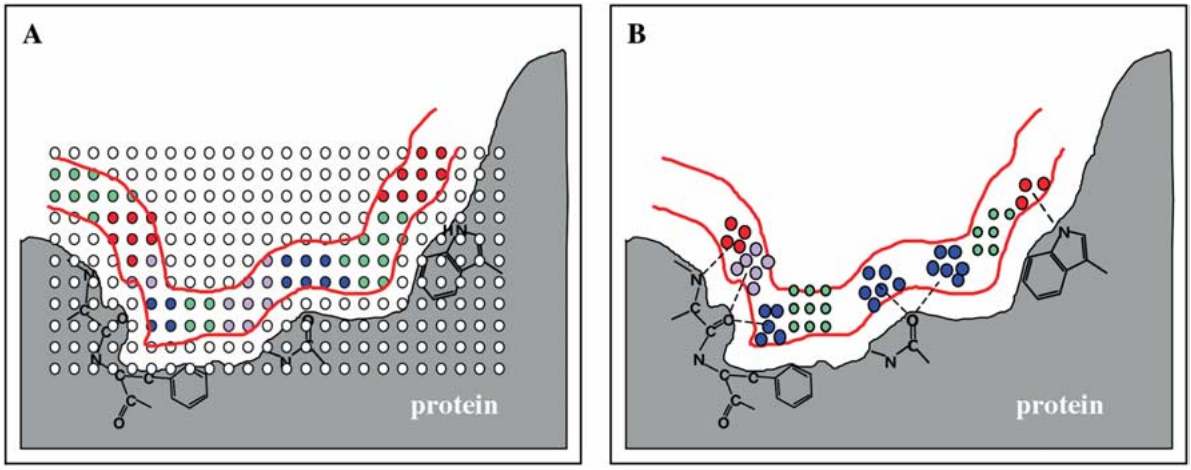


Figure 2. Comparing the (A) grid-based and (B) knowledge-based template generation methods. Template points are generated on a grid in version 1 of SLIDE. The method implemented in SLIDE, version 2 uses a knowledge base to define points where optimal protein–ligand interactions can be made, based on points where the ligand could make optimal hydrogen bonds and hydrophobic interactions with the protein. Template points are colored according to their type: green for hydrophobic, red for acceptor, blue for donor, and purple for donor and/or acceptor points.

points closer than 2.5 Å to a protein atom are discarded. The clustering of hydrogen-bonding template points reduces the number of points by about 10–25%. Points generated by the clustering of a donor and an acceptor point are relabeled as donor/acceptors.

Hydrophobic template points are generated using a grid for initial point placement, as before, but the criteria have been updated for which of these points should be included to represent favorable sites for ligand interactions. Hydrophobic points are those grid points with a hydrophobic enhancement score of at least 3. This score is defined as the number of carbon atoms minus the number of hydrophilic atoms, such as oxygen or nitrogen, within a spherical shell of radius 2.5–5.0 Å from the template point in question. The cutoff value of 3 was found to define the significantly hydrophobic protein surface patches that complement the hydrophobic groups in ligands for a number of 3D protein-ligand complexes.

After they are generated separately, the H-bonding and hydrophobic template points are merged into one template that can be used for docking with SLIDE. If the total number of template points is much larger than 150 (a practical upper limit given the combinatorics of matching ligand interaction points with template points), then the complete linkage clustering feature can be used to reduce neighboring points of the same class to a single point, the cluster centroid. Complete linkage clustering has the desirable features that the clusters can be defined to not exceed a certain diameter (helping control the separation between centroids), and they are guaranteed to be the most densely occupied set of clusters for that diameter [25]. Typically we use a clustering threshold of 4 Å, resulting in hydrophobic template points separated by about 2 Å. When a clustering threshold of x Å is used with complete linkage clustering (where x is typically chosen between 2 and 4 Å), the average distance between the final template points (the centroids of each cluster) is very close to $x/2$. For any uniformly distributed set of points clustered by complete linkage, the centroids of the clusters will be separated by half the cluster diameter (called the clustering threshold in this work), on average.

Ligand interaction points

Hydrophobic ligand interaction points are assigned using a rule-based approach summarized in Figure 4. These rules are designed to place an interaction point at approximately every 1.5 hydrophobic carbon atoms in hydrophobic chains and around the circumference of hydrophobic rings. This density of hydrophobic interaction points is commensurate with the spacing of hydrophobic points in the protein template, using the default clustering criteria. For this approach, carbon and sulfur atoms bonded only to carbon, sulfur or hydrogen atoms are considered to be hydrophobic. Other atoms are taken as hydrophilic. Hydrogen bonding interaction points in the ligand are identified as atoms capable of accepting or donating hydrogen

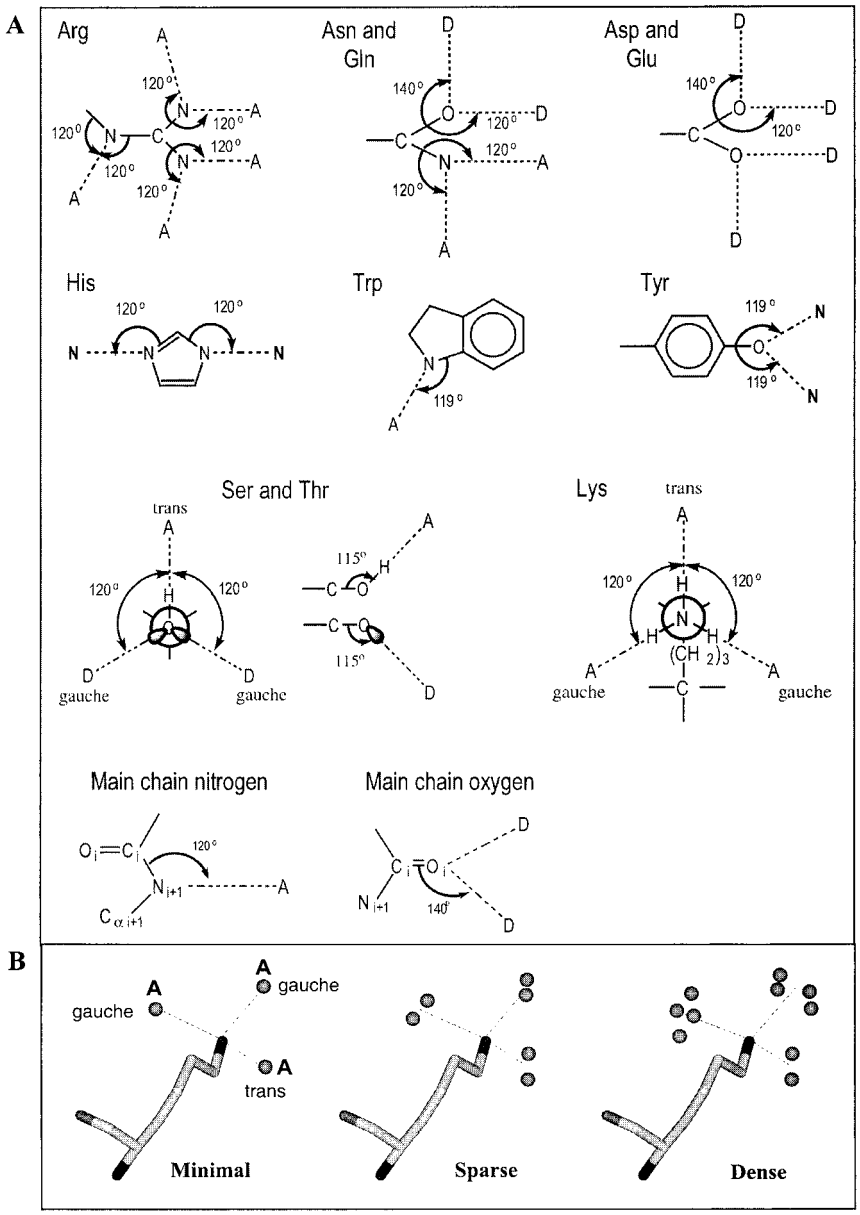


Figure 3. (A) Placement of optimal hydrogen-bonding template points in SLIDE. For each polar side chain, the optimal placement of hydrogen-bond donor (D), acceptor (A) and donor and/or acceptor (N) template points is shown with respect to the donor and acceptor atom positions in the side chain. These template points represent positions where a ligand atom matching the template point can form a hydrogen bond with the protein. A ligand atom matching a donor/acceptor (N) template point must be either a donor or acceptor, or both. These optimal distances and angles are consensus values describing preferred geometries [7, 8] observed in high resolution protein structures from the PDB. The positions of hydrogen atoms in the protein are not assumed in template point placement, since these positions are not available in most crystal structures. Instead, the most favorable positions for hydrogen-bonding partners is measured relative to the geometry of the covalent bonds in the side chains (e.g., trans and gauche positions for Lys), as found from analysis of crystallographic data [7, 8]. (B) Three-dimensional example of template point placement relative to a Lys side chain. The template points defined for minimal, sparse, and dense templates are shown, along with the most-preferred distance and angle for hydrogen bonding, as shown above. The default template specification in SLIDE is dense, and thus there are more possible H-bond template point matches, each of which is shifted by a small amount relative to the optimal position and still allows formation of a near-optimal hydrogen bond between the matched ligand atom and the protein. Sparse and minimal hydrogen-bond templates are alternatives that can be used to decrease the number of hydrogen-bond template points when the complete template for a protein, including hydrophobic points, exceeds the practical limit of about 150 points.

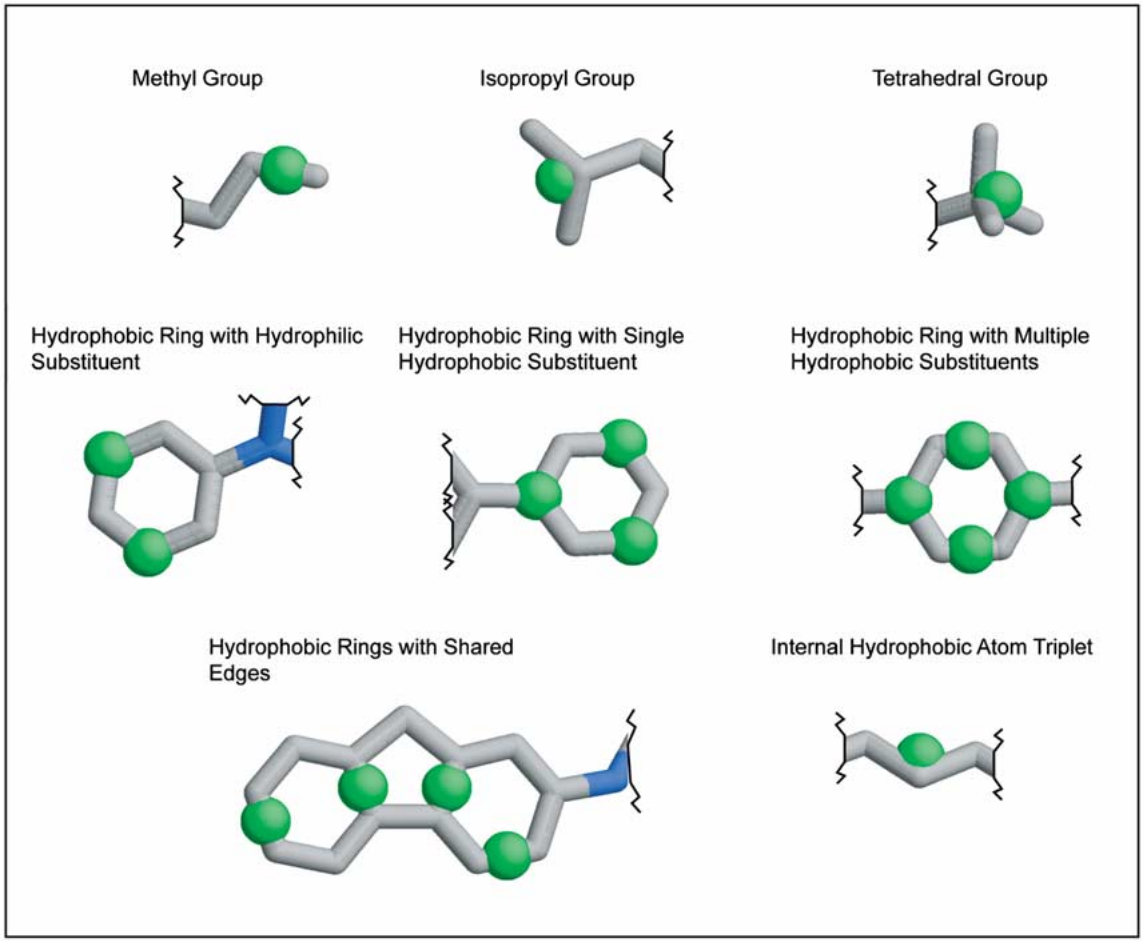


Figure 4. Summary of rules for hydrophobic interaction point assignment. The goal is to place a point at approximately every 1.5 carbon atoms, which is commensurate with the default spacing of hydrophobic points in the template. Hydrophobic interaction points are denoted by green spheres, carbon atoms by gray tubes, and nitrogen atoms, representing hydrophilic atoms, by blue tubes.

bonds, based on the SYBYL atom types in the mol2 file (described at http://www.tripos.com).

Ligand databases

A combined database of known ligands from the PDB and a subset of 14,691 randomly selected CSD compounds was assembled for alpha-thrombin and π-class human GST. The CSD database was prescreened to exclude molecules with excessive molecular weight as well as those containing unusual atoms. The nonredundant subset of known ligands for thrombin contained 42 molecules taken from thrombin–ligand complexes available from the PDB. To screen for ligands to GST, 15 known ligands with PDB crystal structures in complex with human GST were selected. For both thrombin and GST, ligands from crystal structures with a resolution of 3.0 Å or better were included in the known ligand test set. If a ligand was found in multiple structures, the one with the highest resolution was chosen. To ensure that SLIDE can appropriately model the side-chain conformational changes necessary in nature when proteins bind their ligands, structures of thrombin and π-GST determined crystallographically with ligand-free active sites (*apo* structures) were used as the targets for screening and docking (PDB code 1vr1 for thrombin [26] and PDB code 16gs for GST [27]). This also avoided the docking bias that is implicit in redocking experiments (when the ligand-bound structure of the protein, already conformationally biased for that ligand, is used as the basis for docking). Because interactions in a mutant protein structure might change the favored orientation of a ligand relative to its orientation in the wild-type

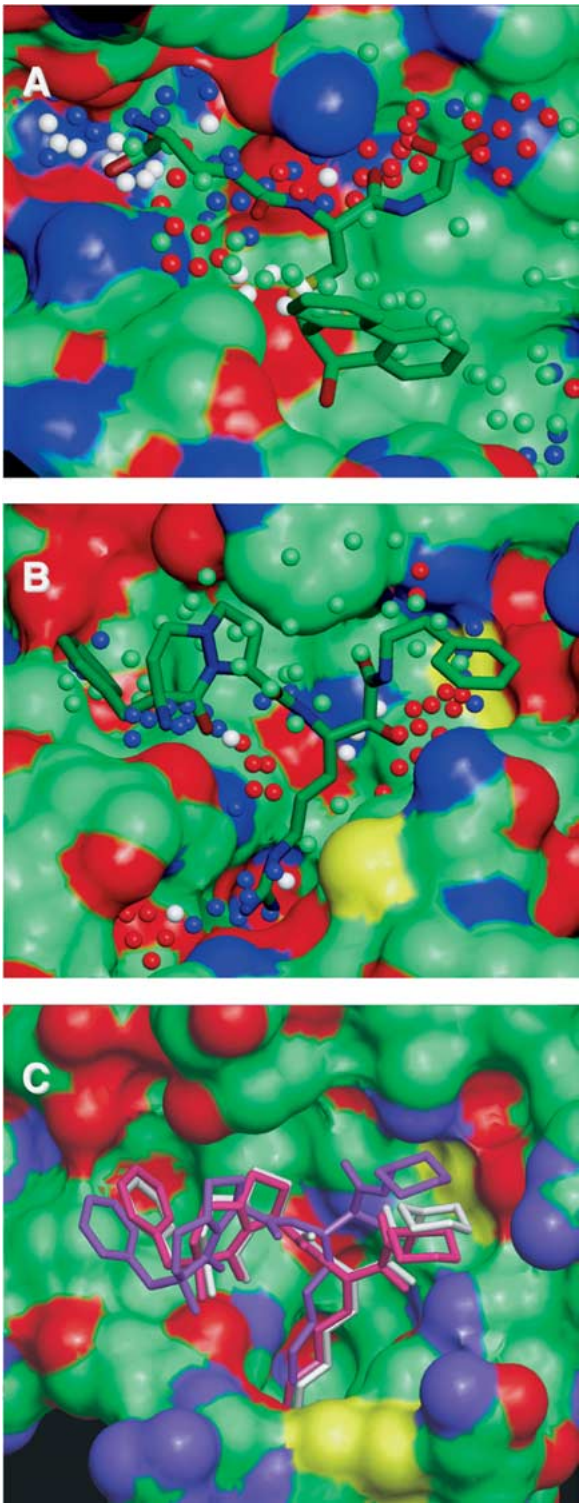
protein (and therefore not allow fair comparison of the docking with the crystallographic complex), ligands from complexes containing a mutant version of π-GST were excluded from the analysis. Four of the GST crystal complexes (PDB codes 13gs, 20gs, 21gs and 2gss [28, 29]) contained two ligands: glutathione, and a smaller hydrophobic ligand bound to the xenobiotic subsite of the active site. Only the hydrophobic ligands from these structures were included in the screening dataset, and glutathione from the GST-glutathione complex 1aqw [30] was used as the single representation of this ligand in the screening set.

In order to focus the large number of orientations that can result from the screening/docking process on productive binding modes, selected template points can be labeled as key points. Template points from parts of the binding site known to be critical for tight and/or specific binding can be marked as key points, and any docking must then include a match to one (not all) of these points. This ensures that docked molecules will at least partially occupy the targeted site. For thrombin, points in the specificity pocket within 5.0 Å of the carboxyl oxygens of Asp 189 were selected as key points. Assignment of key points in the GST binding site was more challenging, as it is made up of two subsites, one for hydrophobic ligands and the other for glutathione, which is fairly polar. SLIDE was run twice on the known ligands in the case of GST: initially with key hydrogen bonding points in a 5.0 Å radius sphere around the side chain hydroxyl oxygen of Ser 65 in the deepest pocket of the glutathione binding site, to capture ligands that can bind to this polar site, then with key hydrophobic points in the area between Tyr 108 and Phe 8, the xenobiotic (hydrophobic) binding site. Screening against the CSD ligands was done using the first set of key points in the glutathione-binding pocket, which includes both hydrophobic and hydrogen-bonding interactions.

Using key points is mainly a convenient way to ensure that ligands make interactions in the deep pockets of the binding site, rather than making less favorable, superficial interactions. Placing key points in the deepest pocket of the thrombin active site would be useful, in the absence of any knowledge of thrombin ligand structure or chemistry, to ensure the absence of a significant, destabilizing cavity in the complex. Ensuring that deep pockets are filled is also a widely used approach in structure-based drug design to increase ligand binding affinity and specificity. For GST, the use of key points allows a convenient analysis of ligand binding to the hydrophobic binding site versus binding to the glutathione site, without specifying which ligands favor which site, or how they bind. We can therefore assess the accuracy of ligand specificity as well as docking for GST: hydrophobic ligands should fit and score well in the hydrophobic site, and score poorly if they also dock into the polar site (when key points are included there, instead), and vice versa for the polar ligands. This allows a more sophisticated analysis for GST, making use of both its binding sites. Key points can also hurt docking results, because not all ligands may make one of the chosen interactions and therefore would either not be docked at all, or would be forced to dock by making a non-native interaction. Thus, using key points is only recommended for predicting the docking of ligands if there is a strong indication as to the location of a key binding pocket within the larger binding site (as is obvious in the case of thrombin, which has a funnel-shaped active site). Another appropriate occasion for including key points is in design applications, when the intent is to control which pocket or binding site is to be probed by a database of ligand candidates or fragments.

Evaluation of these protein and ligand representations in ligand screening and docking

Templates for thrombin and GST were created both with the grid-based and the knowledge-based template generation methods; the knowledge-based templates are shown in Figure 5A,B. Sets of interaction points for the known ligands and the CSD compounds were also identified using both assignment methods. SLIDE was used to screen the known ligands and the CSD compounds against thrombin and GST, first using the grid-based template and the original ligand interaction points, and in a second experiment using the knowledge-based template and the new ligand interaction points. The two methods for representing the protein target and ligand candidates were evaluated in two ways. First, they were evaluated based on how well SLIDE, using these protein and ligand representations, could reproduce the known ligand positions in the structure of the protein-ligand complex. This involved docking the ligands into an *apo* structure of the protein, with side-chain positions not already optimized for the ligands. Secondly, they were evaluated by how well known ligands and nonspecific molecules (in our case, CSD compounds) could be differentiated. The heavy atom root-mean-square-deviation (RMSD) was used to compare the docked ligand orientation with its crystal structure position. Because



scoring remains a major challenge in the field [31–33], and to ensure that the results were not very dependent on the particulars of the scoring function, the dockings were also evaluated using DrugScore as well as the SLIDE score. While SLIDE scores the protein-ligand complex based on the number of hydrogen bonds and the hydrophobic complementarity [1], DrugScore [20] calculates protein-ligand interaction energies employing a knowledge-based potential that reflects the frequency of pair-wise atomic distances observed in protein-ligand complexes from the PDB. The known ligands and CSD compounds were each docked, scored, and sorted by score. Then, the enrichment in selecting known ligands from the random database, based on scores, was calculated as the percentage of known ligands captured as a function of the percentage of the database screened, where the top 1% of the database represented the top scoring compounds.

Results

All four combinations of template and ligand interaction point design were evaluated: grid-based template with original interaction points, grid-based template with new interaction points, knowledge-based template with original interaction points, and knowledge-based template with new interaction points. Both the knowledge-based template design and the new interaction point assignments resulted in improvements individually, but the most improvement was seen upon combining the two. For brevity, we present only the

Figure 5. New knowledge-based template and the corresponding improvement in docking quality. The Connolly solvent-accessible molecular surfaces [45] of the GST (A) and thrombin (B) active sites are shown, color–coded according to atom type (green – carbon, blue – nitrogen, red – oxygen, yellow – sulfur). Known ligands from PDB structures 2pgt (A) and 1a5g (B) were docked into the binding site with SLIDE and are shown as tubes, also colored according to atom type. The template points are represented as spheres, with blue representing hydrogen-bond donor points, red for acceptors, white for donor/acceptors, and green for hydrophobic interaction points. (C) Comparing the docked orientations to the crystal structure position of a β-strand mimetic inhibitor (PDB code 1a46) in the binding site of thrombin. The crystal structure position of the ligand is shown in white, and the docked orientation using the knowledge-based method is in magenta (RMSD 1.03 Å), while the docking obtained with the grid-based method is shown in blue (RMSD 2.48 Å). This is representative of the improvement in docking quality observed for the thrombin and GST ligands in general. The view into the thrombin active site is slightly shifted relative to that in the previous panel.

results obtained with the two most relevant combinations: grid-based protein template with original ligand interaction point assignments (subsequently referred to as method 1, and corresponding to the implementation in SLIDE v. 1), and knowledge-based template with new interaction points (method 2, as now implemented in SLIDE v. 2).

Thrombin

The 42 known thrombin ligands used in this study are listed in Table 1, along with the PDB code of the crystallographic complexes from which they were obtained. SLIDE docked 36 ligands into the binding site of thrombin using both methods. The ligands with no scores listed could not be docked, due to unresolved steric overlaps with the apo-active site thrombin structure (1vr1) except for the case of benzamidine (PDB code 1dwb), which was not docked, due to the unusual proximity of its three interaction points (the two amide N's, and any pair of its three benzene-ring hydrophobic points, were all $<$2.5 Å apart). This caused benzamidine dockings to not meet a default parameter setting in SLIDE which ensures that the minimum edge of any triangle being matched is $>$2.5 Å. This is intended to ensure that ligand dockings are complementary to more than a very local region of the binding site. (If the binding site is small, or the goal is to find small molecules that match very locally, this parameter can be changed easily.) Among the docked ligands, 27 had a heavy atom RMSD smaller than 2.0 Å compared to the crystal structure orientation using method 1, while 33 such dockings were obtained with method 2. As shown in Figure 6A, the dockings were generally closer to the crystal structure position using method 2, as reflected by their lower RMSD values. The mean RMSD for thrombin ligand dockings was 1.83 Å using method 1, and 1.28 Å using method 2. An example of the typical improvement in the quality of docking for thrombin ligands is shown in Figure 5C.

Enrichment plots of the percentage of known ligands docked as a function of the percentage of the database screened (CSD plus thrombin ligands) are shown for SLIDE scores (Figure 7) and DrugScores (Figure 8). Higher enrichment is gained with method 2 compared to method 1, independently of the scoring function used (indicated by a shift to the left of the new curve compared to the original one in panel A in Figures 7 and 8). This means that more known ligands are returned by SLIDE among the top scoring CSD compounds. Based on the SLIDE score, for example, the percentage of the known ligands that ranked among the top scoring 100 molecules increased from 38% (16 out of 42) to 64% (27 out of 42). The results are very similar when using DrugScores: 67% of the known molecules (28 of the 42) ranked among the top scoring 100 molecules with method 2, compared to 33% (14 of the 42) using method 1.

The score distributions also show that the knowledge-based protein and ligand representations provide a better separation between known ligands and randomly chosen CSD compounds for both the SLIDE scores (Figure 7BC) and DrugScores (Figure 8BC). The difference between the mean SLIDE scores of the known ligands and CSD compounds increased from 20.7 score units to 27.1 score units when method 1 was replaced by method 2. DrugScore also mirrors a better discrimination between known ligands and CSD compounds when the knowledge-based method is used.

Glutathione S-transferase

SLIDE was able to find a collision-free orientation for 14 of the 15 known ligands in the active site of GST with method 2, while 13 were docked using method 1 (Table 2). The ligand from the crystal complex 19gs could not be docked for the same reason described for benzamidine in the previous section, whereas the reason for failure of chlorambucil (21gs) to dock was the existence of unresolved steric clashes with the protein. Method 2 resulted in better dockings (lower RMSD values), as illustrated in Figure 6B by the majority of points falling under the diagonal. Only one of the 14 docked ligands had a lower RMSD when method 1 was used, two were docked about equally well, while 10 were docked closer to their crystal structure position with method 2. The number of known ligands docked with an RMSD less than 2.0 Å doubled from five to ten, and the mean RMSD between crystal structure and docked positions decreased from 2.15 Å to 1.00 Å upon introducing the knowledge-based method. The four hydrophobic ligands, shown by the crystal complexes to bind to the hydrophobic subsite of GST (13gs, 20gs, 21gs, 2gss), were docked incorrectly (RMSD $>$ 5.0 Å) when polar template points were used as key points. This is not surprising given that these ligands must make interactions in a region different from where the key points were assigned. However, their docking improved substantially when hydrophobic key points were used in the second run with either method of template generation

Table 1. Comparison of SLIDE scores, DrugScore scores and RMSDs of known thrombin ligands docked into the active site of thrombin by SLIDE using the original template and ligand interaction point generation methods in comparison with the knowledge-based method. The original DrugScore scores are divided by 10^4 to give a comparable order of magnitude to SLIDE scores. For both scores, a larger absolute value means a better score, and 'best' corresponds to the docking with the highest score or lowest RMSD (columns 5–8 and 9–10, respectively).

PDB code of thrombin-ligand complex	Ligand name	DrugScore $\times 10^{-4}$ crystal str. position	SLIDE score crystal str. position	Best SLIDE score		Best DrugScore $\times 10^{-4}$		Best RMSD (Å)	
				grid-based	knowledge-based	grid-based	knowledge-based	grid-based	knowledge-based
1a2c	Aeruginosin298-A	−41.8	60.4	29.2	23.6	−37.4	−29.8	8.33	8.57
1a3b	Borolog1	−56.5	49.8	50.5	55.0	−48.3	−56.0	1.22	0.30
1a3e	Borolog2	−32.7	32.0	–	–	–	–	–	–
1a46	Beta-strand mimetic inhibitor	−62.0	57.3	43.3	61.2	−37.3	−52.2	2.49	1.03
1a4w	Ans-Arg-2ep-Kth	−48.0	71.5	59.8	61.1	−49.5	−50.5	1.39	0.52
1a5g	Bic-Arg-Eoa	−70.9	60.1	55.8	74.0	−65.5	−62.0	0.56	0.97
1a61	Mol-Arg-Lom	−58.1	58.6	54.1	64.3	−38.1	−54.0	1.90	1.02
1ad8	MDL103752	−72.6	31.8	–	–	–	–	–	–
1ae8	Eoc-D-Phe-Pro-Azalys-Onp	−47.5	36.9	45.5	53.1	−44.6	−46.4	0.32	0.65
1afe	Cbz-Pro-Azalys-Onp	−38.5	21.8	35.4	40.5	−36.2	−33.6	1.26	1.05
1aht	p-Amidino-phenyl-pyruvate	−37.2	30.9	22.7	29.9	−31.6	−34.9	2.40	0.81
1ai8	$PhCH_2OCO$-D-Dpa-Pro-boroMpg	−55.1	44.0	–	–	–	–	–	–
1aix	$PhCH_2OCO$-D-Dpa-Pro-boroVal	−51.4	38.1	–	–	–	–	–	–
1awf	GR133487	−44.8	56.8	44.0	28.0	−43.0	−34.2	1.56	11.29
1awh	GR133686	−44.5	37.0	47.5	–	−46.4	–	0.90	–
1ay6	Hmf-Pro-Arg-Hho	−57.2	72.1	55.1	66.8	−48.5	−54.4	1.01	0.71
1b5g	Bcc-Arg-Thz	−56.7	37.9	32.8	57.8	−30.1	−56.9	8.71	0.40
1ba8	Pms-Ron-Gly-Arg	−51.5	58.5	51.6	57.5	−43.2	−46.4	0.98	0.51
1bb0	Pms-Ron-Gly-3ga	−50.7	54.8	51.2	66.8	−44.7	−50.5	1.14	0.65
1bcu	Proflavin	−30.8	24.0	21.9	25.7	−30.5	−27.0	3.90	2.16
1bhx	SDZ 229-357	−47.2	49.3	44.2	55.6	−43.4	−51.2	0.75	0.53
1bmm	BMS-186282	−50.3	53.9	47.8	57.5	−41.2	−52.6	0.71	0.38
1bmn	BMS-189090	−55.5	45.7	38.4	43.5	−48.5	−55.9	0.83	0.29
1dwb	Benzamidine	−26.7	15.9	–	–	–	–	–	–
1dwc	MD-805 (Argatroban)	−42.6	52.8	57.3	45.2	−43.3	−43.3	0.56	1.04
1dwd	NAPAP	−60.5	46.9	43.5	52.7	−52.5	−64.3	0.97	0.44
1fpc	Ans-Arg-Epi (DAPA)	−40.4	46.9	57.0	53.2	−40.3	−39.0	0.94	0.90
1hdt	Alg-Phe-Alo-Phe-CH_3 (BMS-183507)	−53.8	53.0	67.1	61.3	−55.2	−53.6	0.43	0.65
1lhc	Ac-D-Phe-Pro-boroArg-OH	−57.3	52.8	37.2	46.2	−45.2	−52.8	1.18	0.67
1lhd	Ac-D-Phe-Pro-boroLys-OH	−51.1	41.3	32.7	48.5	−35.8	−46.9	1.29	0.77
1lhe	Ac-D-Phe-Pro-boro-N-butyl-amidino-Glycine-OH	−59.9	54.2	51.3	55.1	−49.6	−56.3	1.05	0.71
1lhg	Ac-D-Phe-Pro-borohomoornithine-OH	−46.8	42.8	–	37.3	–	−35.2	–	1.32
1nrs	Leu-Asp-Pro-Arg	−51.9	52.8	43.5	57.7	−34.6	−46.6	1.49	0.75
1ppb	PPACK	−50.9	43.8	46.2	44.4	−32.5	−50.7	1.73	0.71
1tbz	Dpn-Pro-Arg-Bot	−62.8	40.2	41.8	66.5	−35.8	−51.4	4.11	1.10
1tmb	Cyclotheonamide A	−68.4	72.3	54.2	57.9	−38.9	−60.4	2.61	1.00
1tmt	Phe-Pro-Arg	−54.7	47.6	48.9	51.4	−36.1	−49.7	2.18	0.54
1tom	Methyl-Phe-Pro-amino-cyclohexylglycine	−49.2	36.1	43.1	45.5	−39.1	−44.9	1.01	0.74
1uma	N,N-dimethylcarbamoyl-alpha-azalysine	−18.9	20.7	12.9	23.0	−15.1	−22.0	2.80	0.94
3hat	Fibrinopeptide A mimic	−51.0	50.4	46.5	39.2	−29.0	−46.7	1.62	0.89
7kme	SEL2711	−60.4	49.1	52.4	60.7	−63.2	−65.1	0.54	0.40
8kme	SEL2770	−59.0	59.2	47.1	58.6	−54.8	−61.5	1.06	0.79

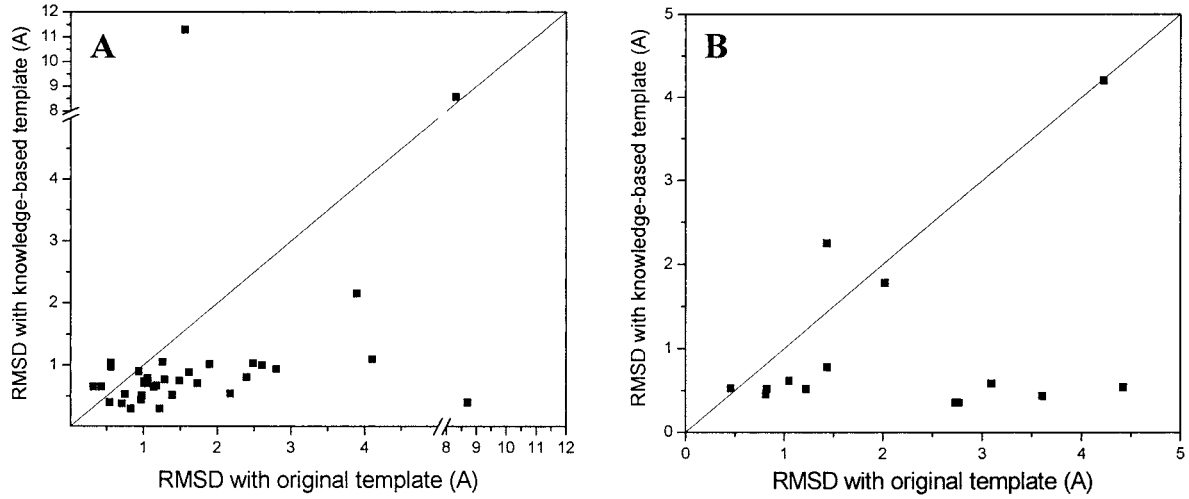


Figure 6. Comparing the RMS deviations between the docked orientations of known ligands and their crystal structure positions resulting from the original and the knowledge-based methods of template and ligand interaction point generation in the case of thrombin (A) and GST (B). Ligands docked better (with lower RMSD) with the knowledge-based method are represented by points below the diagonal line. The significant outlier in (A) with RMSD ∼11.3 Å is a ligand with a neutral side chain occupying the S1 specificity pocket of thrombin in the x-ray structure of the protein-ligand complex (PDB code 1awf [46]). This is a case in which the inclusion of key points can lead to misdocking. The atypical lack of hydrogen-bonding atoms in the portion of the 1awf ligand that binds to the S1 specificity pocket led to the inability of SLIDE to match this part of the molecule to at least one key point in the S1 pocket. The ligand was thus rotated by SLIDE about 180° compared to its crystal structure position, in order to satisfy the key point matching requirement by placing another, polar side chain into the S1 pocket.

and interaction point assignment. Hydrophobic template points can be used as key points for docking smaller sets of ligands to a protein, but this is not a practical alternative when screening large databases. Since matching three template points is sufficient for docking with SLIDE, using hydrophobic key points when screening a large database can result in docking a very large number of small, relatively nonspecific, hydrophobic molecules. They could later be eliminated based on their scores, of course, but this would still result in a considerable increase of the running time and output volume.

Only the results of the first run (with hydrogen bonding key points) were used to construct the enrichment plots for GST (Figure 9A). For brevity, only the enrichment plot for DrugScores is shown; the results were substantially similar using SLIDE scores. DrugScores indicate that more of the known ligands were retrieved among the top scoring molecules (Figure 9A), meaning improved enrichment was achieved with method 2 compared to method 1 for GST. When the SLIDE scoring function was used, 73% of the known ligands (11 out of 15) were ranked among the top scoring 100 of all docked molecules when using method 2, compared to 60% (9 out of 15) among the top 100 with method 1. Using DrugScore, the percentage of the known ligands ranking among the top scoring 100 of all the docked molecules increased from 33% (5 out of 15) to 60% (9 out of 15).

The distribution of scores obtained for the docked known ligands and CSD compounds to GST are shown in Figures 9B and C. The difference between the mean scores of the GST ligands and randomly selected GST molecules increased due to the introduction of the knowledge-based method, independently of the scoring function applied: the means were separated by an additional 7.5 score units using SLIDE scores, and by an additional 9.4×10^4 units using DrugScore. Although the standard deviations of the DrugScores and SLIDE scores also increased, the increased separation of the means was roughly three times greater than the increase in standard deviations.

Discussion

Because the computation time increases nearly exponentially with the size of the template, a compromise must be reached such that the most important features of the binding site are captured with the smallest possible number of template points. Using a knowledge-based approach for identifying the most favorable hydrogen-bonding subsites in the binding site of the protein proved to be superior over grid-based sampling

Table 2. Comparison of SLIDE scores, DrugScore scores and RMSD's of known GST ligands docked into the active site of GST (PDB code 16gs). SLIDE was used with the grid-based template and original ligand interaction point generation methods in comparison with the knowledge-based method. The original DrugScore scores are divided by 10^4 to give a comparable order of magnitude to SLIDE scores. For both scores, a larger absolute value means a better score, and 'best' corresponds to the docking with the highest score or lowest RMSD (columns 5–8 and 9–10, respectively).

PDB code of GST-ligand complex	Ligand name	DrugScore $\times 10^{-4}$ crystal str. position	SLIDE score crystal str. position	Best SLIDE score		Best DrugScore $\times 10^{-4}$		Best RMSD (Å)	
				grid-based	knowledge-based	grid-based	knowledge-based	grid-based	knowledge-based
10gs	Benzylcysteine phenylglycine	−50.2	40.7	42.9	44.6	−23.3	−49.1	2.73	0.36
12gs	S-nonyl-cysteine	−49.7	46.8	40.5	52.2	−24.4	−44.9	2.77	0.36
13gs*	Sulfasalazine	−30.7	34.6	20.6 (38.4)	32.2 (50.9)	−12.4 (−27.3)	−21.7 (−27.7)	8.76 (2.02)	6.42 (1.78)
18gs	1-(S-glutathionyl)-2,4-dinitrobenzene	−41.8	44.5	38.8	47.9	−35.9	−36.2	1.06	0.64
19gs	Phenol-1,2,3,4-tetrabromophthalein-3′,3″-disulfonic acid ion	−12.6	17.7	–	–	–	–	–	–
1aqv	p-Bromobenzylglutathione	−47.7	43.8	40.4	48.6	−18.2	−43.6	3.61	0.44
1aqw	Glutathione	−36.6	31.8	24.6	37.5	−28.0	−32.7	0.82	0.46
1aqx	S-(2,3,6-trinitrophenyl)cysteine	−46.5	37.4	37.5	42.4	−32.0	−49.6	1.44	0.78
1pgt	S-hexylglutathione	−46.1	49.2	33.1	46.4	−42.9	−38.3	0.46	0.53
20gs*	Cibacron blue	−22.2	21.7	19.5 (50.3)	44.0 (56.2)	−22.1 (−24.6)	−28.2 (−25.9)	5.51 (0.83)	5.48 (0.52)
21gs*	Chlorambucil	−22.0	25.4	– (37.6)	12.1 (35.7)	– (−19.6)	−18.5 (−23.0)	– (4.22)	9.11 (4.21)
2gss*	Ethacrynic acid	−19.7	26.9	10.1 (30.7)	20.3 (36.1)	−14.6 (−24.6)	−18.5 (−19.1)	6.21 (1.43)	5.33 (2.25)
2pgt	(9R,10R)-9-(S-glutathionyl)-10-hydroxy-9,10dihydrophenanthrene	−54.8	55.6	35.6	68.5	−28.9	−53.4	4.42	0.54
3gss	Ethacrynic acid-Glutathione conjugate	−52.3	75.0	58.3	71.4	−31.3	−47.7	1.22	0.52
3pgt	Glutathione conjugate of (+)-Anti-BPDE	−51.3	66.1	59.7	64.1	−36.0	−50.6	3.09	0.59

*Ligands that are mainly hydrophobic in character and bind to the hydrophobic subsite of GST. The numbers in parentheses next to these ligands are the scores and RMSD values obtained in a separate run, when hydrophobic template points from their respective binding subsite were selected as key points. In the other screening runs for GST, hydrogen bonding template points were selected as key points. Each ligand is required to match only one of the key points, allowing the screen to focus on ligands predicted to bind in the correct region within the active site.

followed by the selective retention of points where ligand atoms could act as hydrogen-bond donors or acceptors. More known ligands could be docked closer to their known crystal structure positions for both thrombin and GST using the knowledge-based method of template and ligand interaction point generation.

Docking experiments usually return multiple docked orientations per ligand. Ideally, the scoring function will indicate the one closest to the crystal structure by giving it the highest score. Also, when a large database is screened, the scoring function should be able to discriminate between promising ligand candidates and artificial hits. Using the assumption that most CSD compounds are unlikely to be ligands of thrombin and of GST, the ability of SLIDE scores and DrugScores to discriminate between known ligands and CSD compounds was tested. The enrichment plots calculated with both scoring methods showed

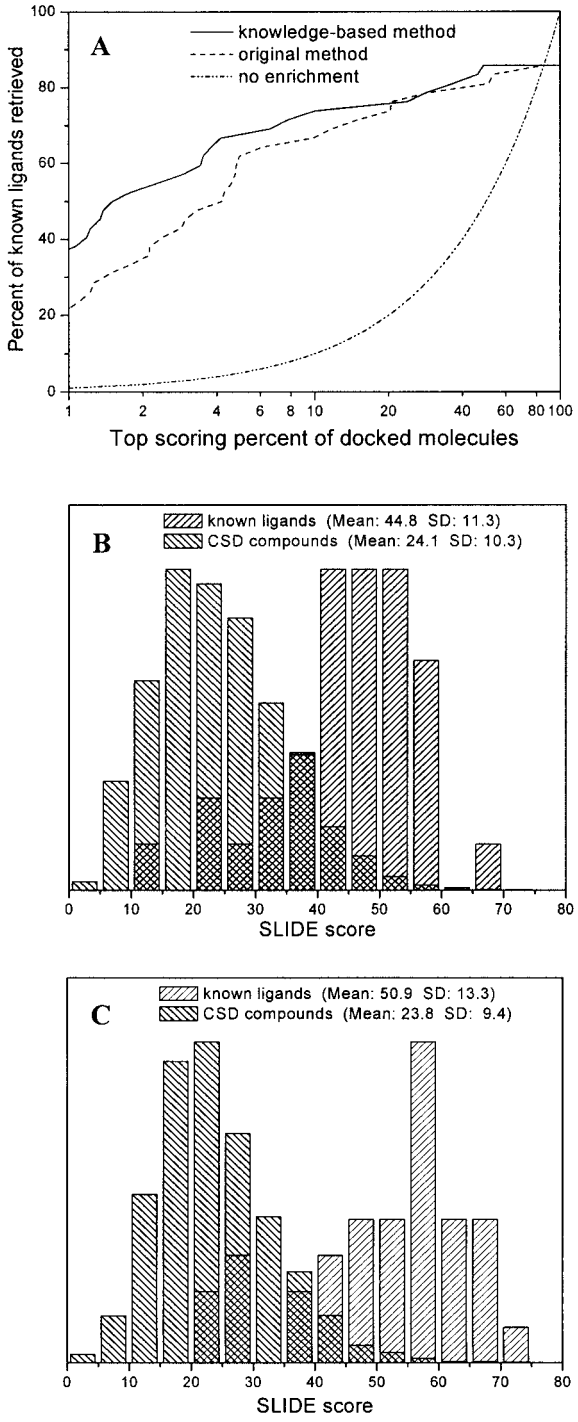


Figure 7. Screening and enrichment improvements for thrombin using the knowledge-based template and new ligand interaction point assignments, as reflected by SLIDE scores (A), where a shift to the left of the curve corresponding to the new method indicates slightly improved enrichment. The distributions of SLIDE scores obtained with the grid-based method (B) and the knowledge-based method (C) show that the knowledge-based method gives a better separation between the scores of known thrombin ligands and random CSD compounds, reflected by a greater separation between the means of their score distributions. Curves that do not reach 100% for the 'Percent of known ligands retrieved' reflect the fact that some ligands were not docked.

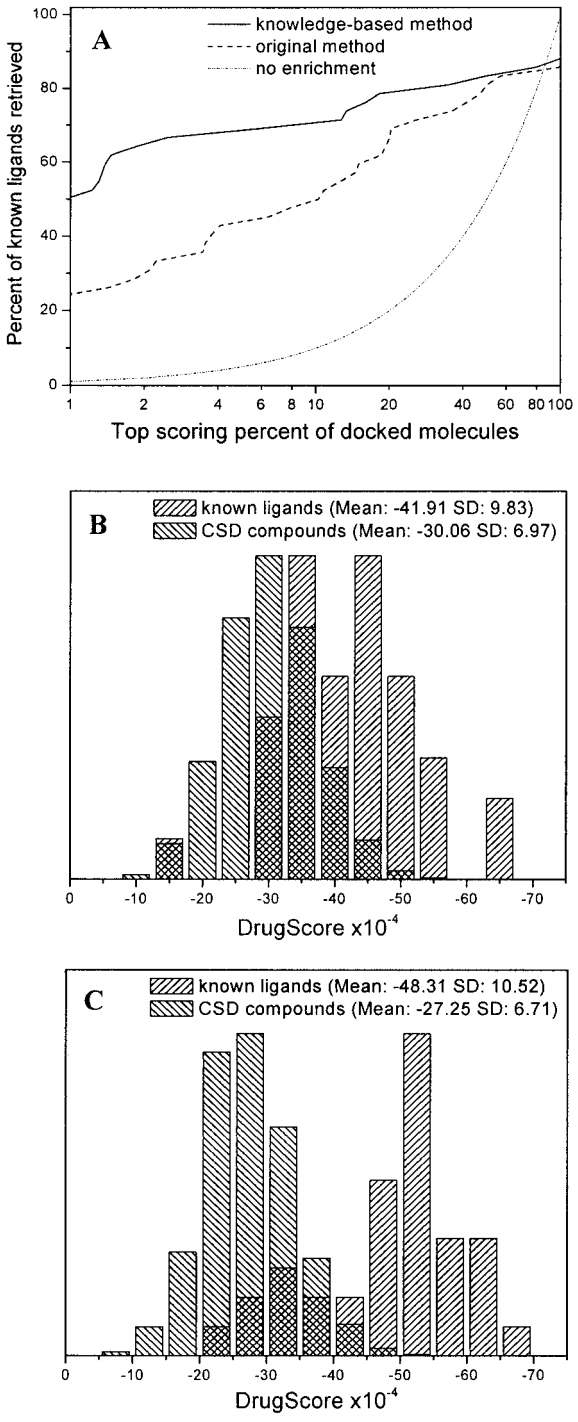


Figure 8. Significant improvement in enrichment for thrombin ligands, as reflected by the scoring function DrugScore (A), where a leftwards shift of the curve corresponding to the knowledge-based method indicates improved enrichment. The distributions of DrugScore scores (divided by 10^4) obtained using the grid-based method (B) and the knowledge-based method (C) show a much better separation between the scores of known thrombin ligands and CSD compounds. This is reflected by a 10-unit increase in separation between the mean DrugScore for ligands and the mean DrugScore for random CSD compounds. Curves that do not reach 100% for the 'Percent of known ligands retrieved' indicate that some ligands were not docked.

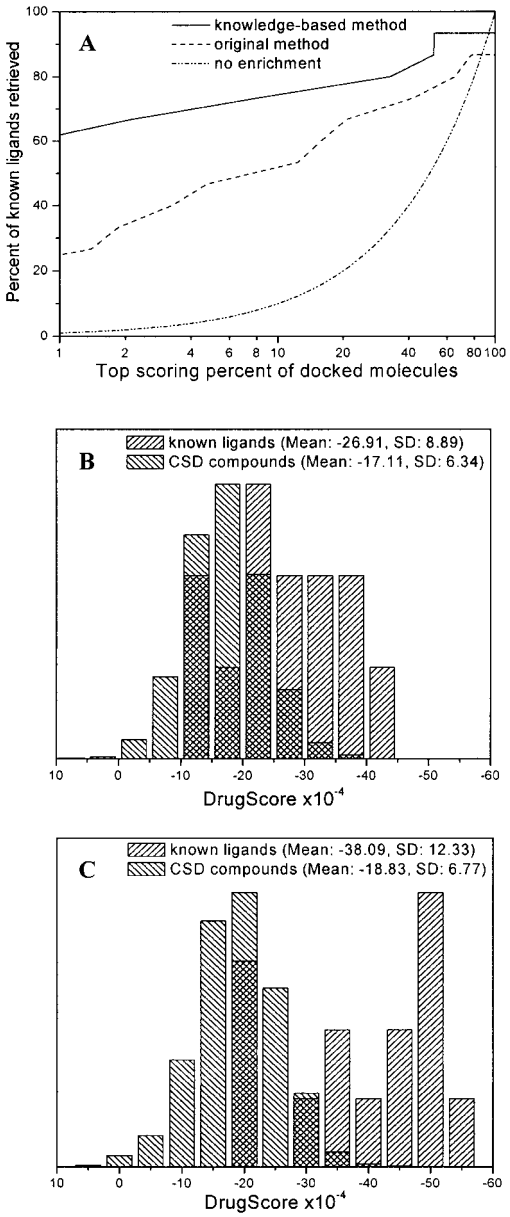


Figure 9. Enrichment for glutathione S–transferase ligands, as reflected by the scoring function DrugScore (A), where the significant leftwards shift of the curve corresponding to the knowledge-based method indicates greater enrichment. The distributions of the scores (divided by 10^4) obtained using the grid-based method (B) and the knowledge-based method (C) again show a better separation between the scores of known GST ligands and CSD compounds, indicated by the large increase of 10 units between the means of these two classes of compounds. Given the smaller sample size (15) of GST ligands, this score distribution is less well defined than those for thrombin (Figures 7 and 8). However, the same trends in improvement are found for both proteins and both scoring functions. Curves that do not reach 100% for the 'Percent of known ligands retrieved' indicate that some ligands were not docked. This percentage decreased with use of the knowledge-based template.

improvement upon replacing the grid-based template with the knowledge-based one, and the separation of scores between ligands and CSD compounds also increased. The reason for this is the ability of SLIDE to dock ligands better with the knowledge-based method, with better dockings receiving higher scores, whereas the CSD compounds received roughly the same scores using both methods.

Precise computational prediction of the binding affinities of a series of ligands for an arbitrary protein target cannot be routinely achieved by any method at this time. Particular challenges remain in the handling of interfacial solvation and protein and ligand flexibility, so scoring functions perform best when the details of the protein-ligand complex are well-resolved. Thus, docking presents a particularly hard case for scoring, and consensus scoring by combining several scoring functions has been suggested to enhance hit rates [31–33]. To compensate for the shortcomings of using a single scoring function, a second, independent scoring function, DrugScore, was also used to score the ligands docked by SLIDE. For thrombin and GST, the two scoring functions showed similar results: increased screening enrichment for known ligands, due to better separation of the ligands from CSD compounds. The correlation between the SLIDE scores and the DrugScores of known ligands also increased (Figure 10). This could be due to both scoring functions being trained on correctly positioned ligands from known protein–ligand complexes. They both perform quite well when the ligand is docked correctly, but may show less consistent performance on slightly misdocked molecules. In fact, our analysis on the relationship between RMSD and score (unpublished results) indicates that as a ligand is shifted from its optimal position, the correlation between RMSD and score is quickly lost. Once the ligand is slightly misdocked (say, due to a 1.5 Å shift from its optimal position), its score may be indistinguishable from a that of a poor docking due to misalignment of key hydrogen bonds and hydrophobic interactions. Thus, the score may not suggest that the docking is close to being correct. This problem would be difficult to solve by focusing on improving the scoring function, since even a perfect scoring function would be quite sensitive to a 1.5 Å shift between the interacting protein and ligand groups. However, this problem *can* be addressed by improving the sampling of orientational space and the modeling of flexibility in docking. Better sampling and flexibility modeling result in testing more accurate dockings, increasing the probability

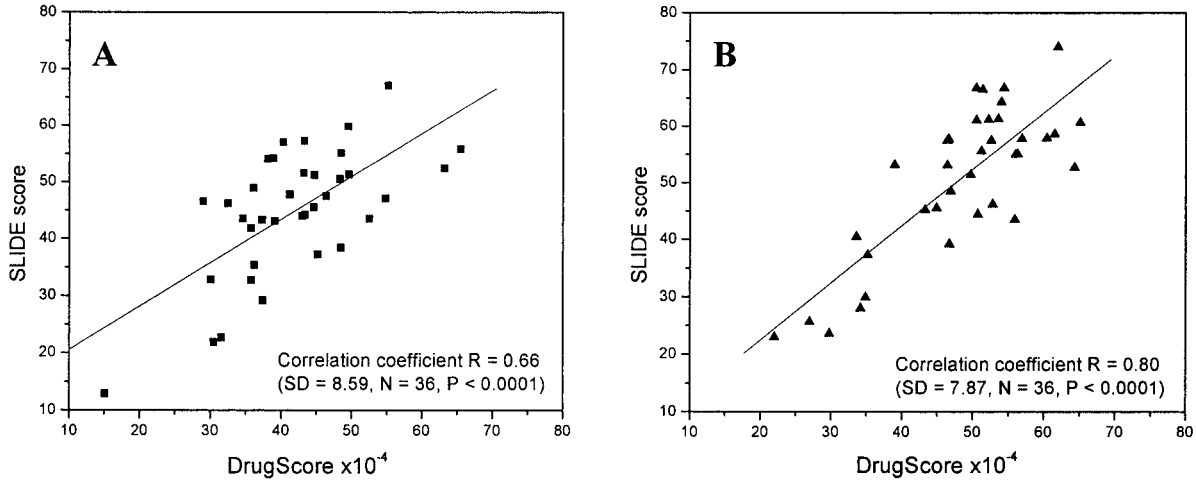


Figure 10. Correlation between SLIDE scores and DrugScores of known thrombin ligands with the grid-based (A) and the knowledge-based method (B). The negative DrugScore scores are shown with positive sign for ease of comparison, so that correlation rather than anticorrelation between DrugScores and SLIDE scores is measured.

that the correct interactions between protein and ligand will be measured and result in high scores. The SLIDE scoring function and flexibility modeling remained the same in versions 1 and 2 (the 'original' and 'new' versions mentioned in Figures 6–9). Therefore, the improvements in the sampling and representation of protein and ligand chemistry alone account for the significant improvements observed in the scores and docking RMSD values with the new version of SLIDE (see Figures 6–9).

The modeling of protein flexibility is also very important to accurate docking. Often, validation studies test redocking, in which the ligand is removed from the co-crystal structure, and the separated protein and ligand structures are used to test the docking program's ability to identify the correct ligand binding orientation in the protein. In that case, the protein is guaranteed to be in the correct conformation for the ligand. This simplifies the docking problem, such that only orientational sampling for the ligand is needed. It also assumes that the correct protein conformation is known for that ligand, which is not true when predicting a protein-ligand complex or designing a new ligand. Only 9 of the 42 thrombin ligands could be docked into the *apo* structure without conformational change in the protein or ligand (data not shown), whereas with SLIDE flexibility modeling of the protein and ligand, 36 of 42 (86%) of the ligands could be docked. For GST, 93% of ligands could be docked with flexibility modeling, but only 60% without. Thus, SLIDE models flexibility appropriately, allowing correct docking of the majority (~90%) of thrombin and GST ligands, as well as discriminating well between ligands and non-ligands in screening. Without protein flexibility modeling, for most ligands docking requires using the pre-conformed protein structure for that ligand, or forcing unnatural, additional flexibility within the ligand.

Both SLIDE score and DrugScore performed significantly better using the knowledge-based protein representation than with the original grid-based template. Regularizing the sampling of hydrophobic interaction points on ligands (another change in version 2 of SLIDE, relative to version 1) also resulted in docking and scoring improvements. One explanation for the observed improvements in scoring could be that neither scoring method was optimized to work with a grid-based template, in which the distances measured between interacting atoms could be non-optimal due to rounding off to the nearest grid point. However, this brings up the important point that the protein template and ligand interaction points in SLIDE are used only for the initial docking of the ligand, whereas scoring by either method is done using the full-atom representation of the ligand docked to the protein, after flexibility modeling (and without reference to the template or interaction points). Thus, *improving the quality of the initial docking,* through improving the representation of the protein and ligand, is what results in the significant improvements in docking accuracy and scoring observed here. These improvements are apparently independent of the scoring function used

(DrugScore and SLIDE score were developed using different paradigms, as discussed below) or on the particulars of the protein and its ligands (thrombin and GST are structurally and chemically quite different).

We have no definitive explanation for why SLIDE score and DrugScore results are apparently so correlated for the thrombin ligands (R = 0.80; Figure 10B). DrugScore is derived from the extent to which a given protein-ligand complex shows favored distances between the protein and ligand atoms. Favorability is gauged from pair-wise atomic distance distributions derived from a large set of protein-ligand complexes from the Protein Data Bank. The SLIDE scoring function is a weighted sum of two terms. The first measures hydrophobic complementarity, calculated as the complementarity in atomic hydrophobicity values of atoms in the ligand with protein atoms that are within a certain radius. This radius was chosen to include the first shell of protein atoms within van der Waals contact of the ligand atom. The atomic hydrophobicity values came from a prior study of the tendency of protein surface atoms to bind water molecules in crystallographic structures [50]. The second term in the SLIDE scoring function, counting intermolecular hydrogen bonds, is based on others' studies of the favored geometries of hydrogen bonds involving protein atoms. Despite counting interactions somewhat differently, SLIDE score and DrugScore are both based on knowledge derived from the geometry of interactions within protein crystallographic structures. This may be the fundamental basis for the observed correlation in their values for the thrombin complexes.

A number of groups have done docking and screening method validations on thrombin [12, 14, 33–40], with a focus on how the docking and scoring methods affect the results. In particular, Stahl and Rarey [33] present a detailed analysis of four different scoring functions in combination with the docking tool FlexX, using thrombin as one of their targets. Depending on the scoring function used, 20–70% of the 67 known thrombin ligands are among the top ranking 10% of their screening database of about 10000 compounds. This percentage improves to 80% when using a combined scoring function. Baxter et al. [34] test the docking accuracy of PRO_LEAD on 70 protein-ligand complexes including 6 thrombin structures, resulting in 79% of the ligands being docked within 2.0 Å RMSD. This program also provides a reasonable separation between the docked scores of the 43 known thrombin ligands and 10000 random molecules from the screening database, with 84% of the known ligands ranking among the top scoring 10% of docked molecules. Knegtel et al. [35] compare the performance of DOCK 4.0 and FlexX 1.5 by docking 32 known ligands to thrombin. For ∼40% of the ligands, fully flexible docking yields orientations within 2 Å of the known binding modes. This increased ligand conformational sampling in DOCK is found to be comparable to rigid docking of about 800 conformers per ligand and increases the docking accuracy somewhat, at the expense of an additional 20 minutes' run time per compound. In another study, Knegtel et al. [36] use DOCK 4.0 to identify thrombin inhibitors from a database of 32 known inhibitors, ten chemically similar but inactive compounds, and 1000 corporate database compounds. The performance is again scoring-function dependent, with 78–94% of actives being ranked among the 10% best scoring molecules, but neither scoring function gave a good differentiation between actives and inactives among the top scoring compounds. In the results presented here, SLIDE screening on the ∼15,000 molecules of the combined thrombin ligand and random CSD compound database identified 64–67% of thrombin ligands (depending on whether SLIDE score or DrugScore was used as the metric) within the top 0.7% of screened compounds. The runtime was about 17 hours for this screening. Although the runtime is determined primarily by the template size, other factors like ligand size and number of rotatable single bonds are also influential. While it is risky to compare methods using different ligand database sizes and degrees of molecular diversity (as described above), these results give some idea of the state of the art for molecular screening and docking of ligands for thrombin and GST.

Other groups have also investigated the influence of protein or ligand representation on docking results. Fradera et al. [37] test two ligand similarity-driven flexible docking approaches by modifying DOCK 4.0 to include the molecular-field matching program MIMIC [41]. The modified methods outperform DOCK by improving the quality of the 31 thrombin ligand dockings by 1 Å RMSD on average and by identifying 1.5–2 times more active molecules among the top-ranked 10% of molecules, for each of the three screening databases used. Their results with MIMIC/DOCK tend to be better than results of DOCK alone and take far less time, but prove to be rather dependent on the choice of the reference ligand. Fox and Haaksma [38] test their approach of combining GRID [16] to map the binding site of thrombin and UNITY (TRIPOS, Inc.) to do a flexible 3D database

search for benzamidine-based thrombin inhibitors, using a database of in-house thrombin inhibitors and a subset of ACD compounds. The method provides accurate docking orientations for 90% of the x-ray conformations of the known inhibitors, although the docking accuracy drops considerably in the case of CORINA-generated conformers [42].

Glutathione S-transferase has been less widely studied as a docking and screening target, although it has been included in some larger docking validations [12, 14, 43]. There are at least 11 different GST isozymes with different substrate specificities, which complicates the comparisons. Koehler et al. [44] use an interesting approach to decipher the key determinants of GST isozyme selectivity. Based on finding that glutathione (GSH) binds to all isozymes in a single bioactive conformation, they superimpose the available GST x-ray structures from the PDB using the bound ligands rather than the protein backbones to compare their binding sites. Their conclusion that the shape and surface hydrophobicity of the binding site are the key determinants of differences in ligand specificity between GST isozymes can be exploited in finding new, more isozyme-specific inhibitors by virtual screening. Such isozyme-specific differences would appear directly in SLIDE's knowledge-based protein templates for different GST isozymes, providing a convenient way to screen for ligands that bind well to one template/isozyme but not another.

Conclusions

Our results show that improving the representation of hydrogen-bonding and hydrophobic interaction points on the ligand and protein by a knowledge-based approach, as implemented in SLIDE, can significantly improve both the quality of docking and the docking scores of known ligands relative to randomly-selected molecules. The resulting unbiased protein template can also provide significant insights into the binding and specificity determinants of the protein, and thus provide a structure-based design template for optimizing ligand functional groups. SLIDE v. 2.0, including source code, is available to academic and industrial researchers; please see http://www.bch.msu.edu/labs/kuhn/web/projects/slide/home.html.

Acknowledgements

The authors thank Judith Guenther, Holger Gohlke, and Gerhard Klebe of University of Marburg, Germany, for providing DrugScore for our use. We would also like to thank Dylan Hirsch-Shell for his review of the manuscript. This work was supported by grants from the American Heart Association (9940091N), the National Science Foundation (DBI–9600831), and the National Institutes of Health (U01 AI-053877).

References

1. Schnecke, V. and Kuhn, L.A., Perspectives in Drug Discovery and Design, 20 (2000) 171.
2. Schnecke, V. and Kuhn, L.A., Intell. Syst. Mol. Biol., (1999) 242.
3. Schnecke, V. and Kuhn, L.A., Thorpe, M.F. and Duxbury, P.M. (Eds.) Rigidity Theory and Applications, Kluwer Academic/Plenum Publishers, New York, NY, 1999, pp. 385-400.
4. Schnecke, V., Swanson, C.A., Getzoff, E.D., Tainer, J.A. and Kuhn, L.A., Proteins, 33 (1998) 74.
5. Hobohm, U. and Sander, C., Protein Sci., 3 (1994) 522.
6. Raymer, M.L., Sanschagrin, P.C., Punch, W.F., Venkataraman, S., Goodman, E.D. and Kuhn, L.A., J. Mol. Biol., 265 (1997) 445.
7. Ippolito, J.A., Alexander, R.S. and Christianson, D.W.J., Mol. Biol., 215 (1990) 457.
8. McDonald, I. and Thornton, J.M., *Atlas of Side-Chain and Main-Chain Hydrogen Bonding*, http://www.biochem.ucl.ac.uk/∼mcdonald/atlas/.
9. Kuntz, I.D., Blaney, J.M., Oatley, S.J., Langridge, R. and Ferrin, T.E., J. Mol. Biol., 161 (1982) 269.
10. Shoichet, B.K. and Kuntz, I.D., Protein Eng., 6 (1993) 723.
11. Jones, G., Willett, P. and Glen, R.C., J. Mol. Biol., 245 (1995) 43.
12. Jones, G, Willett, P., Glen, R.C., Leach, A.R. and Taylor, R., J. Mol. Biol., 267 (1997) 727.
13. Ruppert, J., Welch, W. and Jain, A.N., Protein Sci., 6 (1997) 524.
14. Kramer, B., Rarey, M. and Lengauer, T., Proteins, 37 (1999) 228.
15. Morris, G.M., Goodsell, D.S., Halliday, R.S., Huey, R., Hart, W.E., Belew, R.K. and Olson, A.J., J. Comp. Chem., 19 (1998) 1639.
16. Boobbyer, D.N., Goodford, P.J., McWhinnie, P.M. and Wade, R.C., J. Med. Chem., 32 (1989) 1083.
17. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E., Nucleic Acids Research, 28 (2000) 235.
18. Allen, F.H. and Kennard, O., Chemical Design Automation News, 8 (1993) 1& 31.
19. Bergner, A., Günther, J., Hendlich, M., Klebe, G. and Verdonk, M., Biopolymers, 61 (2002), 99.
20. Gohlke, H., Hendlich, M. and Klebe, G., J. Mol. Biol., 295 (2000) 337.
21. Gohlke, H., Hendlich, M. and Klebe, G., Perspectives in Drug Discovery and Design, 20 (2000) 115.
22. Verdonk, M.L., Cole, J.C., Watson, P., Gillet, V. and Willett, P., J. Mol. Biol., 307 (2001) 841.

23. Boer, D.R., Kroon, J., Cole, J.C., Smith, B. and Verdonk, M.L., J. Mol. Biol., 312 (2001) 275.
24. Moreno, E. and Leon, K., Proteins, 47 (2002) 1.
25. Sanschagrin, P. and Kuhn, L.A., Protein Sci., 7 (1998) 2054.
26. Dekker, R.J., Eichinger, A., Stoop, A.A., Bode, W., Pannekoek, H. and Horrevoets, A.J.G., J.Mol.Biol., 293 (1999) 613.
27. Oakley, A.J., Lo Bello, M., Ricci, G., Federici, G. and Parker, M.W., Biochemistry, 37 (1998) 9912.
28. Oakley, A.J., Rossjohn, J., Lo Bello, M., Caccuri, A.M., Federici, G. and Parker, M.W., Biochemistry, 36 (1997) 576.
29. Oakley, A.J., Lo Bello, M., Nuccetelli, M., Mazzetti, A.P. and Parker, M.W., J. Mol. Biol., 291 (1999) 913.
30. Prade, L., Huber, R., Manoharan, T.H., Fahl, W.E. and Reuter, W., Structure, 5 (1997) 1287.
31. Bissantz, C., Folkers, G. and Rognan, D., J. Med. Chem., 43 (2000) 4759.
32. Charifson, P.S., Corkery, J.J., Murcko, M.A. and Walters, W.P., J. Med. Chem., 42 (1999) 5100.
33. Stahl, M. and Rarey, M., J. Med. Chem., 44 (2001) 1035.
34. Baxter, C.A., Murray, C.W., Waszkowycz, B., Li, J., Sykes, R.A., Bone, R.G., Perkins, T.D. and Wylie, W. J., Chem. Inf. Comput. Sci., 40 (2000) 254.
35. Knegtel, R.M., Bayada, D.M., Engh, R.A., von der Saal, W., van Geerestein, V.J. and Grootenhuis, P.D., J. Comput. Aided. Mol. Des., 13 (1999) 167.
36. Knegtel, R.M. and Wagener, M., Proteins, 37 (1999) 334.
37. Fradera, X., Knegtel, R.M. and Mestres, J., Proteins, 40 (2000) 623.
38. Fox, T. and Haaksma, E.E., J. Comput. Aided. Mol. Des., 14 (2000) 411.
39. Sotriffer, C.A., Gohlke, H. and Klebe, G., J. Med. Chem., 45 (2002) 1967.
40. Murray, C.W., Baxter, C.A. and Frenkel, A.D., J. Comput. Aided. Mol. Des., 13 (1999) 547.
41. Mestres, J., Rohrer, D.C. and Maggiora, G.M., J. Comp. Chem., 18 (1997) 934.
42. Sadowski, J. and Gasteiger, J., Chem. Rev., 93 (1993) 2567.
43. Chen, Y.Z. and Ung, C.Y., J. Mol. Graph. Model., 20 (2001) 199.
44. Koehler, R.T., Villar, H.O., Bauer, K.E. and Higgins, D.L., Proteins, 28 (1997) 202.
45. Connolly, M.L., J. Mol. Graphics, 11 (1993) 139.
46. Weir, M.P., Bethell, S.S., Cleasby, A., Campbell, C.J., Dennis, R.J., Dix, C.J., Finch, H., Jhoti, H., Mooney, C.J., Patel, S., Tang, C.M., Ward, M., Wonacott, A.J. and Wharton, C.W., Biochemistry, 37 (1998) 6645.
47. Böhm, H.-J., *J. Comput. Aided Mol Design*, 6 (1992) 61.
48. Fischer, D., Norel, R., Wolfson, H. and Nussinov, R., Proteins, 16 (1993) 278.
49. Fischer, D., Lin, S.L., Wolfson, H.L. and Nussinov, R., J. Mol. Biol. 248 (1995), 459.
50. Kuhn, L.A., Swanson, C.A., Pique, M.E., Tainer, J.A., Getzoff, E.D., Proteins, 23 (1995), 536.