Research Article

Scoring Ligand Similarity in Structure-Based Virtual Screening

Maria I. Zavodszky[1,2,§], Anjali Rohatgi[1], Jeffrey R. Van Voorst[1-3], Honggao Yan[1,2], and Leslie A. Kuhn[1-3*]

[1]*Department of Biochemistry and Molecular Biology,* [2]*Quantitative Biology Initiative, and* [3]*Department of Computer Science & Engineering, Michigan State University, East Lansing, Michigan*

[§]*Current address: GE Global Research, Biosciences, Niskayuna, New York*

***Corresponding author:** Protein Structural Analysis and Design Lab, Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI 48824-1319. Tel: (517) 353-8745. Fax: (517) 353-9334. E-mail: KuhnL@msu.edu
Web: http://www.bch.msu.edu/~kuhn

Keywords: inhibitor discovery, docking, thrombin, SLIDE, interaction fingerprints, high-throughput screening, affinity scoring, EON, DrugScore, GOLD score

Abbreviations: NCI – National Cancer Institute, DTP – Developmental Therapeutics Program, PDB – Protein Data Bank, ACD – Available Chemicals Directory, CSD – Cambridge Structural Database, GST – Glutathione S-transferase, ET – electrostatic, PB – Poisson-Boltzmann, 2D – 2-dimensional

ABSTRACT

Scoring to identify high-affinity compounds remains a challenge in virtual screening. On one hand, protein-ligand scoring focuses on weighting favorable and unfavorable interactions between the two molecules. Ligand-based scoring, on the other hand, focuses on how well the shape and chemistry of each ligand candidate overlay on a 3-dimensional reference ligand. Our hypothesis is that a hybrid approach, using ligand-based scoring to rank dockings selected by protein-ligand scoring, can ensure that high-ranking molecules mimic the shape and chemistry of a known ligand while also complementing the binding site. Results from applying this approach to screen nearly 70,000 National Cancer Institute compounds for thrombin inhibitors tend to support the hypothesis. EON ligand-based ranking of docked molecules yielded the majority (4/5) of newly discovered, low to mid-micromolar inhibitors from a panel of 27 assayed compounds, whereas ranking docked compounds by protein-ligand scoring alone resulted in one new inhibitor. Since the results depend on the choice of scoring function, an analysis of properties was performed on the top-scoring docked compounds according to five different protein-ligand scoring functions, plus EON scoring using three different reference compounds. The results indicate that the choice of scoring function, even among scoring functions measuring the same types of interactions, can have an unexpectedly large effect on which compounds are chosen from screening. Furthermore, there was almost no overlap between the top-scoring compounds from protein-ligand versus ligand-based scoring, indicating the two approaches provide complementary information. Matchprint analysis, a new addition to the SLIDE screening toolset, facilitated comparison of docked molecules' interactions with those of known inhibitors. The majority of interactions conserved among top-scoring compounds for a given scoring function, and from the different scoring functions, proved to be conserved interactions in known inhibitors. This was particularly true in the S1 pocket, which was occupied by all the docked compounds.

INTRODUCTION

Virtual Screening

Increases in efficiency and reliability of computational tools have enabled virtual screening to become a routine method in pharmaceutical drug discovery, complementing *in vitro* high-throughput screening (Schneider and Bohm 2002; Shoichet et al. 2002; Waszkowycz 2002; Jorgensen 2004). Inhibitors have been identified by computational screening of compound databases for protein tyrosine phosphatase-1B (Doman et al. 2002), nuclear hormone receptors (Schapira et al. 2000; 2001), human carbonic anhydrase (Gruneberg et al. 2002), thymidylate synthase (Shoichet et al. 1993), integrin alphavbeta3 (Zhou et al. 2006), acetylcholinesterase (Mizutani and Itai 2004), glycogen synthase kinase-3beta (Polgar et al. 2005) and thrombin (Massova et al. 1998; Bohm et al. 1999; Fox and Haaksma 2000; Howard et al. 2006). Typically, structure-based virtual screening (in which ligand candidates are docked into the atomic structure of a protein's binding site and evaluated by the resulting protein interactions) and ligand-based screening (in which ligand candidates are aligned and scored for similarity to a prototypical known ligand) are used independently (Oprea and Matter 2004; McGaughey et al. 2007). However, at least one tool, FRED (OpenEye Scientific Software, Santa Fe, NM; http://www.eyesopen.com), enables ligand-based scoring of docked molecules. There is growing awareness that additional information can be gained by combining structure-activity relationships derived from assayed ligands with knowledge of the protein-ligand contacts (Peltason and Bajorath 2007). Here we compare the results of ligand-based and protein-based scoring of small molecules docked by SLIDE into the active site of thrombin, after assaying the top-scoring molecules for inhibitory activity.

Thrombin as a target for docking validation, inhibitor discovery, and ligand optimization

Thrombin is a key player in the blood coagulation cascade: it catalyzes the proteolytic cleavage of the soluble plasma protein fibrinogen to produce fibrin (Stubbs and Bode 1993a). In addition, it activates Factor XIII, which cross-links the linear fibrin monomers, producing insoluble blood clots (Sadasivan and Yee 2000). Thrombin is also a potent platelet activator (Brass 2003). The coagulant activity of thrombin is controlled by thrombomodulin, a thrombin binding protein on the surface of endothelial cells (Rezaie and Yang 2003). Numerous efforts to control the blood clotting process have been directed toward thrombin because of its pivotal role in maintaining the intricate balance between hemostasis and thrombolysis (Davie et al. 1991; Das and Kimball 1995; Tulinsky 1996; Sanderson 1999; Katz et al. 2003; Mackman 2008). Biochemical modeling has been greatly aided by the wealth of structural data on thrombin produced in the past twenty five years (Stubbs and Bode 1993b; 1995; Tulinsky 1996).

Fox and Haaksma (Fox and Haaksma 2000) created 3D pharmacophore models of the thrombin active site which were used for flexible ligand screening of several proprietary databases as well as the Available Chemicals Directory (ACD; Elsevier MDL, San Ramon, CA). Their queries of varying complexities returned between two and 1057 hits enriched in known thrombin inhibitors. Bohm and co-workers (Bohm et al. 1999) screened fragment libraries taken from the ACD against multiple pockets of the thrombin active site and then linked the identified hits as building blocks in a combinatorial fashion. They identified the known thrombin inhibitor *p*-

aminobenzamidine as the top scoring compound docked into the P1 specificity pocket and connected it to 10 various substituted benzaldehydes, selected based on their scores and commercial availability. All 10 compounds designed this way bound to thrombin at least 10 times stronger than *p*-aminobenzamidine (half with binding constants in the nanomolar range) and were new, low molecular weight, non-covalent inhibitors of thrombin. A similar approach was used to identify and link small fragments of micromolar potency to create nanomolar inhibitors for thrombin (Howard et al. 2006). DOCK has also been used to screen the CSD for new inhibitors for thrombin (Massova et al. 1998). Three of their top-scoring compounds were found to be competitive inhibitors of thrombin with K_i values in the range of 170-1700 μM. These compounds are relatively hydrophobic, and the crystal structure of the complex formed by one of the inhibitors with thrombin indicates it does not bind in the dominant specificity pocket, S1.

Hybrid protein-ligand and ligand-based scoring in screening for thrombin inhibitors

In this work, SLIDE (*S*creening *L*igands by *I*nduced-fit *D*ocking, *E*fficiently) is applied in ligand discovery mode, with a focus on comparing the properties of molecules identified by ligand-based versus protein structure-based scoring. SLIDE is a computational tool which can efficiently screen hundreds of thousands of molecules to identify feasible ligand candidates with good steric and chemical complementarity to a chosen site on a protein target (Schnecke and Kuhn 1999; 2000). The realistic modeling of protein side-chain and ligand flexibility and the improved representation of the binding site by knowledge-based template design (Zavodszky et al. 2002; Tonero et al. 2008; in preparation) allow discrimination between true ligands and non-specific compounds and facilitate the discovery of new inhibitors (Sukuru et al. 2006). Presented here are novel thrombin inhibitors discovered by screening the Plated Compounds Database of the National Cancer Institute (NCI). The properties of molecules selected based on SLIDE scoring and EON ligand-based scoring are compared and contrasted. This is followed by a detailed analysis of the chemical properties and interactions selected when a series of different protein-ligand scoring functions are used, and when the reference ligand for ligand-based scoring is varied.

MATERIALS AND METHODS

Structures used for screening

The crystal structure of human alpha thrombin with ligand-free catalytic site, representing the unbiased conformation, was used as the target for screening and docking (PDB entry 1vr1 (Dekker et al. 1999); http://www.rcsb.org/pdb (Berman et al. 2000)). Three-dimensional structures for compounds contained in the NCI/DTP Open Chemical Repository (http://dtp.nci.nih.gov) were downloaded from the ZINC (Irwin and Shoichet 2005) web site (http://blaster.docking.org/zinc6/) and used as the screening database. Filtering was applied to include only drug-like molecules having the following properties: (1) molecular weight of 100-500 Da, (2) maximum of 5 hydrogen bond donor atoms, (3) maximum of 10 hydrogen bond acceptor atoms, and (4) logP value between 0 and 5. To ensure thorough conformational sampling of the ligand candidates, low-energy conformations were generated for each molecule using Omega v1.8β (OpenEye Scientific Software, Santa Fe, NM). The default parameters were

used except for the following: fixcycle: true; includeinput: true; multioutputfiles: false; warts: true; MMFF94s: true; rms: 0.6; finalcut: 5.0; and finalopt: false. As a result, conformers were generated within a range of 5.0 kcal/mol from the lowest energy conformer sampled, as determined by the Merck Molecular Force Field (MMFF94s) (Halgren 1999).

Screening software

The workflow of screening and scoring steps is summarized in Figure 1. SLIDE (Schnecke and Kuhn 2000) uses distance geometry and geometrical indexing to dock ligands into the binding site of the target protein by testing all possible ligand anchor fragments for their ability to match part of the protein template. The template consists of points identified as the most favorable positions for ligand atoms to form hydrogen bonds or make hydrophobic interactions with protein atoms (Zavodszky et al. 2002). The ligand anchor fragments are defined by triplets of interaction points, which are hydrogen-bonding ligand atoms or centers of hydrophobic atom clusters. Every combination of three interaction points in the ligand is compared to every combination of three template points in a search for complementary shape and chemistry, while ligand groups outside the anchor fragment are modeled flexibly. SLIDE models induced complementarity by rotating the protein side chains and ligand groups upon binding, using mean-field optimization to determine the minimal set of rotations needed to allow steric fit (Schnecke and Kuhn 1999). This approach has been shown to closely model ~85% of protein side-chain motions upon ligand binding (Zavodszky and Kuhn 2005).

The knowledge-based binding site representation and side-chain flexibility modeling has enabled SLIDE to reproduce the correct binding mode for 80% of known thrombin and glutathione S-transferase (GST) ligands tested previously (Zavodszky et al. 2002). Screening a database of 15,000 randomly chosen Cambridge Structural Database (Allen 2002), molecules with drug-like size and atomic composition mixed with 42 known thrombin inhibitors resulted in retrieving 64% of the known thrombin inhibitors among the top scoring 100 molecules. A similar experiment for GST ranked 11 out of 15 known inhibitors among the top 100 docked molecules (Zavodszky et al. 2002). The accuracy of the current version (3.0) of SLIDE has also been tested by re-docking a series of 100 diverse ligands into their target structures, with 61% of the top-scoring ligand dockings falling within 2.0 Å RMSD of the crystallographic orientation (Tonero et al. in preparation). This performance (using default settings) is similar to that of the other top-performing methods on the same dataset: GOLD and Surflex (56-57%) (Jain 2003; Verdonk et al. 2003; Kellenberger et al. 2004). A newer version of GOLD (v. 3.2, Cambridge Crystallographic Data Centre, Cambridge, UK) docks 59% of these complexes to within 2 Å RMSD (when all residues within 10 Å of the ligand are used to define the binding site, consistent with how SLIDE was run). Using the NCI/DTP Open Chemical Repository database, SLIDE has identified seven new classes of low to mid-micromolar inhibitors for the drug discovery target asparaginyl-tRNA sythetase from the human parasite *Brugia malayi* (Sukuru et al. 2006), with a hit rate of 15% (7 out of 45 compounds assayed were confirmed as inhibitors).

Scoring functions for ranking orientations and relative affinities of ligand candidates

SLIDE, used here to screen for new inhibitors of thrombin, employs a two-step scoring protocol. The first step identifies the best orientation among all the docked conformations of a ligand candidate. In this study, DrugScore was used to identify the top-scoring orientation from SLIDE. DrugScore (Gohlke et al. 2000) was trained on inter-atomic distance profiles derived from protein-ligand complexes in the PDB (Berman et al. 2000). Given the best orientation of each ligand candidate selected by DrugScore, a second scoring function ranked the ligand candidates according to their estimated binding affinity. Either size-normalized AffiScore or EON (described below) was used as this second scoring function to rank ligand candidates for experimental testing. For analyzing the variations in chemical properties of the top hits retrieved by different scoring methods, AffiScore without size normalization, DrugScore, and GOLD were also employed.

AffiScore in SLIDE v. 3.0 is a weighted sum of favorable interactions (hydrophobic contacts, H-bond, salt-bridge, and metal interactions) and unsatisfied and repulsive polar interactions between the protein and docked ligand. The size-normalized option, AffiScore/N, divides the score by the number of heavy atoms in the ligand candidate, favoring smaller compounds with similarly good interactions. Both versions of AffiScore can be used during docking or for rescoring of existing dockings. Our goal was to test the performance of AffiScore/N for the ability to discover inhibitors, which, due to their lower molecular weight relative to compounds favored by most other scoring functions, would be easier to synthesize and optimize. In separate tests, we found AffiScore (without normalization) useful for ranking compounds, attaining a linear correlation coefficient of r=0.63 with the experimentally measured $\Delta G_{binding}$ values for 273 complexes. This is similar to the 0.68 value for XScore (Wang et al. 2003), the best-performing scoring function on this dataset (Tonero et al. in preparation). However, AffiScore and AffiScore/N perform scoring 35 times faster than XScore and are thus more amenable for use during high-throughput screening on large databases. GOLD score (v. 4.0.1) was also used to rank all the docked compounds. GOLD score with the ligand internal energy, S(int), subtracted, was also used, because it was found by the developers to correlate better with ligand-binding affinity values than the full GOLD score (Verdonk et al. 2003).

For ligand-based scoring, EON (OpenEye Scientific Software, Santa Fe, NM) (Nicholls et al. 2004) was used to calculate the electrostatic similarity between each SLIDE-docked ligand candidate and the known thrombin inhibitor PPACK (PDB entry 1ppb (Bode et al. 1989)). The reported ET_PB score is the electrostatic Tanimoto coefficient calculated using Poisson-Boltzmann electrostatics. For EON scoring, the PPACK ligand from PDB entry 1ppb was aligned into the ligand-free active site using InsightII (Accelrys, San Diego), by least-squares superposition of the Cα atoms of all binding site residues within 9 Å of any PPACK atom in 1ppb onto the corresponding Cα atoms in the 1vr1 structure used for docking. The Cα RMSD was 1.3 Å. The ET_PB score was calculated for the best orientation of each docked ligand candidate (identified by DrugScore) relative to this orientation of PPACK.

Computing chemical properties and diversity of top-scoring candidates

To assess the extent to which highly ranked ligands from protein-ligand scoring functions differed from those in ligand-based scoring, we compared the properties of the top-100 compounds selected by five different protein-ligand scoring functions with EON ligand-based scoring using different ligands as reference compounds. The protein-ligand scoring functions were: GOLD score, GOLD score – S(int), DrugScore, AffiScore/N, and AffiScore not normalized for ligand size. To observe its sensitivity to changing the reference ligand, EON electrostatics scoring was performed using three different thrombin inhibitors as references (Figure 2), in their orientations from the corresponding crystal complexes: Phe-Pro-Arg-chloromethylketone (PPACK; from PDB entry 1ppb; Ki = 0.23 nM (Kettner and Shaw 1979), BIC-Arg-N-phenethyl-formamide (PDB entry 1a5g; Ki = 0.071 nM; (Wu et al. 1993; St. Charles et al. 1999); and D-Phe-Pro-Arg-benzothiazole (PDB entry 1tbz; Ki = 0.16 nM; (Matthews et al. 1996). These three high-affinity inhibitors differ significantly in their chemical groups binding in the S1' pocket, and one has a bicyclic group instead of proline in the S2 pocket.

The molecular weights and 2D polar surface areas of the top-100 compounds with respect to each scoring function were computed with FILTER (OpenEye Scientific Software, Santa Fe, NM) (Wang et al. 1997). Diversity within the top-ranking compounds was evaluated by computing the pair-wise ligand similarity scores within each set by using ROCS (OpenEye Scientific Software, Santa Fe, NM). This chemical similarity measure uses the implicit Mills-Dean color force field, which determines how chemical centers of six types of functional groups can interact (hydrogen-bond donors, hydrogen-bond acceptors, hydrophobic centers, anions, cations, and rings). The model of hydrogen bonding in this force-field is based on a survey of small molecule hydrogen-bonding geometry (Mills and Dean 1996).

Interaction fingerprinting

A new SLIDE utility for calculating interaction fingerprints, or "matchprints", between proteins and docked (or crystallographically oriented) ligands was encoded, incorporating SLIDE's tabulation of protein-ligand interactions. This is similar to the SIFt method of analyzing interactions (Deng et al. 2004). In the matchprint method, a hydrophobic contact is formed by a hydrophobic center in the protein – a carbon or sulfur atom not covalently bonded to a polar (oxygen or nitrogen) atom – that is within 4.5 Å contact distance of a carbon or sulfur in the ligand. Hydrogen bonds are identified as interactions between acceptor and donor atoms at a distance of 2.5-3.5 Å, with a donor-H-acceptor angle of 120°, and a pre-acceptor-acceptor-H angle of 90-180°. Salt bridges, which have a substantial Coulombic component and less angular dependence than hydrogen bonds, are identified for oppositely charged atoms, each with a charge magnitude of ≥ 0.3 (as defined by the partial charges assigned in ZINC), and falling within a distance of 2.5-4.5 Å. For comparing a set of protein-ligand interactions in a matchprint, only those interactions that appeared in at least one of the complexes were tabulated. Forty-two thrombin ligands from the PDB and 27 tested inhibitors were compared with respect to shared interactions with thrombin, along with the top-100 docked ligand candidates from each scoring function. This matchprint utility has been prepared for release in SLIDE v. 3.2, which also screens twice as fast as SLIDE v. 3.0. For information on SLIDE licensing (which is free to academic groups), see Software under http://www.bmb.msu.edu/~kuhn.

Experimental reagents

For enzymatic activity assays, human alpha-thrombin was purchased from Enzyme Research Laboratories, South Bend, IN. The chromogenic substrate Sar-Pro-Arg-*p*-nitroanilide and the known thrombin inhibitor *p*-aminobenzamidine were purchased from Sigma Aldrich. The Drug Synthesis and Chemistry Branch, Developmental Therapeutics Program, Division of Cancer Treatment and Diagnosis, National Cancer Institute provided samples of compounds from the Open Chemical Repository.

Measuring inhibition of top-scoring docked candidates

To select a moderate number of compounds to assay for inhibition, the 100 top-scoring ligand candidates from EON and also the top 100 from size-normalized AffiScore were visualized using VIDA (OpenEye Scientific Software, Santa Fe, NM). Final candidates were selected based on placing an amine or amide in the S1 specificity pocket (a known feature of all thrombin inhibitors) and not containing predominantly hydrophobic (carbon) atoms, since such molecules are typically insoluble. (Using a logP filter for pre-processing the screening database is not consistently effective at identifying soluble compounds, in our experience.)

Peptide cleavage activity of human thrombin was measured in a chromogenic assay using Sar-Pro-Arg-*p*-nitroanilide as the substrate. Stock solutions of thrombin (500 nM), the chromogenic substrate (6 mM) and the known inhibitor (positive control) *p*-aminobenzamidine (100 mM) were prepared in 50 mM TRIS/HCl pH 7.8, 100 mM NaCl, 0.1% PEG 8000 buffer (Hopfner et al. 1993). The final concentration of thrombin in the reaction mixture was 5 nM, and the substrate concentration was 60 μM. 10 μl of thrombin stock solution and 10 μl of inhibitor stock solution were added to 970 μl of TRIS buffer and incubated for 5 minutes at room temperature. Then the reaction was initiated by the addition of 10 μl of the chromogenic subtrate stock solution. Initially, 100 mM stock solutions of the inhibitor candidates were prepared in DMSO. Subsequently, they were further diluted with DMSO such that after the addition of 10 μl to the reaction mixture, the final inhibitor concentrations ranged from 0.005 μM to 1.0 mM. As a negative control, buffer was used to substitute for the highly soluble known inhibitor *p*-aminobenzamidine. For the NCI compounds, DMSO was used as a negative control. The initial rate of substrate cleavage by thrombin was determined by monitoring the absorbance change at 404 nm as a function of time, using a Hewlett Packard 8452A spectrophotometer. The percent enzymatic activities at various inhibitor concentrations were obtained by comparing the initial reaction velocities (rates) for thrombin cleavage of the chromogenic substrate in the presence of inhibitors to those measured for the negative control (used to define 100% activity). The percent activity values were plotted against the concentrations of the inhibitor on a logarithmic scale. The data was fit with a sigmoidal curve using the program Origin 7.5, and the inhibitor concentration corresponding to 50% thrombin activity (IC_{50}) was determined.

A potential problem with small molecule inhibition assays was noted by Shoichet and colleagues. They discovered that some small molecules appearing to inhibit several different proteins proved not to be site-specific. They actually formed sub-micrometer aggregates that non-specifically inhibited or engulfed the enzymes (McGovern et al. 2002; McGovern et al. 2003). To assess the possibility of non-specific aggregating inhibition according to their

recommended protocol, the enzyme assay for each apparent inhibitor was repeated in the presence of 0.02% Triton detergent. This disrupts the formation of any ligand aggregates. Thus, any promiscuous or aggregating inhibitors can be identified by the loss of inhibition in the presence of this low concentration of detergent.

RESULTS AND DISCUSSION

The NCI dataset downloaded from the ZINC database contained 66,777 molecules after drug-like filtering, from which 648,641 low-energy conformers were generated with Omega. The minimum number of conformers generated for any compound was 1, the maximum was 400, and the mean was 11 (std. dev. 21). From the SLIDE dockings of all conformers for each compound, DrugScore selected the best-scoring orientation, yielding 7,618 docked compounds. The total screening time for the 648,641 conformers was 64.6 hours.

Inhibition assays

Visual inspection of the top-100 compounds from AffiScore/N and the top-100 compounds from EON using PPACK from 1ppb as a reference was performed to filter out compounds that were primarily hydrophobic or did not place an amine group in the S1 pocket. Based on these criteria, 29 compounds were selected for experimental testing. Of these 29 compounds, only one was present in both the AffiScore and EON top-100 lists. This compound and one other were no longer available. One of the 27 compounds received from NCI was found to be insoluble when the stock solution (in DMSO) was diluted in TRIS buffer. Of the 26 soluble compounds, five proved to inhibit human thrombin with IC_{50} values ranging from 635 nM to 716 μM (Table 1 and Figure 3). Repeating the inhibition assays in the presence of Triton detergent showed no evidence of inhibition by aggregation (engulfing the protein) for any of the five compounds (Figure 4). The IC_{50} value measured for *p*-aminobenzamidine, a known thrombin inhibitor acting as positive control, was 96 μM, and the K_M value was 83 μM. Using these values and the substrate concentration of 60 μM, the calculated K_i for *p*-aminobenzamidine was 56 μM, similar to previously published values (65-80 μM) (Markwardt et al. 1968; Evans et al. 1982).

Comparing the newly discovered inhibitors with known thrombin inhibitors

The SLIDE-predicted binding modes of the five newly identified thrombin inhibitors (Table 1) overlap with binding modes observed for other ligands in the crystal structures of thrombin complexes (Figure 5). While they share chemical substructures with existing thrombin inhibitors, none was a previously known inhibitor, and they exhibited some novel interactions. The second-highest affinity inhibitor was identified by normalized AffiScore, whereas the other four inhibitors were identified by EON/1ppb. Compound 1 was the strongest inhibitor, at 635 nM. Its bicyclic coumarin moiety was positioned into the S1' pocket (Figure 5A), where it made more extensive interactions than other known thrombin inhibitors. Adding functional groups to this scaffold to fill the S2 and S3 pockets (where the phenyl group and modified proline side chain bind in a typical PPACK-like inhibitor, shown in green tubes on the left side of Figure 5A) could potentially result in very strong inhibitors. Compound 4 (Figure 5D) is similar to the 6-fluorotryptamine described earlier as a thrombin inhibitor (Nienaber et al. 2000), except it has a

more polar tail and lacks fluorine. The binding mode of compound 4 differed in that the aromatic ring of 6-fluorotryptamine (Nienaber et al. 2000) binds in the S1 specificity pocket, whereas the same ring system in compound 4 interacted closer to the S2 pocket because there was insufficient room in S1 to accommodate both the guanidinium group and the indole. Three of the five new inhibitors (compounds 3-5, Figure 5C-E) placed an aromatic group at the branch point between the S1 and S2 pockets, which is novel relative to most thrombin inhibitors. These three compounds also contain an aminoguanidine group binding deep in the S1 pocket. Such compounds are known to be effective inhibitors for West Nile and Dengue virus serine proteases (Ganesh et al. 2005).

The five confirmed ZINC inhibitors (Table 1) were also generally much smaller than the PDB thrombin inhibitors. By comparing their interaction matchprints with the 42 inhibitors from the PDB (Supplementary Table 1), the five newly discovered inhibitors made many of the same interactions. At least 4 of the 5 molecules interacted with the Trp 60D side chain, Asp 189 OD1 and OD2, Ala 190 O and side chain, and Gly 219 O. However 4 out of 5 also interacted with the Glu 192 side chain, which is rarely observed for known inhibitors, and these molecules missed many additional opportunities for interaction (hydrophobic interactions with Tyr 60A, Leu 99, Ile 174, Trp 215, and Glu 217 side chains, and hydrogen bonds with Ser 214, Gly 216, and Gly 219, all lining the S2 and S3 pockets). Of the ZINC compounds that were assayed (Supplementary Table 10) but did not inhibit, more than three-quarters missed at least two of the common interactions for the new and previously known inhibitors (Trp 60D side chain, Asp 189 OD1 and OD2, Ala 190 O and side chain, and Gly 219 O). Thus, one possible approach for decreasing false positives would be to filter dockings to identify compounds that can make all, or all but one, of these interactions.

Properties of compounds highly ranked by different scoring functions

To assess the influence of the scoring function choice on the properties of compounds (in the absence of filtering on additional criteria by molecular graphics), we compared the top-scoring 100 compounds from each of the 8 scoring functions with respect to: (1) the molecular weight distribution of the compounds, (2) their 2D polar surface areas, (3) the percentage of compounds in common between each pair of scoring functions, (4) the chemical diversity within each top-100 set of molecules, and (5) patterns in their interactions as docked with thrombin, relative to interactions observed in thrombin-inhibitor crystal structures. As shown in Figure 6A, AffiScore/N tended to select low molecular weight compounds (mostly under 200 D), whereas EON using three different reference compounds resulted in compounds with intermediate molecular weights (150-300 D), and AffiScore, DrugScore, GOLD, and GOLD-S(int) selected relatively high molecular weight compounds (250-450 D). Many known thrombin inhibitors are similar with respect to the groups they place in the S1 and S3 pockets (guanidinium in S1, and phenyl in S3), so the EON reference inhibitors were chosen as having similar affinity values but different chemistry of binding in the two other pockets, S1' and S2 (Figure 2). For polar surface area distributions of top-scoring compounds (Figure 6B), the same separation according to scoring function type was observed: AffiScore/N selected compounds with low polar surface area (mostly 20-60 $Å^2$), EON selected compounds of intermediate polar surface area (mostly 20-100 $Å^2$), and the four other protein-ligand scoring functions, AffiScore, DrugScore, GOLD, and GOLD-S(int), selected compounds with a greater range of polar surface areas (mostly 50-140 $Å^2$). The normalization of AffiScore/N by the number of heavy atoms in the ligand resulted in a

strong bias towards smaller compounds with less polar surface area, as well as compounds with a higher efficiency, per atom, in making favorable interactions.

One surprise was that the five protein-ligand scoring functions tended to select different compounds as top-scoring, though they nominally measure similar features of hydrophobic, hydrogen-bond, and mismatched polar/non-polar interactions. AffiScore/N showed minor overlap (3 compounds) with the top-100 compounds from each of the three EON lists and had no compounds in common with the four other protein-ligand scoring functions (Table 2). These four showed moderate overlap in their top-100 lists, from a minimum of 7 compounds in common for GOLD and AffiScore, to 24 compounds in common for DrugScore and GOLD, to 60 in common for GOLD and GOLD-S(int). Thus, the choice of protein-ligand scoring function can have an enormous effect on the top compounds identified by screening. Even the three protein-ligand scoring functions not normalized by ligand size, which all perform well at predicting ligand orientation (Tonero et al. in preparation), apparently either identify different features for scoring or weight their importance differently, as reflected by selecting different compounds as top-ranking. However, EON ligand-based electrostatic scoring using three different reference inhibitors tended to choose overlapping sets of compounds, despite the inhibitors' differences in the S1' and S2 pockets. There was 76-84% overlap (Table 2) between the top-100 compounds from EON using these three different references (Figure 2).

How chemically diverse were the top-scoring compounds for each scoring function? To assess this, each pair of 3-dimensional structures within the top-100 list was optimally aligned by ROCS (OpenEye Software, Santa Fe, NM) according to the Mills-Dean color forcefield, which measures 3D chemical similarity (Supplementary Materials Figure 1). The maximal similarity value is proportional to the size of the molecules. Thus, a pair of identical large molecules will score higher than a pair of identical small molecules, and the score of a pair of different-sized molecules will be limited by the size of the smaller molecule. As a result, the observed trends in chemical similarity within the top-100 sets can generally be explained by their molecular weight distributions: AffiScore/N compounds were less chemically similar to each other than were the EON top-100 compounds, and the protein-ligand scoring function compounds showed the highest degree of similarity within their top-100 lists. This data can also be used to help prioritize compounds for testing, either to select the compounds that are most chemically diverse within the top-scoring set, or to test compounds that are chemically related to confirmed inhibitors.

Trends in thrombin interactions for the top-100 ligands from different scoring functions

Does each scoring function favor certain ligand interactions with thrombin (as docked by SLIDE), and does it select similar interactions to known thrombin inhibitors from crystal complexes? To address these questions, matchprints that tabulate SLIDE-detected hydrophobic, hydrogen-bond, and salt-bridge interactions were used. Hydrogen bonds and salt bridges were listed as a contact with the corresponding polar atom. Unfavorable interactions were not included (though they could be), and all interactions were weighted equally. The predominant contacts of the 42 known inhibitors from the PDB (Supplementary Materials Table 1) were with thrombin Tyr 60A side chain, Trp 60D side chain, Leu 99 side chain, Ile 174 side chain, Asp 189 OD1 and OD2, Ala 190 O and side chain, ~Gly 193 N, ~Ser 195 N, Val 213 side chain, Ser 214 O, Trp 215 side chain, Gly 216 N and O, ~Glu 217 side chain, Gly 219 O, and ~Cys 220. "Side chain" indicates a hydrophobic contact between that side chain and the ligand, and "~" indicates

less common interactions, based on matchprint analysis of the 42 inhibitors from the PDB. Residues in thrombin are numbered by structural correspondence to chymotrypsin, a convention typically used for serine proteases.

For top-scoring EON compounds using the reference 1ppb inhibitor (one of the scoring methods used to select compounds for inhibition assays), interactions with Trp 60D side chain, Asp189 OD1 and OD2, Ala 190 O and CB, Glu 192 side chain, Gly 219 O and Cys 220 side chain were very common in the S1 pocket (Supplementary Materials Table 3). PPACK, the inhibitor in 1ppb, makes the same interactions, except for the Glu 192 and Cys 220 side chains, plus many additional interactions due to its larger size. (See 1ppb entry in Supplementary Table 1.) The same patterns of interaction with thrombin were found when PDB 1a5g and 1tbz inhibitors were used as EON references (Figure 2 and Supplementary Tables 4-5), despite their having very different chemical groups in the S1' pocket (which is surrounded by Leu 41, Cys 42, His 57, and Cys 58). The smaller ZINC compounds did not tend to contact these residues. All four highly ranked, confirmed inhibitors from EON/1ppb scoring were also found when 1tbz was used as a reference ligand, but one confirmed inhibitor (ZINC01692407) was lost from the top-100 list when 1a5g (with a larger, phenyl side chain in the S1' pocket) was used as reference.

Most of the contacts observed for known PDB inhibitors in the S1 specificity pocket were also made by the top-scoring compounds selected by each of the eight scoring functions (Supplementary Tables 3-10). Compounds that could spatially fit in this pocket were selected during screening by assigning SLIDE template points in the base of the S1 pocket as key points, requiring that each docked compound match at least one of these points. (Note that this spatial filtering was done for all docked compounds and scoring methods, whereas the filtering for the presence of amine or amide groups in S1 was used only to select a manageable number of compounds for experimental assays.) Asp 189 OD1 or OD2 and Ala 190 O and side-chain interactions in the S1 pocket were frequent in all top-scoring lists, though DrugScore (Supplementary Table 6) generally missed the Asp 189 OD1 interaction. A serious missed contact by all scoring functions was with Ser 214 O, near the amide N of the ligand residue binding in the S1 pocket, which is common for known inhibitors. The other major missed opportunities were with groups lining the S2 and S3 specificity pockets, due to the limited size of the ZINC compounds. As a result, interaction with the Trp 215 side chain in the S3 pocket, found for virtually all known inhibitors, was seldom observed for ZINC compounds. Gly 216 N and O (interacting with the amide group in many inhibitors between S2 and S3) and Glu 217 side chain interactions in S3 were also missed. In general, a greater number and diversity of interactions in the S1', S2, and S3 pockets were found for AffiScore (Supplementary Table 7) relative to EON or AffiScore/N (Supplementary Tables 3-5 and 8). While similar to AffiScore interaction preferences, DrugScore missed some key interactions of known PDB inhibitors in the S1 pocket, and selected compounds with more diverse and less common interactions in S1', S2, and S3 pockets. GOLD and GOLD-S(int) both showed a stronger preference for hydrophobic interactions relative to the compounds ranked highly by EON, AffiScore/N or AffiScore, and less preference for the polar interactions shared by many known PDB inhibitors (Supplementary Tables 9-10).

CONCLUSIONS

Structure-based screening can greatly enhance the success rate and decrease the cost relative to performing large *in vitro* screens, since far fewer compounds must be tested to discover lead compounds. Structure-based screening, which assesses protein-ligand interactions, also provides a predicted binding mode for each inhibitor. This is a powerful guide for hypothesis-driven optimization of the inhibitors for higher affinity and protein selectivity (e.g., binding to thrombin over other coagulation and digestive serine proteases). However, ligand-based scoring has become increasingly popular in high-throughput screening and is showing success in pharmaceutical applications (McGaughey et al. 2007). Here we have implemented hybrid scoring, using the docking mode provided by SLIDE to align each ligand candidate relative to a reference inhibitor for ligand-based scoring. Traditional ligand-based scoring methods both align and score the candidates relative to the reference ligand, while not necessarily guaranteeing that the ligands can fit in the binding site. In this study, five of the 26 soluble ligand candidates tested for binding proved to be inhibitors of human thrombin. This success rate is similar to those achieved by other groups, which typically also employ some post-scoring filtering (Shoichet et al. 1993; Massova et al. 1998; Fox and Haaksma 2000; Schapira et al. 2000; 2001; Doman et al. 2002; Gruneberg et al. 2002; Mizutani and Itai 2004; Polgar et al. 2005; Zhou et al. 2006). For comparison, experimental high-throughput screening without structure or ligand-based screening typically has a hit rate of less than 1% (Doman et al. 2002; Polgar et al. 2005; Marx et al. 2006; Schepetkin et al. 2006).

Using EON and AffiScore/N to rank ligand candidates led to the discovery of five new thrombin inhibitors, with the EON-based ranking of dockings yielding four out of the five hits, including the 0.6 μM inhibitor. SLIDE's normalized AffiScore yielded the smallest, thus potentially easiest to optimize, compound with mid-micromolar affinity. These new molecules can now be utilized as starting points for structure-based optimization to increase affinity and specificity. Optimization can benefit from matchprints, which define additional opportunities for making interactions that are characteristic of a number of known thrombin inhibitors. Structure-based AffiScore/N and ligand-based EON score selected top-100 compounds with different physicochemical properties, especially in molecular weight and polar surface area, and there was unexpectedly little overlap in chemical identity of compounds between the top-scoring lists from ligand-based and protein-ligand scoring functions. We were intrigued to find little overlap even among the top-100 compounds from the five different protein-ligand scoring functions. The success of combining protein-ligand docking with ligand-based scoring for inhibitor discovery, and their complementarity in scoring, encourages further testing on other protein targets.

The ZINC compounds that were assayed and proved to be thrombin inhibitors differed from non-inhibitors in the extent to which they made interactions that are conserved among previously known inhibitors. Thus, exhibiting the same interactions as known inhibitors correlated with the success of docked and scored compounds. Those interactions conserved among top-scoring compounds for a given scoring function, or conserved between the different scoring functions' top compounds, tended to be conserved interactions in known inhibitors. This was particularly true in the volume occupied by all docked compounds, the S1 pocket.

While both EON and AffiScore/N tended to favor compounds that made many interactions in common with known inhibitors, most of these compounds missed contacting Tyr 60A, Trp 60D, Leu 99, Ile 174, and Trp 215 in thrombin, which are key hydrophobic contacts found in PDB inhibitors. This knowledge can guide the addition of favorable functional groups to the novel inhibitor cores. For proteins with several available inhibitor or substrate-bound complexes, we also envision filtering high-scoring dockings for their ability to match a set of frequently observed interactions represented by matchprints. Alternatively, thermodynamic hot-spot data on critical sites of interaction could be represented as matchprints and used for filtering dockings. Combining scoring of protein-ligand interactions (which allows new kinds of compounds and interactions to be identified), ligand-based scoring (which favors molecules with overlapping shape and chemistry in 3-dimensional space, while not necessarily selecting structurally similar compounds), and matchprint-based filtering (which identifies compounds that can make a core set of interactions characteristic of known inhibitors) may represent the best of all possible worlds. This enhanced hybrid approach can include more information in screening and scoring, and may help decrease the false positive rate.

We observed that even scoring methods appearing quite similar in basis (e.g., protein-ligand scoring functions that measure hydrophobic and polar interactions) can select different sets of compounds as top candidates from screening. This indicates there is still work to be done in representing and measuring interactions in a physically meaningful way. Hybrid, step-wise scoring (as presented here, using EON scoring to rank docked compounds) and consensus scoring functions (in which independently developed scoring functions are applied concurrently to rank compounds) may be emphasizing the most relevant and accurately measured physical interactions from the individual scoring methods and canceling out the noisy features. Thus, identifying which scoring functions work well together across a number of molecular systems can also enhance our understanding of molecular recognition, by identifying features or feature representations to emphasize or de-emphasize in future scoring function development.

Acknowledgements

The authors sincerely thank Gerhard Klebe (University of Marburg, Germany) and Holger Gohlke (Christian-Albrechts-University, Kiel, Germany) for providing DrugScore, and OpenEye Scientific Software (Santa Fe, NM) for providing Omega, EON, FILTER, ROCS, and VIDA for our use. Enrico Di Cera and Leslie Bush from Washington University and David Huffman from Western Michigan University were generous in sharing thrombin assay protocols, and Michal Zolkiewski (Kansas State University) and Houria Hassouna (Michigan State University) were important contributors in the early stages of the experimental assays. We are grateful to the National Cancer Institute for preparing the Plated Compound Database in electronic format, and for providing samples of compounds for assays. We also thank Brian Shoichet and John Irwin (UCSF) for their ZINC database of available compounds, including the NCI/DTP Open Chemical Repository preprocessed as 3D structures. This work was supported by NIH grants GM-67249 and AI-53877 to L.A.K. and by start-up funds from Michigan State University to M.I.Z. The SLIDE screening and docking software, and its scoring functions in stand-alone form, are available for free academic use and commercial licensing; see Software under http://www.bch.msu.edu/~kuhn and contact KuhnL@msu.edu.

REFERENCES

Allen, F.H. 2002. The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallogr. B* **58:** 380-388.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. 2000. The Protein Data Bank. *Nucleic Acids Res.* **28:** 235-242.

Bode, W., Mayr, I., Baumann, U., Huber, R., Stone, S.R., and Hofsteenge, J. 1989. The refined 1.9 A crystal structure of human alpha-thrombin: interaction with D-Phe-Pro-Arg chloromethylketone and significance of the Tyr-Pro-Pro-Trp insertion segment. *Embo J* **8:** 3467-3475.

Bohm, H.J., Banner, D.W., and Weber, L. 1999. Combinatorial docking and combinatorial chemistry: design of potent non-peptide thrombin inhibitors. *J. Comput. Aided. Mol. Des.* **13:** 51-56.

Brass, L.F. 2003. Thrombin and platelet activation. *Chest* **124:** 18S-25S.

Das, J., and Kimball, S.D. 1995. Thrombin active site inhibitors. *Bioorg. Med. Chem.* **3:** 999-1007.

Davie, E.W., Fujikawa, K., and Kisiel, W. 1991. The coagulation cascade: initiation, maintenance, and regulation. *Biochemistry* **30:** 10363-10370.

Dekker, R.J., Eichinger, A., Stoop, A.A., Bode, W., Pannekoek, H., and Horrevoets, A.J. 1999. The variable region-1 from tissue-type plasminogen activator confers specificity for plasminogen activator inhibitor-1 to thrombin by facilitating catalysis: release of a kinetic block by a heterologous protein surface loop. *J. Mol. Biol.* **293:** 613-627.

Deng, Z., Chuaqui, C., and Singh, J. 2004. Structural interaction fingerprint (SIFt): A novel method for analyzing three-dimensional protein-ligand binding interactions. *Journal of Medicinal Chemistry* **47:** 337-344.

Doman, T.N., McGovern, S.L., Witherbee, B.J., Kasten, T.P., Kurumbail, R., Stallings, W.C., Connolly, D.T., and Shoichet, B.K. 2002. Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. *J. Med. Chem.* **45:** 2213-2221.

Evans, S.A., Olson, S.T., and Shore, J.D. 1982. p-Aminobenzamidine as a fluorescent probe for the active site of serine proteases. *J. Biol. Chem.* **257:** 3014-3017.

Fox, T., and Haaksma, E.E. 2000. Computer based screening of compound databases: 1. Preselection of benzamidine-based thrombin inhibitors. *J. Comput. Aided Mol. Des* **14:** 411-425.

Ganesh, V.K., Muller, N., Judge, K., Luan, C.H., Padmanabhan, R., and Murthy, K.H.M. 2005. Identification and characterization of nonsubstrate based inhibitors of the essential Dengue and West Nile virus proteases. *Bioorg. Med. Chem.* **13:** 257-264.

Gohlke, H., Hendlich, M., and Klebe, G. 2000. Knowledge-based scoring function to predict protein-ligand interactions. *J. Mol. Biol.* **295:** 337-356.

Gruneberg, S., Stubbs, M.T., and Klebe, G. 2002. Successful virtual screening for novel inhibitors of human carbonic anhydrase: Strategy and experimental confirmation. *J. Med. Chem.* **45:** 3588-3602.

Halgren, T.A. 1999. MMFFVI. MMFF94s option for energy minimization studies. *J. Comput. Chem.* **20:** 720-729.

Hopfner, K.P., Ayala, Y., Szewczuk, Z., Konishi, Y., and Di Cera, E. 1993. Chemical compensation in macromolecular bridge-binding to thrombin. *Biochemistry* **32:** 2947-2953.

Howard, N., Abell, C., Blakemore, W., Chessari, G., Congreve, M., Howard, S., Jhoti, H., Murray, C.W., Seavers, L.C., and van Montfort, R.L. 2006. Application of fragment screening and fragment linking to the discovery of novel thrombin inhibitors. *J. Med. Chem.* **49:** 1346-1355.

Irwin, J.J., and Shoichet, B.K. 2005. ZINC--a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **45:** 177-182.

Jain, A.N. 2003. Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. *J. Med. Chem.* **46:** 499-511.

Jorgensen, W.L. 2004. The many roles of computation in drug discovery. *Science* **303:** 1813-1818.

Katz, B.A., Elrod, K., Verner, E., Mackman, R.L., Luong, C., Shrader, W.D., Sendzik, M., Spencer, J.R., Sprengeler, P.A., Kolesnikov, A., et al. 2003. Elaborate manifold of short hydrogen bond arrays mediating binding of active site-directed serine protease inhibitors. *J. Mol. Biol.* **329:** 93-120.

Kellenberger, E., Rodrigo, J., Muller, P., and Rognan, D. 2004. Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins* **57:** 225-242.

Kettner, C., and Shaw, E. 1979. D-Phe-Pro-ArgCH2Cl – A selective affinity label for thrombin. *Thrombosis Res.* **14:** 969-973.

Mackman, N. 2008. Triggers, targets, and treatments for thrombosis. *Nature* **451:** 914-918.

Markwardt, F., Landmann, H., and Walsmann, P. 1968. Comparative studies on the inhibition of trypsin, plasmin, and thrombin by derivatives of benzylamine and benzamidine. *Eur. J. Biochem.* **6:** 502-506.

Marx, C., Berger, C., Xu, F., Amend, C., Scott, G.K., Hann, B., Park, J.W., and Benz, C.C. 2006. Validated high-throughput screening of drug-like small molecules for inhibitors of ErbB2 transcription. *Assay Drug Dev. Technol.* **4:** 273-284.

Massova, I., Martin, P., Bulychev, A., Kocz, R., Doyle, M., Edwards, B.F., and Mobashery, S. 1998. Templates for design of inhibitors for serine proteases: application of the program DOCK to the discovery of novel inhibitors for thrombin. *Bioorg. Med. Chem. Lett.* **8:** 2463-2466.

Matthews, J.H., Krishnan, R., Costanzo, M.J., Maryanoff, B.E., and Tulinsky, A. 1996. Crystal structures of thrombin with thiazole-containing inhibitors: Probes of the S1' binding site. *Biophys. J.* **71:** 2830-2839.

McGaughey, G.B., Sheridan, R.P., Bayly, C.I., Culberson, J.C., Kreatsoulas, C., Lindsley, S., Maiorov, V., Truchon, J.F., and Cornell, W.D. 2007. Comparison of topological, shape, and docking methods in virtual screening. *J Chem Inf Model* **47:** 1504-1519.

McGovern, S.L., Caselli, E., Grigorieff, N., and Shoichet, B.K. 2002. A common mechanism underlying promiscuous inhibitors from virtual and high-throughput screening. *J. Med. Chem.* **45:** 1712-1722.

McGovern, S.L., Helfand, B.T., Feng, B., and Shoichet, B.K. 2003. A specific mechanism of nonspecific inhibition. *J. Med. Chem.* **46:** 4265-4272.

Mills, J.E., and Dean, P.M. 1996. Three-dimensional hydrogen-bond geometry and probability information from a crystal survey. *J Comput Aided Mol Des* **10:** 607-622.

Mizutani, M.Y., and Itai, A. 2004. Efficient method for high-throughput virtual screening based on flexible docking: discovery of novel acetylcholinesterase inhibitors. *J. Med. Chem.* **47:** 4818-4828.

Nicholls, A., MacCuish, N.E., and MacCuish, J.D. 2004. Variable selection and model validation of 2D and 3D molecular descriptors. *J. Comput. Aided Mol. Des.* **18:** 451-474.

Nienaber, V.L., Boxrud, P.D., and Berliner, L.J. 2000. Thrombin inhibitor design: X-ray and solution studies provide a novel P1 determinant. *J Protein Chem* **19:** 327-333.

Oprea, T.I., and Matter, H. 2004. Integrating virtual screening in lead discovery. *Curr Opin Chem Biol* **8:** 349-358.

Peltason, L., and Bajorath, J. 2007. Molecular similarity analysis uncovers heterogeneous structure-activity relationships and variable activity landscapes. *Chemistry & Biology* **14:** 489-497.

Polgar, T., Baki, A., Szendrei, G.I., and Keseru, G.M. 2005. Comparative virtual and experimental high-throughput screening for glycogen synthase kinase-3beta inhibitors. *J. Med. Chem.* **48:** 7946-7959.

Rezaie, A.R., and Yang, L. 2003. Thrombomodulin allosterically modulates the activity of the anticoagulant thrombin. *Proc. Natl. Acad. Sci. U.S.A.* **100:** 12051-12056.

Sadasivan, C., and Yee, V.C. 2000. Interaction of the factor XIII activation peptide with alpha - thrombin. Crystal structure of its enzyme-substrate analog complex. *J. Biol. Chem.* **275:** 36942-36948.

Sanderson, P.E. 1999. Small, noncovalent serine protease inhibitors. *Med. Res. Rev.* **19:** 179-197.

Schapira, M., Raaka, B.M., Samuels, H.H., and Abagyan, R. 2000. Rational discovery of novel nuclear hormone receptor antagonists. *Proc. Natl. Acad. Sci. U.S.A* **97:** 1008-1013.

Schapira, M., Raaka, B.M., Samuels, H.H., and Abagyan, R. 2001. In silico discovery of novel retinoic acid receptor agonist structures. *BMC. Struct. Biol.* **1:** 1-7.

Schepetkin, I.A., Khlebnikov, A.I., Kirpotina, L.N., and Quinn, M.T. 2006. Novel small-molecule inhibitors of anthrax lethal factor identified by high-throughput screening. *J. Med. Chem.* **49:** 5232-5244.

Schnecke, V., and Kuhn, L.A. 1999. Database screening for HIV protease ligands: The influence of binding-site conformation and representation on ligand selectivity. *Proc. Int. Conf. Intell. Syst. Mol. Biol.***:** 242-251.

Schnecke, V., and Kuhn, L.A. 2000. Virtual screening with solvation and ligand-induced complementarity. *Perspectives in Drug Discovery and Design* **20:** 171-190.

Schneider, G., and Bohm, H.J. 2002. Virtual screening and fast automated docking methods. *Drug Discov. Today* **7:** 64-70.

Shoichet, B.K., McGovern, S.L., Wei, B., and Irwin, J.J. 2002. Lead discovery using molecular docking. *Curr. Opin. Chem. Biol.* **6:** 439-446.

Shoichet, B.K., Stroud, R.M., Santi, D.V., Kuntz, I.D., and Perry, K.M. 1993. Structure-based discovery of inhibitors of thymidylate synthase. *Science* **259:** 1445-1450.

St. Charles, R.S., Matthews, J.H., Zhang, E., and Tulinsky, A. 1999. Bound structures of novel P3-P1' beta-strand mimetic inhibitors of thrombin. *J. Med. Chem.* **42:** 1376-1383.

Stubbs, M.T., and Bode, W. 1993a. A model for the specificity of fibrinogen cleavage by thrombin. *Semin. Thromb. Hemost.* **19:** 344-351.

Stubbs, M.T., and Bode, W. 1993b. A player of many parts: the spotlight falls on thrombin's structure. *Thromb. Res.* **69:** 1-58.

Stubbs, M.T., and Bode, W. 1995. The clot thickens: clues provided by thrombin structure. *Trends Biochem. Sci.* **20:** 23-28.

Sukuru, S.C., Crepin, T., Milev, Y., Marsh, L.C., Hill, J.B., Anderson, R.J., Morris, J.C., Rohatgi, A., O'Mahony, G., Grotli, M., et al. 2006. Discovering new classes of Brugia malayi asparaginyl-tRNA synthetase inhibitors and relating specificity to conformational change. *J. Comput. Aided Mol. Des.* **20:** 159-178.

Tonero, M.E., Zavodszky, M.I., He, L., Van Voorst, J.R., Arora, S., Namilikonda, S., and Kuhn, L.A. in preparation. Effective scoring functions for predicting ligand binding mode and affinity in docking and high-throughput screening.

Tulinsky, A. 1996. Molecular interactions of thrombin. *Semin. Thromb. Hemost.* **22:** 117-124.

Verdonk, M.L., Cole, J.C., Hartshorn, M.J., Murray, C.W., and Taylor, R.D. 2003. Improved protein-ligand docking using GOLD. *Proteins* **52:** 609-623.

Wang, R., Lu, Y., and Wang, S. 2003. Comparative evaluation of 11 scoring functions for molecular docking. *J. Med. Chem.* **46:** 2287-2303.

Wang, R.X., Fu, Y., and Lai, L.H. 1997. A new atom-additive method for calculating partition coefficients. *Journal of Chemical Information and Computer Sciences* **37:** 615-621.

Waszkowycz, B. 2002. Structure-based approaches to drug design and virtual screening. *Curr. Opin. Drug Discov. Devel.* **5:** 407-413.

Wu, T.P., Yee, V., Tulinsky, A., Chrusciel, R.A., Nakanishi, H., Shen, R., Priebe, C., and Kahn, M. 1993. The structure of a designed peptidomimetic inhibitor complex of alpha-thrombin. *Protein Eng.* **6:** 471-478.

Zavodszky, M.I., and Kuhn, L.A. 2005. Side-chain flexibility in protein-ligand binding: the minimal rotation hypothesis. *Protein Sci.* **14:** 1104-1114.

Zavodszky, M.I., Sanschagrin, P.C., Korde, R.S., and Kuhn, L.A. 2002. Distilling the essential features of a protein surface for improving protein-ligand docking, scoring, and virtual screening. *J. Comput. Aided Mol. Des.* **16:** 883-902.

Zhou, Y., Peng, H., Ji, Q., Qi, J., Zhu, Z., and Yang, C. 2006. Discovery of small molecule inhibitors of integrin alphavbeta3 through structure-based virtual screening. *Bioorg. Med. Chem. Lett.* **16:** 5878-5882.

Table 1. Experimentally confirmed thrombin inhibitors identified among the NCI/DTP Open Chemical Repository of compounds. "E" denotes high ranking according to EON scoring, and "A" denotes high ranking according to normalized AffiScore.

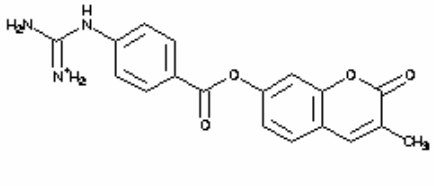
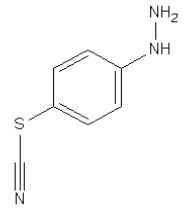
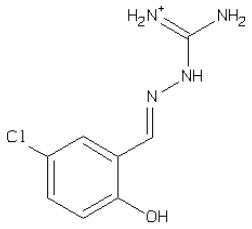
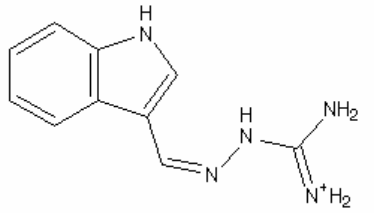
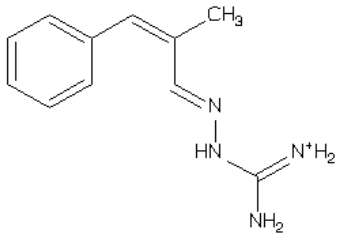
#	NCI/ ZINC Code	Chemical Name	2D Structure	IC_{50} (μM)	Scoring protocol
1	NSC162404 01627553	3-methyl-2-oxo-2H-chromen-7-yl 4-((amino(imino)methyl) amino)benzoate		0.635	E
2	NSC164904 00393806	4-hydrazinophenyl thiocyanate		430	A
3	NSC65385 01692407	2-(5-chloro-2 hydroxybenzylidene) hydrazine carboximidamide		463	E
4	NSC67122 01694232	2-(1H-indol-3-ylmethylene) hydrazine carboximidamide		663	E
5	NSC65811 01280746	2-(2-methyl-3-phenyl-2-propenylidene) hydrazine carboximidamide		719	E

Table 2. Number of compounds in common between the top-100 lists for each pair of scoring functions used to rank SLIDE dockings.

	EON/1a5g	EON/1ppb	EON/1tbz	AffiScore/N	AffiScore	DrugScore	GOLD	GOLD–S(int)
EON/1a5g	100	76	81	3	0	0	0	0
EON/1ppb	76	100	84	3	0	0	0	0
EON/1tbz	81	84	100	3	0	0	0	0
AffiScore/N	3	3	3	100	0	0	0	0
AffiScore	0	0	0	0	100	13	7	11
DrugScore	0	0	0	0	13	100	24	32
GOLD	0	0	0	0	7	24	100	60
GOLD –S(int)	0	0	0	0	11	32	60	100

FIGURE LEGENDS

Figure 1. Workflow of docking, ranking, and assaying compounds for thrombin inhibition.

Figure 2. Three known inhibitors used as reference ligands for EON scoring, from PDB entries (A) 1a5g, (B) 1ppb, and (C) 1tbz. The phenyl groups of all three compounds, which bind in the S3 pocket, are shown at upper left. They are followed by proline (1ppb and 1tbz) or another backbone restraining group (1a5g; lower left) binding in the S2 pocket, then a guanidinium group (upper right) binding in the S1 pocket. A large aromatic group pocket in 1a5g and 1tbz (far right) binds in the S1' pocket, whereas PPACK (in 1ppb) has no side chain in this position.

Figure 3. Enzyme activity as a function of inhibitor concentration (on a logarithmic scale) for (A) compound 5 and (B) the known inhibitor, *p*-aminobenzamidine. Using the sigmoidal fit utility of Origin 7.5, the IC_{50} value was determined as the inhibitor concentration corresponding to 50% enzymatic activity.

Figure 4. Assay for aggregating inhibitor behavior, monitoring any significant change in inhibition observed when detergent is added to the assay solution. DMSO was used as a negative control (it causes no inhibition), and the enzymatic inhibition assay of compound 1 was repeated in the presence of 0.02 % Triton in the reaction mixture, to disintegrate any possible aggregates of compound 1. Such aggregates could indirectly inhibit the enzyme by engulfing the thrombin molecules. No changes in enzyme inhibition were observed upon adding Triton, indicating that inhibition was specific, not by aggregation of the compound around thrombin. The same result was observed for the other five compounds in Table 1.

Figure 5. SLIDE-predicted binding modes of the confirmed inhibitors (carbon atoms colored in magenta) compared with known thrombin inhibitors with similar binding modes defined by crystallography (carbon atoms colored in green): (A) compound 1 with ligand 1a46, (B) compound 2 with ligand from PDB entry 1dwd, (C) compound 3 with ligand from PDB 1a46, (D) compound 4 with ligand from PDB 1ca8, and (E) compound 5 with ligand from PDB 1ppb. All are shown in the ligand-free binding site of thrombin (PDB 1vr1) used for screening. The PDB inhibitor shown for reference in each figure is one with a similar binding mode to the ZINC compound in the specificity pockets that are occupied by both compounds. The scoring protocols favored identifying smaller inhibitors than the PDB inhibitors shown for reference.

Figure 6. Comparison of properties of the top-100 compounds selected from eight different scoring functions. (A) For each molecular weight range, 100-150D, 150-200D, etc., the number of top-ranking compounds that falls within that range is given for each scoring function. (B) Within each polar surface area range, 20-30 $Å^2$, 30-40 $Å^2$, etc., the number of top-ranking compounds is also given for each scoring function.