PROTEINS: Structure, Function, and Genetics 42:296–318 (2001)

Design, Docking, and Evaluation of Multiple Libraries Against Multiple Targets

Michelle L. Lamb,[1] Keith W. Burdick,[2] Samuel Toba,[1] Malin M. Young,[1] A. Geoffrey Skillman,[1] Xiaoqin Zou,[1] James R. Arnold,[1] and Irwin D. Kuntz[1]*

[1]*Department of Pharmaceutical Chemistry, School of Pharmacy, University of California, San Francisco, California*
[2]*Graduate Group in Biophysics, University of California, San Francisco, California*

ABSTRACT **We present a general approach to the design, docking, and virtual screening of multiple combinatorial libraries against a family of proteins. The method consists of three main stages: docking the scaffold, selecting the best substituents at each site of diversity, and comparing the resultant molecules within and between the libraries. The core "divide-and-conquer" algorithm for side-chain selection, developed from an earlier version (Sun et al., J Comp Aided Mol Design 1998;12:597–604), provides a way to explore large lists of substituents with linear rather than combinatorial time dependence. We have applied our method to three combinatorial libraries and three serine proteases: trypsin, chymotrypsin, and elastase. We show that the scaffold docking procedure, in conjunction with a novel vector-based orientation filter, reproduces crystallographic binding modes. In addition, the free-energy-based scoring procedure (Zou et al., J Am Chem Soc 1999;121:8033–8043) is able to reproduce experimental binding data for P_1 mutants of macromolecular protease inhibitors. Finally, we show that our method discriminates between a peptide library and virtual libraries built on benzodiazepine and tetrahydroisoquinolinone scaffolds. Implications of the docking results for library design are explored. Proteins 2001;42:296–318.**

Key words: combinatorial library; docking; scoring; drug design; serine protease; screening

INTRODUCTION

The essential task of medicinal chemistry is discovering new therapeutic agents. A critical step is finding novel lead compounds against specific targets. Historically, discovery was a serial process, screening and optimizing compounds one at a time. The advent of combinatorial chemistry and high-throughput screening has significantly altered the approach to drug discovery. The rapid generation and screening of combinatorial libraries is now a common undertaking in both academic and industrial settings. This approach has resulted in significant savings of time and resources in selecting potent inhibitors as potential drug candidates.[3–5] However, the hit rates when screening unbiased libraries are often very low, and many libraries have not yet been associated with interesting biological activity. This implies a need for a rational means of customizing combinatorial libraries for particular targets and for screening efficiently one or more libraries against a large number of targets.

One strategy for screening is to use computational methods. Structure-based drug design has been shown to work well for retrieving potential ligands from databases.[6,7] It can also be used in concert with combinatorial chemistry. Several algorithms for the structure-based design and evaluation of combinatorial libraries have been developed,[1,8–10] and the advantage of structure-based, directed library design over designs based only on molecular diversity has been shown.[8] It has been estimated[11] that the number of organic compounds with molecular weights $<$ 750 a.m.u. approaches 10^{200}, whereas 10^{12}–10^{15} is a practical limit on the size of a library that may be screened on a computer.[12] The challenge is to search this vast chemical space for molecules with the desired properties. The virtual library must be reduced to a smaller, syntheti-

Abbreviations: Standard three- and one-letter amino acid abbreviations are used, with the exception of Ace, acetyl and NMe, N-methyl peptide capping groups and non-natural amino acids (numbers refer to heavy atoms of side chain): Abu (2), α-aminobutyric acid; Ape (3), α-aminopentanoic acid = norvaline; Ahx (4), α-aminohexanoic acid = norleucine (Nle); Ahp (5), α-aminoheptanoic acid; Hse (hS), homoserine = α-amino-γ-hydroxybutyric acid; OMTKY3, turkey ovomucoid third domain; BPTI, bovine pancreatic trypsin inhibitor (Kunitz); C/E-1, *Ascaris* chymotrypsin-elastase inhibitor 1; PPT, tripeptide scaffold/library; BZP, 1,4-benzodiazepin-2-one scaffold/library; THQ, tetrahydroisoquinolinone scaffold/library; CHYM, bovine chymotrypsin; PPE, porcine pancreatic elastase; TRYP, bovine trypsin; RMSD, root mean squared deviation; SAR, structure-activity relationship; GB/SA, generalized-Born/buried surface area

The Supplementary Material referred to in this article can be found at http://www.interscience.wiley.com/jpages/0887-3585/42_3/296/v42.3.296.htm.

Grant sponsor: National Institutes of Health; Grant numbers: GM31497, GM56531, and CA 72006.

Michelle L. Lamb's present address is Molecular Design Group, DuPont Pharmaceuticals Research Laboratories, San Francisco, CA 94111.

Geoffrey Skillman's present address is OpenEye 335c Winische Way, Sante Fe, NM 87501. http://www.eyesopen.com.

Malin M. Young's present address is Biosystems Research, Sandia National Laboratories, Livermore, California.

Xiaoqin Zou's present address is Dalton Cardiovascular Research Center and Department of Biochemistry, University of Missouri, Columbia, Missouri.

*Correspondence to: Irwin D. Kuntz, 533 Parnassus Avenue, U-80, P.O. Box 0446, San Francisco, CA 94143-0446. E-mail: kuntz@cgl.ucsf.edu

Received 23 March 2000; Accepted 26 September 2000

cally feasible, library while retaining the compounds most likely to be active against a given target.

Designing and evaluating multiple libraries against a single target is conceptually straightforward and will be considered in detail below. More attention is required for organization of targets. Just as ligands in a combinatorial library are related through a common scaffold, it is accepted that most proteins can be grouped into families possessing common geometric frameworks (folds) and common functional residues. Moreover, the same compounds often inhibit proteins in the same family; hence, specificity is a constant challenge for medicinal chemists. The positive side is that a library could be targeted to probe an entire protein family rather than a single protein target. Many libraries that have activity against multiple members of a protein family have already been synthesized. For example, libraries have been synthesized that are active against kinases,[13] aspartyl proteases,[14] and metalloproteases.[15] Generalized, this strategy would lead to probe libraries directed against arrays of biochemical activities.

This vision leads naturally to the proposal of simultaneous virtual screening of many libraries against many protein targets. It has been suggested that proteins can be described with a profile of SAR data for different ligands,[16,17] and the reciprocal process of profiling ligands by describing their binding to a set of targets has also been performed.[18,19] Libraries could be subjected to a similar analysis. They would be placed on a continuum ranging from universal libraries that contain hits against every target to completely directed libraries that contain hits against only one target. Libraries with greater breadth may increase the economy with which information about binding preferences is gathered in the lead discovery process. In subsequent optimization phases, libraries with narrow specificities may be desirable as the focus on well-defined properties becomes germane. Having the knowledge of how virtual libraries perform across a spectrum of targets should lead to a much broader understanding of the advantages and limitations of combinatorial strategies and unavoidable side effects.

Major concerns for large-scale target-based approaches to virtual screening are (a) the availability of target structures, (b) the accuracy of the scoring function, and (c) the computational resources required. Coordinates for nearly 10,000 crystal structures have been deposited in the Protein Data Bank to date,[20] and the number of solved protein structures increases daily. In addition, structural genomics and homology modeling efforts promise to provide structural models for a significant fraction of the known sequences.[21] The performance of different scoring functions has been extensively covered by Charifson et al.[22] For this study, we adopted a scoring function based on a force field and a solvation correction that uses the generalized-Born/buried surface area (GB/SA) formulation of Still et al.[23] and Zou et al.[2] We show that this scoring function used here can accurately reproduce the experimental results. Currently, it remains a formidable computational task to screen known libraries against multiple targets. In 1999 alone, ~300 combinatorial libraries were published.[24,25] Let us assume each library has three points of diversity for which 100 substituents must be evaluated. By using existing docking algorithms, which require ~100 s/mol for flexible docking, and a single processor, the computational time needed to screen these libraries against all known protein structures is on the order of 10^7 years. Clearly, a better approach is needed.

In this article, we present an accurate, efficient, and general target-based approach to the design, docking, and virtual screening of multiple libraries against multiple targets. Our method consists of three main stages: docking the scaffold, selecting the best substituents at each site of diversity, and comparing the resultant molecules within and between the libraries. We make repeated use of this divide-and-conquer strategy and incorporate judicious pruning of the library at each step of the library design. It is fast enough to allow a survey of dozens of libraries against hundreds of targets on a small multiprocessor system.

For demonstration purposes and to provide a careful evaluation of all phases of the library design, we applied our method to three combinatorial libraries and three proteins from the serine protease family: trypsin, chymotrypsin, and elastase. This family of enzymes has been well studied; particularly relevant is the abundance of structural and thermodynamic information, which includes a wide variety of both covalent and macromolecular inhibitors.[26] Laskowski[28–30] and Otlewski[31] and their respective coworkers have investigated the detailed thermodynamics of the recognition macromolecular inhibitors within the primary specificity subsite (S_1 in Schechter and Berger notation[27]) of each protease.[28–31] Mutations at the P_1 position in ovomucoid third domain (OMTKY3) and pancreatic trypsin inhibitor (BPTI) include all the natural amino acids. Thus, we have an excellent basis for comparison of predicted and experimental binding affinities for a set of peptide-like substituents (Fig. 1).

We first establish that our scoring compares well to the experimental results for the macromolecular inhibitors. We then show that our procedure for peptide scaffold docking retains geometries analogous to the crystal structures and that the substituent evaluation phase suggests appropriate side chains for the S_1 site. We next explore two virtual libraries (Fig. 1), a benzodiazepine library (BZP) and a tetrahydroisoquinolinone library (THQ), that share the substituent set of the peptide library. Finally, we show that our scoring discriminates between the peptide library (PPT) and the two non-peptide libraries.

MATERIALS AND METHODS

Our protocol is shown in Figure 2. As indicated above, the basic strategy uses docking to place minimal scaffolds. The best scaffold positions are then elaborated with flexible side chains in two stages—first as single substituents and then as full molecules. A number of parameter choices have been made, both within the protocol and within the DOCK 4.0.1 program,[32,33] which permit us to test many aspects of library design. We have adopted several "greedy algorithm" filters to improve efficiency. We have also made

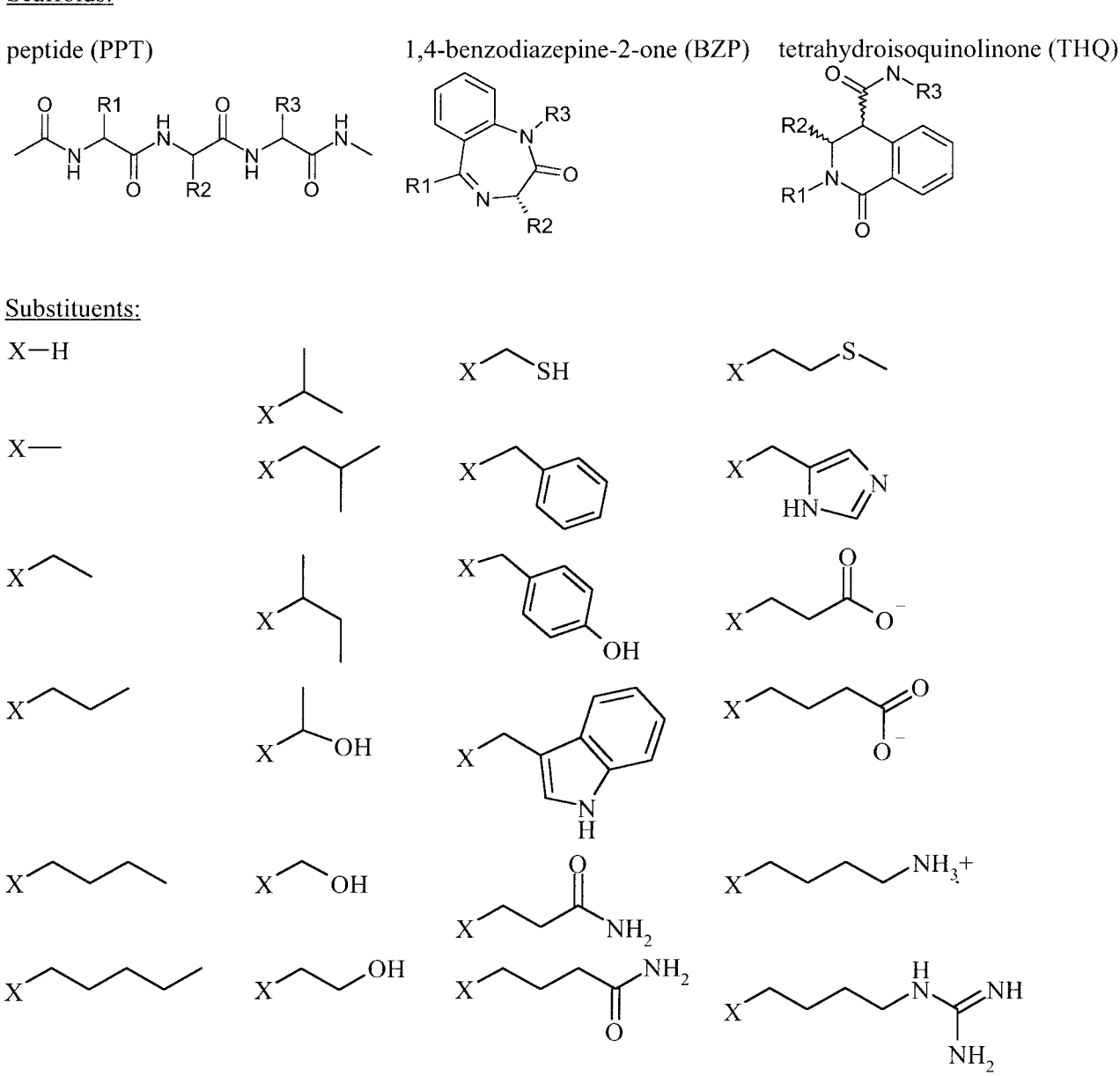

Fig. 1. Scaffolds and amino acid side chains for combinatorial libraries. X denotes the point of attachment to the scaffold position.

use of the SYBYL[34] software and programming language to build molecules. Alternative software and scripting languages, such as the Daylight Toolkit,[35] could be used to carry out the same operations.

We divide this section into four subsections: target selection, preparation, and site characterization; library preparation; the docking protocol; and scoring validation.

Target Selection, Preparation, and Site Characterization

We used several criteria for selecting the serine protease structures used in this study. The crystal structures chosen were of high atomic resolution (1.9–2.1 Å) and were complexed with non-covalent, substrate-like inhibitors, to ensure that the active sites were well formed and not biased toward recognition of irreversible inhibitors. Structures selected were chymotrypsin-APPI (P_1 = Arg, 1CA0),[36] elastase-elafin (P_1 = Ala, 1FLE),[37] and trypsin-A90720A, which has an arginine-like substituent in the P_1 position (1TPS).[38] When these structures are compared with the large number of other ligand-bound structures for these enzymes, we see relatively little perturbation in the S_1 site arising from induced fit except for elastase where the Ala-S_1 pocket is somewhat smaller than that with a leucine ligand (see below). The inhibitors were removed,

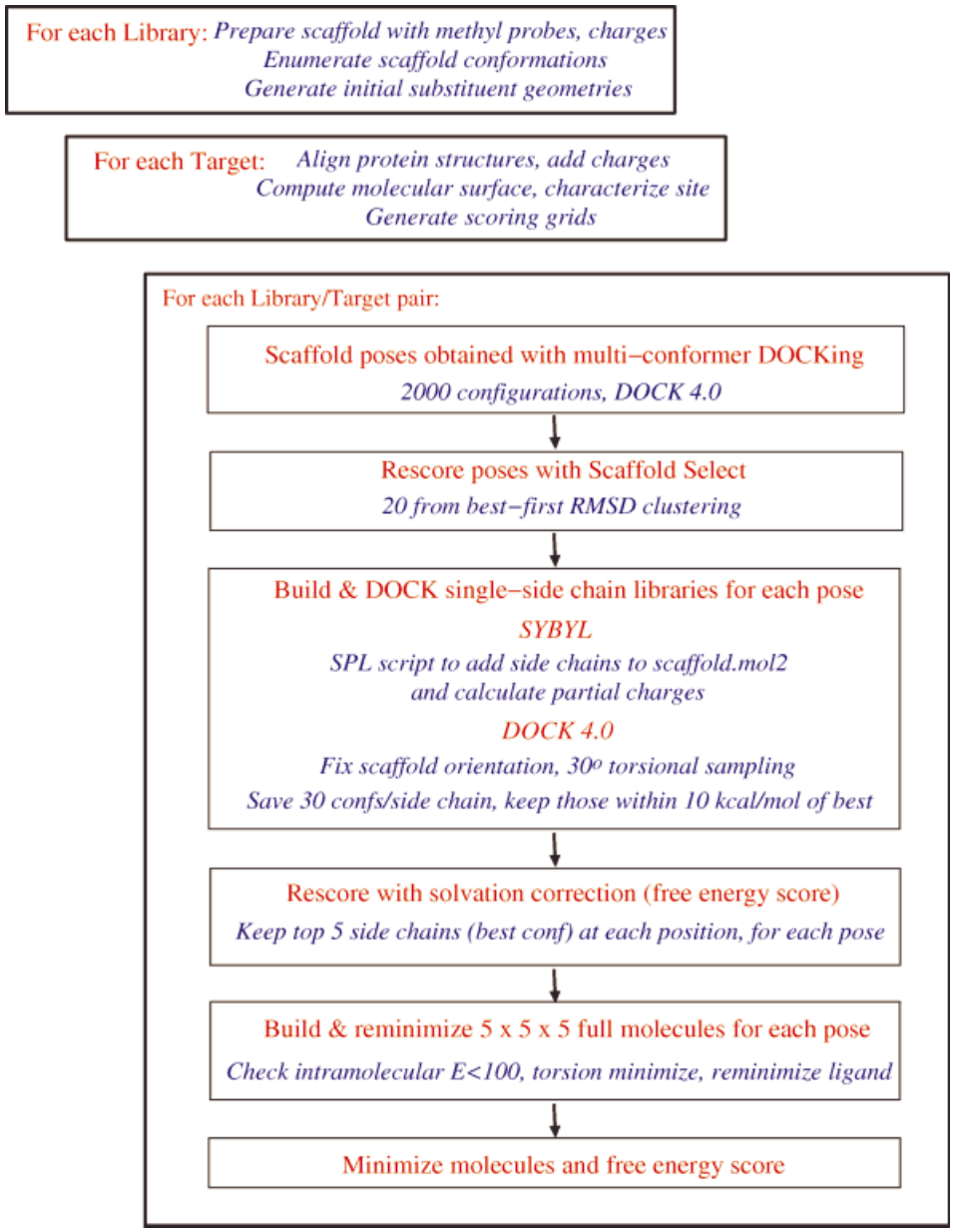

Fig. 2. Protocol for docking multiple libraries against multiple targets.

along with any crystallographic water molecules. The three proteins were then aligned by matching heavy atoms of the catalytic triad (Ser 195, Asp 102, and His 57) to those of trypsin. Hydrogens, disulfide bonds, and AMBER charges[39] were added via SYBYL, resulting in net charges of +6, +3, and +4 for trypsin, chymotrypsin, and elastase, respectively. A 0.15 Å scoring grid was generated for each target by using the *grid* module of DOCK 4.0.[40] In preliminary calculations, we found that the 0.15 Å grid spacing gave a balanced compromise between computational speed and accuracy. DOCK force field is based on the AMBER[39] intermolecular force field of VDW and electrostatic energy (Eq. 1):

$$E = \sum_{i=1}^{lig} \sum_{j=1}^{rec} \left[\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} + 332.0 \frac{q_i q_j}{\varepsilon r_{ij}} \right]. \quad (1)$$

The active sites for each protease were further characterized to generate a template for the orientation search in DOCK. The molecular surface[41–43] for each protein was computed by using the *dms* program distributed with MidasPlus.[44] Spheres that fill surface indentations were calculated (within 10 Å of the catalytic serine OG atom) with the program SPHGEN.[45,46] This sphere set (N_{spheres} = ~500–800) was subjected to several steps of enrichment and pruning to improve catalytic site representation. First, the heavy atoms of the crystallographic inhibitor

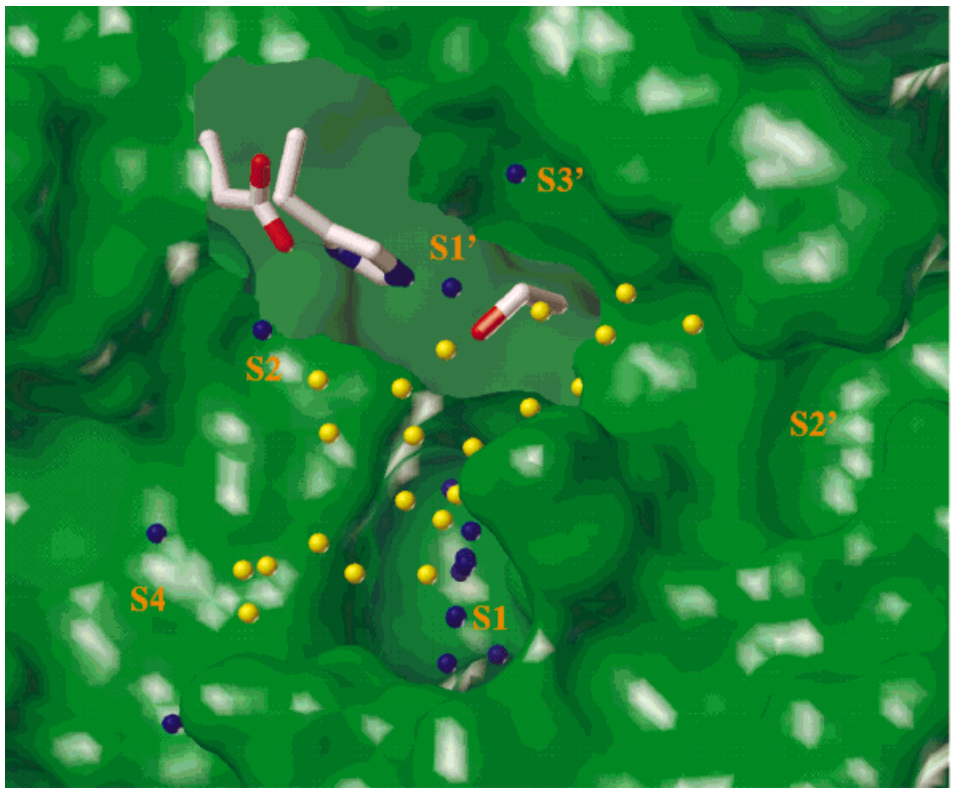

Fig. 3. The active site of chymotrypsin, with catalytic residues Asp102, His57, and Ser195 displayed. Yellow spheres fill the space appropriate for scaffold docking, whereas blue spheres fill the labeled subsites. The MOLCAD-generated surface[34] of the protein is shown in green.

were added to the list of sphere centers. Next, the resultant sphere set was filtered by DOCK contact score[47]; spheres with a contact score ≤ -1.0 units were retained. Further reduction was obtained by clustering the spheres into pairs of nearest neighbors and removing the sphere in each pair with the more positive contact score. If sphere centers close to the inhibitor P_1 residue carbonyl oxygen atom were removed, a site point representing this atom was added back into the sphere set to mark the protease "oxyanion hole." At this stage, the active site of each protease was populated with 35–40 spheres. Finally, to focus the scaffold orientation search away from protease subsites (pockets), spheres in these regions were manually deleted from the sphere set for each protease. These "scaffold" sphere sets contained ~25 spheres each (Fig. 3).

Library Preparation

Minimal scaffold

Initial scaffold geometries with methyl groups at each of the three diversity or attachment sites were created by using CONCORD[48,49] within the SYBYL suite of programs. Hydrogens were added, and Gasteiger-Marsili charges were assigned to the methylated scaffold. We used a multiconformer approach for scaffolds. Scaffold conformations (up to 1,500) were generated with DOCK in a target-independent fashion by using a torsion-driving routine. The key variables set for this procedure were flexible_ligand, torsion_drive, clash_overlap (0.35), and intramolecular_score. Bond lengths, bond angles, and ring conformations are not modified. The torsion-driving algorithm uses a look-up table of suitable torsion angle values for each torsion type, given in the default files, flex.defn and flex_drive.tbl.[47] For example, as the default, the dihedral angle through an sp^3-sp^3 bond would sample *gauche*$^+$ (60°), *gauche*$^-$ (−60°) and *trans* (180°) conformations. Ligand conformational energies are evaluated as intramolecular non-bonded energies only; no explicit torsional energy barriers are added.

Full side chains

For all of these studies, we have used the set of natural and non-natural amino acid side chains shown in Figure 1. These side chains were prepared in SYBYL with the initial CONCORD geometries (from the α-carbon, labeled *X*, onward) written to a mol2 file for later attachment to a scaffold (see Docking Protocol, below).

Docking Protocol

Our protocol for docking multiple combinatorial libraries to multiple targets uses DOCK as shown schematically in the flowchart of Figure 2. In our force field calculations, a 10 Å interatomic distance cutoff and distance-dependent dielectric ($\varepsilon = 4r$) were used. All-atom parameters for both proteases and ligands were applied. Once the side

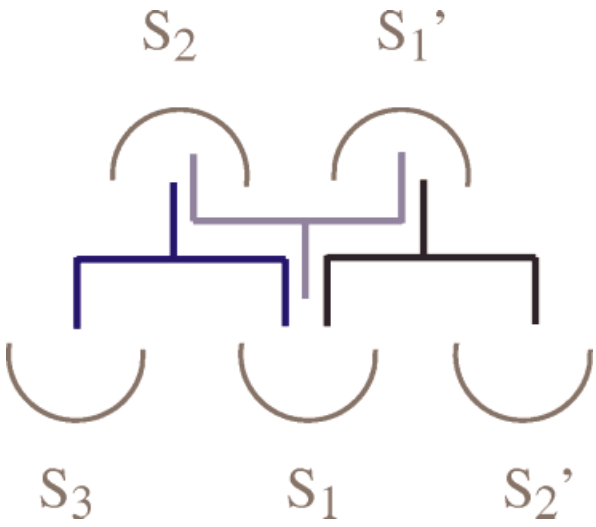

Fig. 4. A schematic of the three possible in-register tri-substituted scaffold orientations within serine protease subsites S_3 through S_2'.

chains had been sampled and evaluated with the force field, the best scoring conformations were reevaluated with a continuum solvent model-based scoring function[2] to account for the desolvation effects. Below, we enumerate the docking details at each stage of the protocol.

1. Scaffold docking. One complication of the protease active sites is that a tripeptide can dock in multiple registrations to the specificity pockets within the active site (Fig. 4). Because our ultimate goal is a general method for the design of libraries, we wanted to explore alternative geometries, which span the S_3-S_2' subsites. In early stages of scaffold docking, however, we observed many docked orientations in which the acetyl or N-methyl portion of the scaffold itself had been placed into a subsite. Although this may be a reasonable conformation for the peptide to adopt, especially in the case of a small S_1 pocket such as that of elastase, it does not provide a productive orientation for library designs. Hence, we developed alternative filtering mechanisms that limit the scaffold configurations to ones expected to result in good side chain presentations.

We started with a database of scaffold conformations, described above. This database was rigidly docked to the set of "scaffold" site spheres (Fig. 3) of each protease active site, so that the scaffold was explicitly not oriented within the subsite cavities of the active site. The automated-matching DOCK 4 algorithm was used in these calculations, and Simplex minimization (initial translation 1.5, maximum iterations 50) was applied to the docked scaffold orientations.[32,33] The 2,000 best orientations for each scaffold conformation were retained. Finally, the scaffold intramolecular score and scaffold-protein intermolecular score were added together, and the scaffold "poses" (conformations and orientations) were ranked accordingly. These ranked poses for each target were then filtered based on surface normal vector scoring (below).

2. Vector scoring of scaffolds with "Scaffold Select." The Scaffold Select method is designed to find productive scaffold orientations in protein targets with defined subsite pockets such as the proteases studied here. The set of 2,000 scaffold poses for each scaffold target pair was pruned down to 20 in the following manner (Fig. 5). The molecular surface of the protein, originally generated for use with SPHGEN, also contains unit vectors normal to the surface at each point.[41–43] This set was processed to obtain the set of vectors within 15 Å of the catalytic serine hydroxyl oxygen atom (e.g., for chymotrypsin, N_{vectors} = 8422). The unit vectors were extended in 1 Å increments to a maximum length of 15 Å. Vectors that intersected the surface of the protein or which projected from convex surfaces were pruned, leaving behind vectors local to concave regions of the active site (N_{vectors} = 1515). Next, our "Scaffold Select" program was used to identify the set of vectors that were complementary to each substituent attachment bond for each scaffold configuration (i.e., for the peptides, the C_α—C_β bonds). A vector was deemed complementary if its closest-approach distance to a bond segment was $\leq$ 1.0 Å and its angle of approach was between 120 and 180°. For each complementary vector, the number of protein atoms within 3.4 Å of the vector was calculated, yielding a "vector contact" score. If more than one vector matched an attachment segment, the vector with the highest contact score was selected as the set representative. Only scaffold poses having vectors complementary to all three attachment segments were then retained (e.g., N_{poses} = 729 for the peptide scaffold in chymotrypsin).

The score for each scaffold pose was determined by using an empirically weighted sum of vector-protein contacts and hydrogen bonds between the protease and scaffold (Eq. 2):

$$\text{Vector Score} = 0.1*(\text{vector contacts}) + 0.5*(\text{hydrogen bonds}). \quad (2)$$

Schnecke and coworkers[50] use a similar scoring function for peptide docking, in that it also weights the number of hydrogen bonds formed and a hydrophobic complementarity term (weighting = 1:1.2). This function could be modified for systems in which hydrogen bonding is less important for ligand recognition. For our purposes, hydrogen bonds were defined as three-point representations between donor atom (D), hydrogen atom (H), and acceptor atoms (A), the H $\cdots$ A distance $\leq$ 2.8 Å and an D-H $\cdots$ A geometric angle between 90 and 180°. The scaffold poses were rank-ordered by this vector score and then RMSD clustered (2.0 Å cutoff) by using a best-first algorithm to increase diversity within the set (N_{clusters} = 131, peptide-chymotrypsin). Only the cluster heads for the first 20 clusters were carried forward to the next step.

3. Addition and incremental growth of side chains at each position on each scaffold pose. Each substituent was joined to each scaffold pose for each target at a single diversity position at a time, by using SYBYL. For example, three sublibraries of Ace-Xaa-Ala-Ala-NMe, Ace-Ala-Xaa-Ala-NMe, and Ace-Ala-Ala-Xaa-NMe (where Xaa is the natural or non-natural amino acid) were created for each of the 20 peptide poses. Charges for each molecule were recomputed as well. After

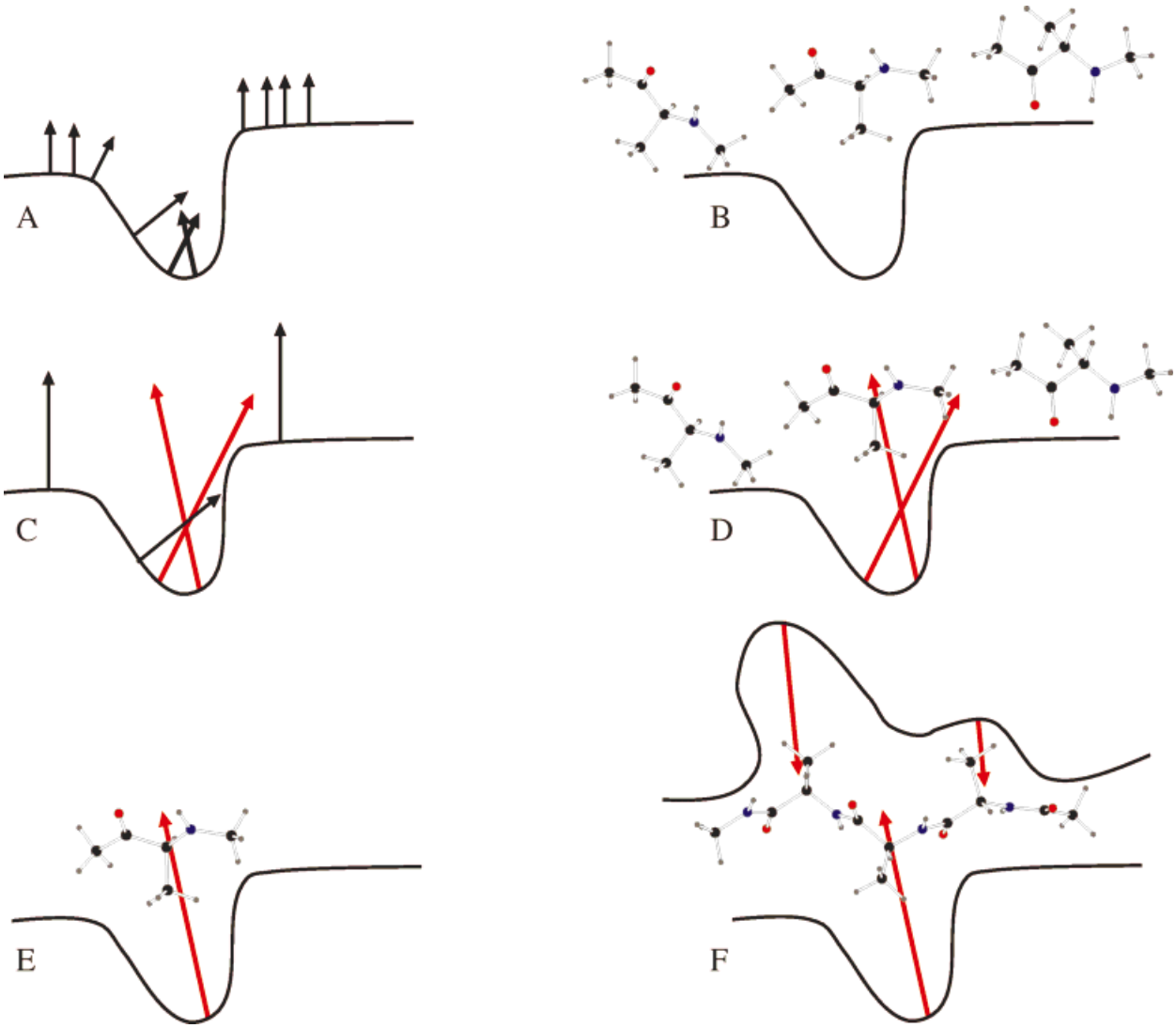

Fig. 5. Evaluation of scaffold orientations using a surface normal vector method. A: Compute molecular surface and unit normal vectors, as normal preparation of active site for docking. B: DOCK scaffold (here a single peptide unit) into active site. C: Extend the unit normal vectors; red vectors survive the filtering process. D: Determine vector matches for each scaffold. E: Evaluate vector score for each scaffold (Eq. 2). F: Represents the evaluation of a full tripeptide scaffold, matching three vectors. (See text for further details.)

this step, conformations of each side chain in each sublibrary were sampled with DOCK. In this calculation, the scaffold atoms were designated as the rigid "anchor" for each molecule, and no orientation search or minimization of the scaffold pose was performed. Each added side chain was flexible, however, and enhanced dihedral angle sampling was used (e.g., a C—C single bond was sampled in 30° increments). A clash-overlap ratio of 0.3 and torsional minimization were allowed; the minimization included bonds two layers before the currently sampled rotatable bond. Up to 100 conformations of each side chain were saved during each cycle of growth, and a maximum of 30 conformations per side chain was written out once growth was completed. The conformations of each side chain were ranked based on intra- and intermolecular score, whereas the molecules in each sublibrary were compared with one another on the basis of intermolecular score alone. Finally, for each side chain, only conformations that were within 10 kcal/mol of the best scoring conformer were retained for use in the next step.

4. Free energy scoring of single-side chain sublibraries. Estimating desolvation energies of ligands and protein active sites is essential for calculations of binding affinity.[2] Currently, such calculations (using a GB/SA model) are too computationally expensive to serve as the primary scoring method. However, they are feasible as a post-processing filter on DOCK results. No minimization was performed at this stage. The empirical parameters of $\beta = 0.6$, $\sigma_1 = 0.025$, $\sigma_2 = 0.02$ were used in Equation 3, where SA_{hp} and SA represent the hydrophobic and total solvent-accessible surface areas, respectively; VDW is the van der Waals interaction between protein and ligand; and G_{pol} uses the

generalized Born equation to represent the electrostatic contribution to the free energy.[2]

$$\Delta G_{\text{binding}} = \sigma_1 \Delta(\text{SA}_{\text{hp}}) + \beta^*(\text{VDW}) - \sigma_2 \Delta(\text{SA}) + G_{\text{pol}}. \quad (3)$$

Because the S_1 subsite in each protease is the major determinant of specificity, a penalty term for molecules that do not adequately fill this cavity was included in the calculation of G_{pol}; the P_1 side chain atoms of the crystallographic ligand were used to define the cavity. The details of the penalty term, as well as the implementation of the free energy scoring method, have been described previously by Zou et al.[2]

5. Merging and minimizing $5 \times 5 \times 5$ sublibraries. In the previous steps, we docked and scored isolated scaffolds and evaluated individual side chains. Next, we generated complete molecules by selecting the best five side chains from free energy scoring at each attachment position (step 4). Because the scaffold position for each pose had not changed, the methyl groups at diversity positions could be replaced with the best conformer of each of the selected side chains using SYBYL. A fast intramolecular energy calculation was performed to remove any full-built molecules that contained a clash (i.e., DOCK force field score $>$ 100 kcal/mol). The remaining molecules ($\sim$ 70–125 per scaffold pose) in the sublibrary were then energy minimized to remove any further bad intramolecular contacts.

6. Final free energy minimization of $5 \times 5 \times 5$ sublibraries. In this last stage, the sublibraries generated in step five were flexibly minimized against the solvation scoring function, and the final free energy scores were calculated as described for step 4 above.

Scoring Validation

Protein structures for these calculations were chosen to closely parallel the protease-inhibitor complexes studied experimentally.[28,31] For bovine chymotrypsin, binding data were available for P_1 mutant OMTKY3 and BPTI inhibitors, as were crystal structures of the protease-inhibitor complexes: 1CHO (OMTKY3 wild type, P_1 = Leu), 1HJA (P_1 = Lys mutant), and 1MTN (BPTI wild type, P_1 = Lys).[30,51,52] In contrast, no structural data for porcine pancreatic elastase bound to OMTKY3 were available; hence, we relied on complexes with the human inhibitor elafin (1FLE, P_1 = Ala) and the chymotrypsin/elastase inhibitor (C/E-1) from *Ascaris* (1EAI, P_1 = Leu).[37,53] Inhibition of trypsin by BPTI is a classic example of serine protease inhibition, and the mode of binding was reported early on (2PTC).[54] In each of the protein-ligand systems, the macromolecular scaffold was reduced to an Ace-P_2-P_1-P_1'-NMe peptide for scoring purposes. The calculations for growth of all 24 P_1 residues were conducted as described in steps 3 and 4 of the protocol above, with *no minimization of scaffold position.* The consequences of the assumption that a given scaffold position is relevant for all substituents is discussed below.

RESULTS

Overview

In this section we first explore the capability of our sampling and scoring procedures to reproduce experimental results and then turn to the library design issues of side chain and scaffold selection. As a critical test, we investigated the ability of our docking tools to adequately sample and score amino acid side chain conformations for the trypsin-like serine protease family. For this evaluation, the experimental binding data accumulated carefully by Lu et al.[28] and Krowarsch et al.[31] were invaluable. These tests were generally successful (see below). We then expanded the side chain-selection procedure to three sites of substitution, generating capped tripeptides. Most of these calculations use the 24 side chains described above, but we also briefly explored a set of 100 side chains. We found that appropriate side chains are selected for the S_1 site. Our results for the other subsite preferences are consistent with the limited data available. Finally, we designed libraries using non-peptide scaffolds and the same set of side chains, and compared these libraries to our peptide library results.

Mutations at P_1 in the Context of a Truncated Macromolecular Inhibitor Scaffold

To test the accuracy of substituent selection, we made use of exhaustive P_1 mutational data for macromolecular inhibitors of serine proteases. We focus on binding constants determined for BPTI with bovine chymotrypsin and trypsin[31] and for OMTKY3 with both chymotrypsin and porcine pancreatic elastase.[28] Using the crystal structure geometries of the peptide inhibitor in these complexes, we added the substituents corresponding to 24 P_1 mutants, and for each one used solvation scoring to predict the differential free energy of binding relative to glycine. The investigation of three proteases allows us to evaluate a common scaffold and a single set of side chains in diverse receptor environments. Comparisons to the observed values for each protease complex are given in Table I. Looking first at the overall results, we see very good agreement for the aliphatic hydrocarbons, underestimation of the interaction for aromatic side chains, and modest overestimation for polar, uncharged side chains. We examine these findings in detail below.

Variation with increasing aliphatic side chain length of binding affinity

The scoring of straight-chain aliphatic side chains in the hydrophobic S_1 pocket of chymotrypsin provides a pure test of our treatment of solvation—uncomplicated by issues such as partial atomic charges or hydrogen bond geometry. Figure 6 shows excellent performance in these cases for the OMTKY3 scaffold. The contribution to binding from each additional methylene group averages 1.3 kcal/mol, observed, and 1.1 kcal/mol, calculated. The maximal binding affinity contributed per methylene unit in well-packed complexes has been previously estimated as 1.5 kcal/mol,[55] making this system close to ideal. Predicted results for aliphatic amino acids presented from a BPTI-derived scaffold to chymotrypsin parallel those from the OMTKY3 scaffold (Table I), with the contribution per methylene group averaging 1.3 kcal/mol.

Similar results were generated for trypsin, although the data were only linear through norleucine. The average

TABLE I. Comparison of Experimental and Computed Relative Free Energies (kcal/mol) of Binding for P_1 Mutants

P_1 residue[a] heavy atoms		Chymotrypsin					Elastase			Trypsin	
		OMTKY3			BPTI		OMTKY3	C/E-1	Elafin	BPTI	
		$\Delta\Delta G$exp	$\Delta\Delta G$calc[b]	$\Delta\Delta G$calc[c]	$\Delta\Delta G$exp	$\Delta\Delta G$calc[d]	$\Delta\Delta G$exp	$\Delta\Delta G$calc[e]	$\Delta\Delta G$calc[f]	$\Delta\Delta G$exp	$\Delta\Delta G$calc[g]
0	G	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	A	−1.3	−1.8	−1.0	−1.9	−0.7	−2.2	−1.9	−1.7	−1.7	−1.6
2	2	−3.0	−3.6	−3.1		−2.8	−3.5	−3.4	5.8		−2.8
	S	−1.1	−2.8	−2.2	−0.7	−3.5	0.0	−2.4	−2.6	−1.6	−1.4
	C	−3.4	−4.0	−3.1		−3.1	−1.9	−3.0	−0.6		−3.5
3	3	−4.7	−5.1	−4.4		−4.3	−3.3	−4.9	[h]		−4.3
	V	−1.8	−2.9	−0.7	−2.0	1.2	−1.3	−3.0	[h]	−0.6	14.7
	hS	−3.2	−4.6	−5.2		−6.0	−1.2	−5.1	0.0		−4.1
	T	−1.6	−3.7	−3.2	−2.0	−2.8	−2.0	−3.8	4.3	−1.7	−2.4
4	4	−5.5	−5.5	−5.6		−5.5	−3.1	−5.5	10.0		−5.6
	L	−6.0	−5.7	−3.9	−5.7	−2.0	−2.2	−5.7	[h]	−3.4	−4.6
	I	−1.6	−3.3	−0.1	−1.2	1.1	−1.1	−4.2	[h]	−1.2	[h]
	M	−5.6	−5.7	−5.8	−5.5	−6.3	−1.6	−4.7	3.0	−4.6	−6.7
	N	−2.7	−4.8	−4.1	−2.8	−4.5	1.5	−6.0	2.7	−4.3	−5.0
	D	1.1	[h]	[h]	1.1	[h]	5.5	[h]	[h]	−0.8	[h]
	D°	1.1	−0.6	1.0	1.1	0.0	5.5	1.1	5.2	−0.8	1.3
5	5	−6.3	−5.7	−6.4		−5.8	0.8	−3.7	[h]		−5.5
	K	−1.6	5.9	6.0	−4.1	6.4	5.8	12.1	[h]	−12.0	7.5
	K°	−1.6	−2.13	−3.4	−4.1	−2.9	5.8	−1.8	[h]	−12.0	3.03
	Q	−3.1	−4.1	−3.8	−3.8	−3.7	1.8	−5.3	11.2	−2.9	−5.0
	E	0.6	[h]	[h]	−0.2	[h]	5.4	[h]	[h]	−2.9	[h]
	E°	0.6	0.1	1.4	−0.2	1.9	5.4	0.03	[h]	−2.9	0.13
6	H	−3.0	−7.0	−5.6	−4.0	−7.0	5.9	−2.4	[h]	−3.5	−5.6
7	F	−7.5	−2.6	−4.2	−6.1	−7.1	5.7	7.5	[h]	−5.3	7.6
	R	−2.1	0.9	−0.8	−4.7	1.0	7.1	[h]	[h]	−12.7	−11.6
	R°	−2.1	−0.9	0.8	−4.7	1.0	7.1	10.3	[h]	−12.7	2.3
8	Y	−8.2	[h]	[h]	−6.7	−1.5	6.8	[h]	[h]	−5.4	[h]
10	W	−7.7	−0.2	−8.5	−6.5	−9.5	6.1	[h]	[h]	−3.6	3.6

[a]Residues labeled numerically refer to the number of methylene units within the side chain; Abu (2), Ape (3), Ahx (4), Ahp (5). The homoserine residue is represented by hS. A superscript ° denotes a neutralized residue.
[b]Crystal structure 1CHO (OMTKY3 wild-type, P_1 = L).
[c]1HJA (OMTKY3 mutant, P_1 = K).
[d]1MTN (BPTI, P_1 = K).
[e]1EAI (*Ascaris* chymotrypsin/elastase-1 inhibitor, P_1 = L).
[f]1FLE (Elafin, P_1 = A).
[g]2PTC (BPTI, P_1 = K).
[h]Site does not accommodate side chain presented by this scaffold ($\Delta\Delta G > 15$ kcal/mol).

contribution from each additional methylene group was 1.4 kcal/mol. To our knowledge, there is no complete set of binding data for straight-chain aliphatic mutants for BPTI-trypsin. However, a ΔG of −1.7 kcal/mol for the glycine to alanine mutation was observed,[31] and Beckmann et al.[56] report a ΔG of −1.1 kcal/mol for the norvaline to norleucine mutation. Our corresponding calculated values are −1.6 and −1.3 kcal/mol, respectively.

Although there has been no published crystal structure of porcine pancreatic elastase (PPE) with OMTKY3, the chymotrypsin/elastase inhibitor of the *Ascaris* fold-family has P_1 = Leu and has been crystallized with PPE.[53] In addition, the structure of elafin, a human elastase-specific inhibitor, in complex with PPE is known.[37] The results differed considerably, depending on which enzyme structure was used (Table I). As expected, we find that elastase crystallized with leucine in the S_1 pocket was able to accommodate larger side chains than does the complex with alanine in the pocket (elafin). Of course, the docking calculations do not include the cost of rearranging a small pocket to allow a larger P_1 side chain. Although agreement between predicted and experimental relative free energies of binding for porcine elastase is less quantitative, the trend in binding as side chain length increases is correct, and it is clear from the experimental data that the elastase structure is unable to accommodate side chains longer than norleucine (Fig. 6).

Scoring of additional P_1 *mutants*

The full set of $\Delta\Delta G$ results for P_1 mutants is shown in Figure 6 for each protease. Agreements are reasonable for most side chains across the three targets except for tyrosine and the charged amino acids. Trends for elastase are correct, but accuracy suffers for larger side chains that are predicted to bind too well (see below).

If we look carefully at tyrosine docking to chymotrypsin, we see that the side chain conformations of Phe and Tyr are very similar, with the phenyl ring orientation governed

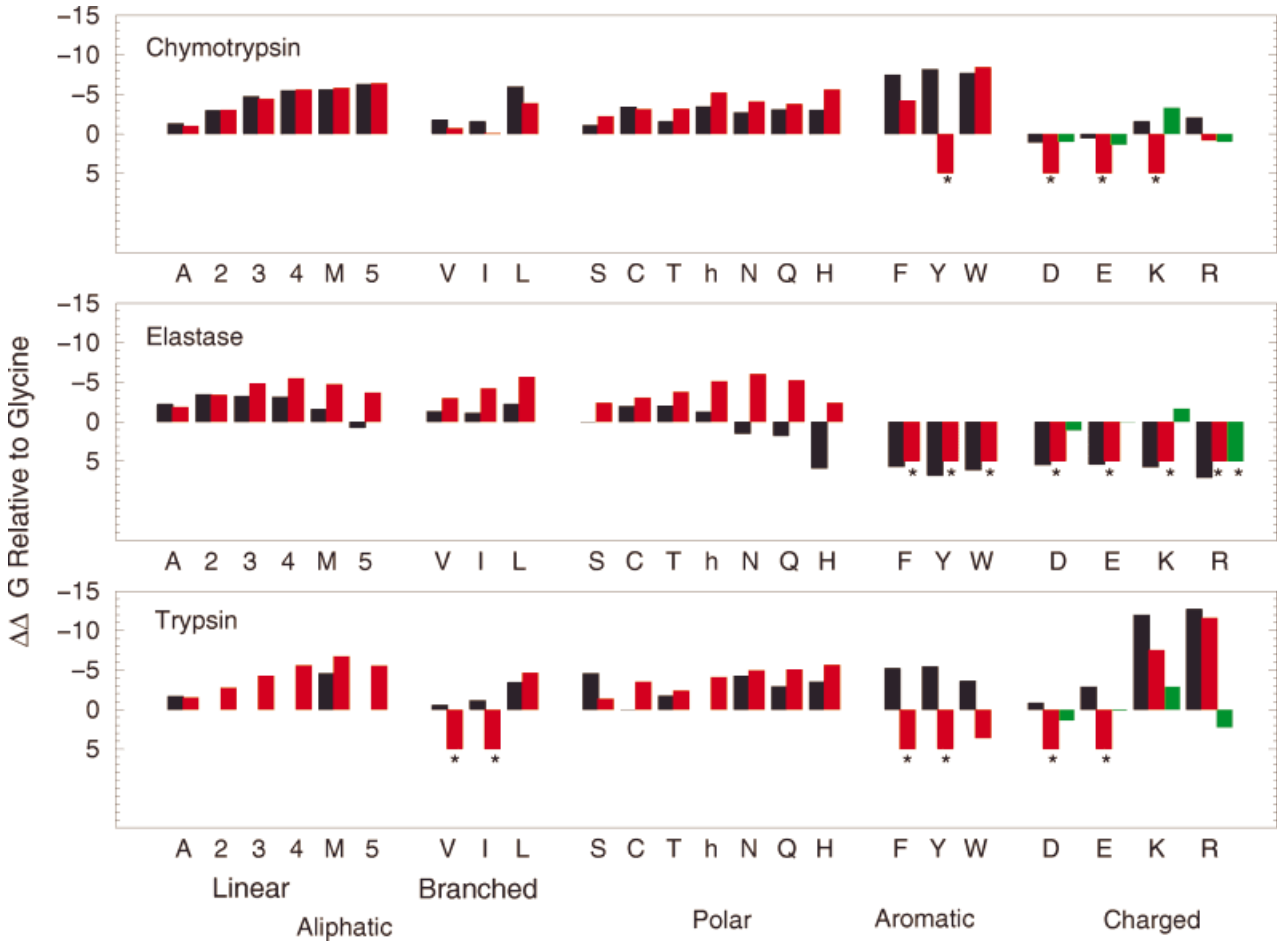

Fig. 6. Experimental (black) and computed (red) relative free energies of binding, relative to glycine, for all P_1 mutations. Results for neutralization of substituents normally charged at physiological pH are plotted in green. Anomalous results noted with asterisks have been truncated at 5 kcal/mol (see text for details).

by the position of Met 192. The tyrosine hydroxyl is too close (2.4 Å) to the backbone carbonyl of Ser 217, however, and even side chain minimization under the solvation scoring function is not sufficient to remove this clash. This is not unexpected, because the calculation has not allowed either protein flexibility or the recovery of ligand conformations by changes in scaffold position. In fact, the more flexible tripeptide docking results, described later, are able to identify tyrosine as a favored library component (see Table II below, and Supplement I). Phage display experiments have shown selection of His and Asn for chymotrypsin,[36] and these more polar residues also score well in our calculations (Table I).

For elastase, we correctly reproduced the general trend favoring small side chains. With the structure crystallized with P_1 = Leu, the calculation predicts leucine to bind well to the protease. When using a structure containing alanine as the P_1 ligand, a smaller binding site is seen and longer, branched amino acids cannot be accommodated. Again, because we do not model protein flexibility, the important point is that the calculation matches the binding of small side chains well. The calculations rank polar, uncharged side chains somewhat more favorably than do the experimental data.

Similarly, for trypsin, fixing the protein and scaffold (backbone) configurations results in poor $\Delta\Delta G$ predictions for the aromatic and branched residues; predictions for Val, Ile, Phe, and Tyr were all in error by >10 kcal/mol. However, we have partially avoided this limitation in our library design protocol by optimization of the scaffold. Another way to mediate this difficulty may be to use a scoring function that is less sensitive to small changes in geometry. As with chymotrypsin, the library docking results allow scaffold flexibility and displacement and yield reasonable rankings of these side chains.

Charged amino acids. The computed results for trypsin show good agreement for binding P_1 = Arg but show a preference for arginine over lysine greater than that observed for BPTI,[31] despite the use of a structure crystallized with lysine in the S_1 pocket. Peptides examined through phage display techniques have shown a preference for Arg as well.[36] In the crystal structure, two water molecules mediate the interaction with Asp 189 at the base of this pocket, but these waters were removed before the

TABLE II. Characteristics of the Best-Scoring Molecules From Each Scaffold Pose for the Peptide Libraries Docked Against the Serine Proteases[†]

	Molecule			Scaffold select	ΔG		Registration					Oxyanion
Library rank	R_1	R_2	R_3	rank	kcal/mol	RMSD[b]	S_3	S_2	S_1	S'_1	S'_2	C═O
Chymotrypsin												
1	W	F	W	7	−23.2	0.9		W	F	W		yes
2	Y	Y	W	1	−22.5	0.8			Y	Y	W	yes
3	S	F	F	15	−19.6	1.4		Ace	F	F		yes
4	W	F	W	5	−19.4	1.0			W	F	W	yes
5	W	W	M	2	−17.7	1.2		W	W	M		yes
6	S	W	W	3	−17.3	0.6	Ace	W	W			yes
7	M	W	W	19	−15.6	1.7	M	W	W			yes
8	5	5	W	10	−15.1	4.6		5	W		5	
9	W	T	W	14	−15.1	3.7			W	Ace		
10	5	Y	N	18	−14.0	6.0		Y	Ace	5		
11	Y	Q	N	17	−12.4	6.0		N	Q	Y		yes
12	T	Y	W	8	−10.9	5.4		Y	NMe			
13	H	hS	W	4	−10.7	6.0		W	NMe	hS	Ace	
14	5	W	hS	11	−8.2	3.0		W	5	NMe		
15	W	W	N	9	−8.0	3.9			Ace	W		
16	Q	Q	W	16	−8.0	6.4	W		Q			
17	C	W	W	6	−7.0	6.2	W		C			
18	W	hS	hS	20	−5.9	5.0			hS	Ace		
19	R	5	W	12	−5.7	5.2		W	NMe		5	
20	W	5	W	13	−5.1	5.2			NMe		5	
Elastase												
1	S	F	W	2	−13.6	1.3			S	F	W	yes
2	W	M	N	9	−13.2	6.3	N	M	Ace	W		
3	W	I	T	12	−12.8	2.0	W	I	T			yes
4	R	M	hS	5	−12.2	3.6	M	R	hS			yes
5	M	N	S	6	−12.0	1.3		M	N	S		yes
6	G	S	4	8	−12.0	1.2		G	S	4		yes
7	W	W	T	15	−11.6	2.3	Ace	W	T			
8	W	Y	hS	1	−10.4	2.1	W	Y		NMe		
9	F	Q	Q	3	−9.6	1.0	Ace	Q	Q	NMe		
10	W	W	I	20	−9.4	2.8	W	W		I		
11	W	R	5	11	−9.3	6.6		NMe	5			
12	5	Q	5	19	−9.0	1.6		Ace	5	Q	5	
13	4	H	W	13	−8.8	2.0		Ace	4	H	W	
14	Q	3	I	4	−8.6	1.9		Q	3	I		yes
15	Q	W	W	16	−8.1	6.8		W	Ace			
16	Q	5	H	14	−7.0	5.9		H	5	Q		
17	Y	W	H	10	−6.8	3.9	W	H				
18	5	S	R	17	−4.0	6.2						
19	W	Q	H	18	−3.3	2.3		Q		NMe	H	
20	hS	M	F	7	−1.4	5.6		F	M	NMe		
Trypsin												
1	R	hS	R	6	−18.0	1.8			R	hS	R	yes
2	R	G	R	2	−12.9	1.2			R	G	R	yes
3	W	M	R	7	−8.9	2.0			R	M	W	yes
4	A	R	D	3	−8.7	1.1		Ace	R	D		yes
5	W	Y	R	5	−8.3	0.9	W	Y	R			yes
6	F	K	T	8	−7.7	1.0		F	K	T		yes
7	F	R	hS	4	−6.2	0.7		F	R	hS		yes
8	R	W	C	15	−3.5	6.0	C	W	R			
9	T	A	R	1	−1.0	4.2			R	T		yes
10	5	Q	R	13	0.2	1.0	Ace		R			
11	Q	5	S	19	1.9	5.0		5	NMe	Q		
12	R	G	F	16	2.8	1.9			R			
13	T	L	G	20	4.4	5.2		L	K			
14	5	3	Q	10	4.4	1.9			5	3		
15	E	W	W	11	6.1	5.5		W	NMe			
16	Q	5	F	12	6.2	5.8		F	5			
17	W	5	W	9	6.4	5.5		W	5		W	
18	E	C	H	18	8.2	5.8						
19	W	Q	A	17	11.4	2.0						
20	F	hS	Y	14	12.1	5.6		Y	NMe			

[a]For abbreviations, refer to footnote, Table I.

[b]Best RMSD to one of the three canonical registrations of OMTKY3 from crystal structures (see text). The "Registration" column shows the occupation of the subsites, but no attempt has been made to represent how the sites are occupied (i.e., direction of entry, degree of burial, scaffold conformation).

docking calculation.[54] Arginine, with a longer chain, extends further into the pocket and also makes favorable contacts through its guanidinium moiety with residues that lysine cannot reach.

As noted above for all three proteases, aspartate and glutamate side chains received very unfavorable free energy scores, many kcal/mol removed from the experimental results (Table I). Smaller discrepancies were seen for lysine and arginine. It had been proposed by Lu et al.[28] that these side chains may become uncharged on binding. To explore this possibility, we calculated the free energy of protonation/deprotonation in solution, shown in Scheme 1. Steps one and two represent the binding free energies of unprotonated (A^-) and protonated (HA) ligands to an enzyme, as estimated with the DOCK free energy scoring function. The free energy of neutralization of a given ligand in solution may be calculated through Equation 4, where pH is the pH of the experiment (pH = 8.3) and pKa refers to the value for the specific amino acid side chain of interest. In this calculation, we have used the following pKa data: Asp = 3.9, Glu = 4.3, Lys = 11.1, and Arg = 12.0.[57,58]

Scheme 1.

$$\begin{array}{ccc} & 1 & \\ H^+ + A^- & \rightarrow & H^+ + A \cdot E \\ 3\downarrow & & 4\downarrow \\ & 2 & \\ HA & \rightarrow & HA \cdot E \end{array}$$

$$\Delta\Delta G \text{ (kcal/mol)} = 2.303RT|\text{pH} - \text{pKa}|. \quad (4)$$

We then calculated the binding free energy (relative to glycine) of the uncharged side chain and added the appropriate correction term. The results are in much better agreement with experiment when this decharging effect is allowed, especially for chymotrypsin where there are no charges within the S_1 site (Table I, Fig. 6). It should be noted that the calculations suggest that Arg and Lys are *charged* in the S_1 pocket of trypsin.

In conclusion, the free energy scoring function performs well for a variety of side chains and these three targets, especially because we did not optimize the position of the scaffold and did not allow the binding pocket to relax. This has been the most complete evaluation of the solvation scoring option in DOCK to date. It is clear that the choice of protease structure used for the calculations will influence the predicted inhibitor side chain rankings. In principle, the target-structure bias can be mitigated with the addition of protein flexibility. Inclusion of protein flexibility in docking calculations has been reviewed recently.[59] When considering large libraries, however, the approaches investigated thus far would add unacceptable computational expense.

Peptide Libraries

Compared with the general problem of docking libraries, the macromolecular P_1 calculations were easy to evaluate. A single scaffold configuration was fixed by the experimentally determined structure of the complex, at least one P_1 side chain conformation was known, and there were experimental binding data directly comparable with the calculations. None of these constraints is available to evaluate the library designs that we now consider.

First, a small tripeptide can productively dock in many configurations that do not match the "canonical" geometry.[27,60] Even when the scaffold configuration presents side chains into canonical enzyme subsites, the registration of the presentation can vary (Fig. 4). Second, we do not have experimental geometric information about side chain conformations to compare with the simulations, although it is reasonable to expect the best binders from the macromolecular scaffolds to appear in plausible conformations that explore the S_1 subsite. Third, we have no binding data for the proposed libraries. Again, we anticipate that the side chains in S_1 will correspond to known preferences. Thus, we must use surrogate markers to evaluate the success of our filtering protocols and the plausibility of the designed libraries.

Scaffold docking and Scaffold Select

We first examine the scaffold configurations. We track three criteria for plausible scaffold presentations: backbone conformation; subsite registration (configuration, pose); and orientation for mechanism-based inhibition. We take the OMTKY3-chymotrypsin crystal structure[51] as our standard for the "canonical" peptide conformation and alignment and compute the RMSD for our scaffold configurations to the heavy atoms of the three P_1-containing canonical registrations (P_3-P_2-P_1; P_2-P_1-P'_1; P_1-P'_1-P'_2). The smallest of the three RMSD values is then reported as the RMSD of a given configuration.

The range of scaffold geometries was fixed by the initial scaffold conformation search. The initial peptide (trialanine) conformations were generated by exhaustive conformational analysis, ranked by internal energy, and the best 1,500 were retained for docking. This set contained 57 canonical conformations—conformations with five of the six φ and Ψ angles within 30° of the OMTKY3 crystal structure. All three canonical registrations were represented in the set of 1,500. The conformations were not reproduced exactly because of sampling at finite (60°) increments. The torsion angle deviations ranged from 0.2 to 23°.

The scaffold conformations were individually docked to each of the three targets by using energy (force field) scoring. We retained 2,000 configurations for each target. For chymotrypsin, this set contained 201 canonical conformations, a threefold enrichment; 168 of the configurations (8.4%) were within 1.5 Å RMSD of one of the three canonical alignments within the enzyme target site.

To evaluate methods for selecting docked scaffold configurations, we compared the Scaffold Select (vector) score to the initial DOCK energy ranking, along with a subsequent free energy score ranking. An enrichment plot for chymotrypsin is shown in Figure 7; the best Scaffold Select peptide pose is contrasted to the best DOCK-scored pose with trypsin in Figure 8. The Scaffold Select filter produced a higher enrichment of structures from the set of

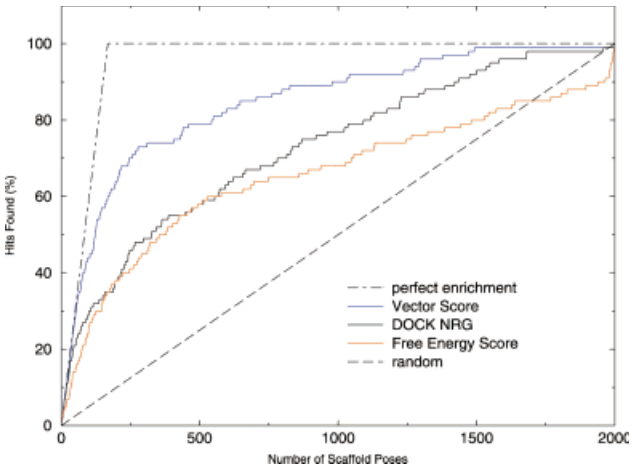

Fig. 7. Enrichment of low RMSD peptide configurations from the set of 2,000 docked into chymotrypsin by the vector scoring method compared with force field and free energy scoring methods.

168 "good" configurations above, compared with the more traditional DOCK energy and free energy scoring functions. Of the scoring methods, Scaffold Select yielded the lowest RMSD for its best-ranked configuration. Furthermore, the pose with the smallest RMSD was ranked #41 by DOCK and #74 by free energy scoring, but was ranked #8 by vector scoring. DOCK and free energy scoring favor scaffold configurations that fill the S_1 pocket, but incorporation of a measure of "opportunity for side chain placement" using Scaffold Select results in more substrate/inhibitor-like backbone configurations. When the Scaffold Select filtering and clustering stage was completed, we had reduced the accepted poses to a conservative level of 20 for each target. Of these poses for chymotrypsin, six (30%) had canonical alignments to the site (RMSD $<$ 1.5 Å), and six had canonical backbone conformations. This represents roughly a twofold enrichment of good configurations and a threefold enrichment in conformations compared with the original set of 2,000. A similar trend was observed for the peptide scaffold in trypsin and elastase.

Two additional criteria allow us to evaluate a docked geometry for mechanism-based protease inhibition. For these serine proteases, we have chosen oxyanion hole occupancy and positioning of the scaffold probe atom (C_β) for growth of side chains into the S_1 pocket. Of the 20 peptide poses retained for each target, 40% occupy the oxyanion hole and 75% place C_β into the S_1 pocket (see Table II below).

Overall, for the peptide scaffold, we found that a combination of traditional docking methods, along with a novel "look-ahead" method based on surface normal vectors, adequately enriched the number of promising scaffold conformations and orientations from a set of many dissimilar configurations that score equally well. This allowed us to proceed to the substituent selection phase of our protocol with the peptide scaffold and gave us a basis for later evaluation of non-peptide scaffold configurations.

Side chain selection for the peptide scaffold

At this stage, side chains were added and evaluated individually at each position, whereas the other diversity sites retained probe methyl groups. The goal was to select a small set of side chains at each position to carry forward to a detailed study of full molecules. This would result in a significant increase in efficiency, because the number of full molecules built is a combinatorial function of the number of side chains. In the previous section we showed that we could select appropriate side chains when using scaffold configurations taken from crystal structures. Using docked configurations produces similar substituent rankings for S_1, as is apparent in the fully built molecule libraries described in detail in the next section.

At the outset, we decided to retain five side chains at each position on each scaffold configuration. To validate our threshold, an exhaustive analysis for *one peptide pose*

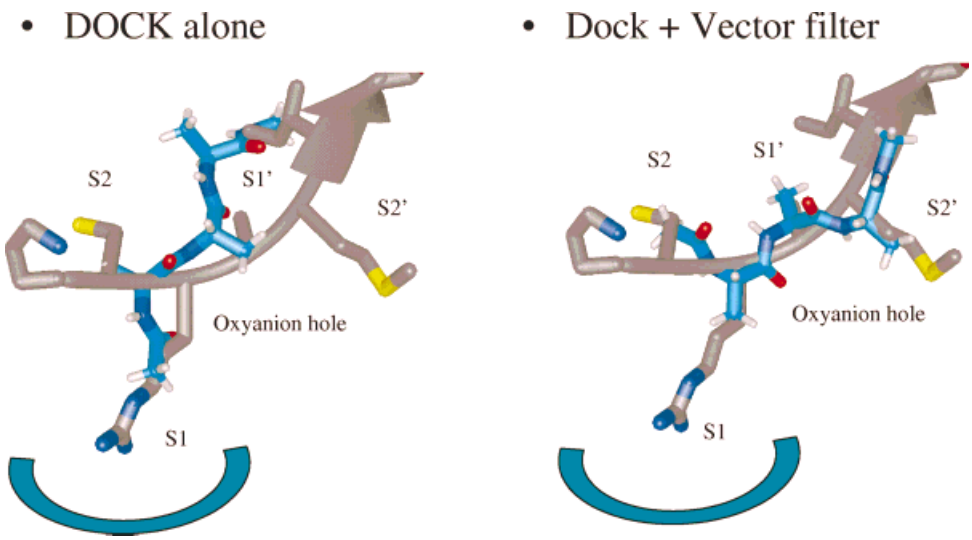

Fig. 8. Best DOCK-scored pose versus best vector-scored pose for PPT complexed with trypsin.

TABLE III. Effect on the Free Energy Scores of the Number of Substituents Retained for Subsets of a Single-Orientation (Pose) 12 × 12 × 12 Peptide Library Evaluated Against Chymotrypsin

No. of substituents (N) retained	No. of molecules (N^3) in library	Best free energy score, kcal/mol	Mean of top 20 free energy scores, kcal/mol
1	1	−16.0	−16.0
2	8	−17.5	−15.8
3	27	−18.2	−15.9
4	64	−18.8	−16.5
5	125	−19.0	−16.9
6	216	−19.0	−17.0
7	343	−19.0	−17.2
8–12	512–1728	−19.1	−17.5

was performed by using a set of 12 representative side chains. By using the best-ranked Scaffold Select pose for chymotrypsin, side chain configurations were explored as above, ranked, and all side chains retained, resulting in 1,728 fully built molecules. Finding the best scoring molecule required that the best eight side chains be retained at each position (Table III). Even so, retaining only five substituents captured most of the best molecules in the 8 × 8 × 8 sublibrary.

To show the potential to screen larger libraries, a set of 100 substituents was attached to the best peptide pose at the site corresponding to P_1. This larger set was evaluated in the same manner as was the set of 24 side chains, and for each target, the top scoring molecules were compared. For elastase, none of the new side chains scored as well as the top scorers from the original set, primarily because the new set consisted of larger side chains. The best-scoring substituents for chymotrypsin and trypsin are shown in Figure 9. This list contains P_1 side chains from natural substrates—arginine for trypsin and phenylalanine for chymotrypsin. In addition, it also contains substructures from known inhibitors, such as a benzamidine isomer for trypsin[54] and quinoline for chymotrypsin.[61]

In summary, we have shown that the free energy scoring function allows reasonable substituent rankings when conformations are sampled by DOCK. Retention of five side chains at this stage of the protocol is sufficient to generate the top-scoring final molecules for a given scaffold pose, and the method selects appropriate substituents for the protease S_1 pocket when applied to larger reagent sets as well.

Full molecule results

The detailed results for the best-scoring peptide from each pose sublibrary are given in Table II for each protease. We note that the best scoring *poses* typically used one of the canonical registrations and make use of the oxyanion hole. The residue in the S_1 pocket, for these poses, was an aromatic side chain for chymotrypsin or a positively charged side chain for trypsin. (The 20 best-scoring peptides for each target can be found in Supplement I.) The best scoring *molecules* came from five or fewer poses for each protease. The S_1 pocket was always filled with the expected types of side chains.

In addition, we have explored the average backbone RMSDs (relative to the OMTKY3-crystal structure, as described above) for the peptide libraries after full molecule minimization and scoring is done. For trypsin and chymotrypsin, the average RMSDs (±SD) were 1.5 ± 0.9 Å and 1.0 ± 0.3 Å, respectively. In contrast, the value for peptides docked to elastase was 3.0 ± 2.1 Å. The larger deviation for elastase reflects peptide orientations for which large substituents make general contacts within the active site rather than interacting with the small S_1 pocket in a canonical manner.

The most common feature of the best sequences found for chymotrypsin was an aromatic side chain interacting with the S_1 pocket (Supplement I), but this corresponded to different sequence positions in the various configurations (Table II). The best-scoring peptide was Ace-Trp-Phe-Trp-NMe, which was presented in a substrate-like fashion with the central phenylalanine in S_1 and had a free energy score of −23.2 kcal/mol. This was only slightly better than the S_1-S_1'-S_2' registration, for which the sequence Ace-Tyr-Tyr-Trp-NMe had the best free energy score, −22.5 kcal/mol. The third registration, S_3-S_2-S_1, had a much poorer predicted energy, either −17.3 or −15.6 kcal/mol, depending on whether the acetyl cap or the substituent was directed into S_3.

The free energy scores for molecules binding to elastase were less favorable than those for chymotrypsin, increasing from −13.6 kcal/mol for the best-scoring peptide, Ace-Ser-Phe-Trp-NMe. In this case, the registration was S_1-S_1'-S_2', which placed the serine side chain into the S_1 pocket and extended the peptide into the prime side of the active site. The other registrations, S_2-S_1-S_1' and S_3-S_2-S_1, had scores of −12.0 and −13.2 kcal/mol, respectively. The most common sequence features were a small, polar group in S_1 and a large aromatic group in S_2'.

For trypsin, the two best-scoring poses were very similar, placing R_1 = Arg in S_1, small groups in S_1', and R_3 = Arg in S_2'. The best-scoring molecule was Ace-Arg-Hse-Arg-NMe, with ΔG = − 18.0 kcal/mol. The third-ranked pose library presented the R_3 group to S_1 with the peptide backbone running through the site in the opposite direction of a canonical inhibitor. Two other major registrations explored, S_2-S_1-S_1' and S_3-S_2-S_1, had poorer scores, −7.7 and −8.3 kcal/mol, respectively.

Figure 10 shows the averages and ranges of scores for the 5 × 5 × 5 set of sequences for each scaffold configuration for each target. The general trend is a greater variance in scores among scaffold configurations than

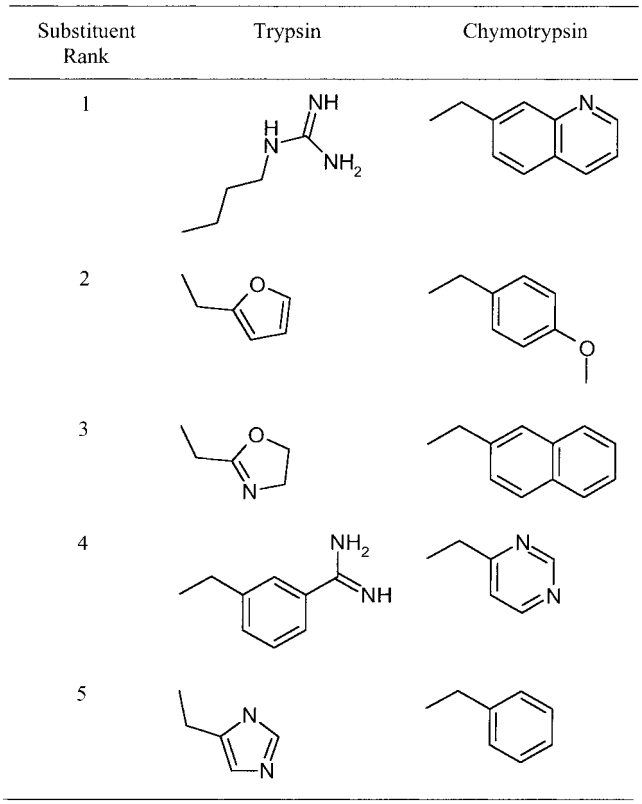

Fig. 9. Top-ranked substituents from a set of 100 evaluated within the S_1 pockets of trypsin and chymotrypsin.

among sequences for a single pose. In other words, *a good scaffold configuration can score well with many possible sequences, but no sequence can rescue a poor scaffold configuration.* All sequences enumerated for the best configuration against chymotrypsin scored better than any sequence for the 14th best pose (e.g., Fig. 10a). The range of scores was largest for trypsin. The sublibraries containing the molecules with the best scores, dominated by Arg and Lys contributions, also have molecules with unfavorable free energy scores (Fig. 10c). Clearly, the inclusion of three suboptimal substituents directed into S_1 greatly spreads the scores.

An important procedural question presents itself at this stage—can we determine the number of scaffold poses we need to retain for future applications of this protocol? Twenty poses for each target was a conservative choice. As one measure of scaffold pose success, the lowest backbone RMSD (from one of the three canonical OMTKY3 registrations) is noted for each peptide pose in Table II. Within the table, each library has been ranked by the free energy score of its best final molecule. Scaffold configurations with RMSD $<$ 2.0 Å have clearly developed into the better scoring libraries for all three target proteins. Thus, for cases in which an experimental binding configuration is known, applying an RMSD filter could significantly reduce the number of poses retained. We also considered the effect of retaining $<$20 peptide poses on the PPT library results for chymotrypsin. We have compared free energy score

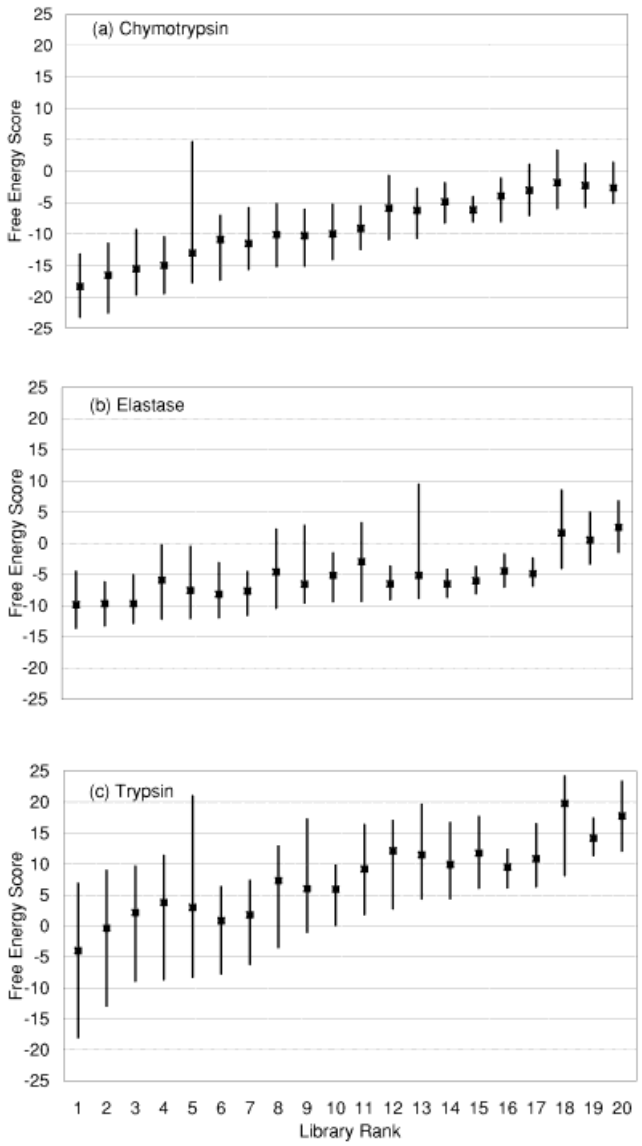

Fig. 10. High, low, and average free energy scores for each PPT pose sublibrary versus (a) chymotrypsin, (b) elastase, and (c) trypsin (filled square = average score).

histograms by using only the best 5, 10, 15, or 20 Scaffold Select poses in Figure 11. The distribution for 15 poses matches the profile found by using all 20 poses for the most negative free energy scores. Here, too, it is clear that filtering to 10 scaffold poses based on vector score results in no significant loss of the top scoring molecules but would yield a 50% time savings.

We can also use the data from the collection of $5 \times 5 \times 5$ sublibraries docked against each target to characterize the binding subsites for each target. If we consider the best 20 of the ($\sim$125 $\times$ 20 poses = $\sim$2,500) molecules in each set (Supplement I), we can ask: How often is each subsite occupied, and how frequently do particular classes of substituents appear in a given protease subsite? The answer is given in Figure 12, where occupancy of a site by a

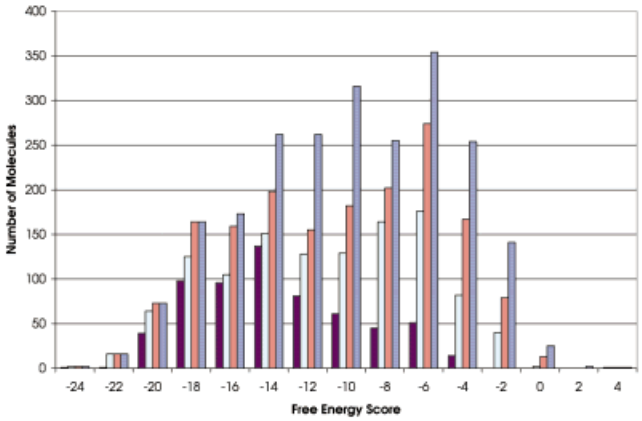

Fig. 11. Effect of the number of scaffold poses retained on the free energy histograms for the PPT library docked to chymotrypsin. First five poses (purple), first 10 poses (cyan), first 15 poses (peach), all 20 poses (blue).

substituent was determined by the scaffold registration and visual inspection of the best-scoring molecules. The S_1 site of chymotrypsin is fully occupied by aromatic residues (including one occurrence of histidine). Elastase shows Arg appearing in S_2 or S'_1 on occasion and excludes aromatic residues from S_1, as we would expect from the small pocket. Aromatic side chains do appear to contribute to binding affinity at the S_3, S'_1, and S'_2 sites, however. Trypsin clearly selects basic (Arg^{20}) side chains for its S_1 pocket but also allows charged side chains in the S'_1 and S'_2 subsites. Unlike chymotrypsin and elastase, aromatic residues do not appear the S'_1 site of trypsin.

We have further characterized substituents by size in Figure 13. For chymotrypsin, the S_2 to S'_2 pockets appear to exclude midsize side chains. In contrast, elastase is tolerant of a range of substituent sizes within its pockets with the exception of the S_1 pocket, which accepts only small side chains, as expected. As seen above with aromatic versus non-aromatic side chains, the S'_1 pockets of both chymotrypsin and elastase are capable of accommodating substituents of six or more heavy atoms. However, for trypsin, the S'_1 pocket favors three or fewer heavy atoms. The ability to characterize substituent class preferences and pocket size based on our library results should be valuable tool for further library optimization.

Thus, analysis of the full peptide sequences shows that the side chain selection phase of the protocol selected appropriate substituents for binding into the S_1 pocket. Furthermore, the number of poses necessary to obtain the best scoring molecules may be reduced in half from the 20 poses we investigated. The substituents of the best-scoring sequences have been used to characterize the protease-binding sites, which we intend to exploit for future design.

Performance comparison

We compared our method to the traditional, one-molecule-at-a-time flexible docking of the library. We enumerated a smaller, $12 \times 12 \times 12$ library of tripeptides and separately docked each of the 1,728 molecules to chymotrypsin. To be appropriate for library design, a docking method would have to process this set in a few hours or, at most, a few days. For comparison, therefore, we allocated a similar amount of computer time as was used in the "scaffold + side chains" protocol described above (Table IV). The difference in results was dramatic. The library design protocol, using all 20 peptide poses and a *superset* of the substituents in the enumerated library, found as its best score −23.2 kcal/mol for the sequence Ace-Trp-Phe-Trp-NMe. For the individually docked peptides, the best free energy score was −13.0 kcal/mol, for the sequence Ace-Leu-His-Leu-NMe. The Ace-Trp-Phe-Trp-NMe molecule, docked individually, therefore scored worse than −13.0 kcal/mol. Hence, neither the score nor the sequence produced by individual docking was optimal. As the size of the library grows, of course, docking individual members of the library becomes even less attractive and is completely infeasible for large libraries.

Furthermore, we compared our side chain selection scheme with the library design program CombiDOCK.[1] This program differs from DOCK in that it explores the conformational space of side chains by using pregenerated rotamer libraries rather than incremental growth. This results in a sixfold speedup, but at the cost of forgoing free energy scoring during side-chain selection. When the molecules generated by CombiDOCK were rescored with the free energy function, the best scores generated by CombiDOCK were −19.2, −10.7, and −13.5 kcal/mol for chymotrypsin, elastase, and trypsin, respectively. The corresponding values generated with our method were −23.2, −13.6, and −18.0 kcal/mol. In short, the present version of CombiDOCK runs severalfold faster but generates fewer high-scoring molecules than our method.

Comparing the approach in this article with two others we have worked with, the compromises between scalability and accuracy are evident. Docking each molecule of a library individually is the least scalable method and thus is inappropriate for large-scale library design. At the other extreme, CombiDOCK is more efficient and more scalable but is less accurate. The method reported here is efficient enough to be useful for library design while retaining the necessary accuracy to select appropriate side chains.

Nonpeptide Libraries in Comparison With Peptides

The primary application of our method is to design and compare libraries. To that end, we have evaluated two additional libraries against the serine proteases. The benzodiazepine scaffold was chosen for its well-known medicinal properties and its prior use as a template for combinatorial libraries.[62] The tetrahydroisoquinolinone scaffold is similar in size but has a more crowded substituent display relative to BZP.[63] In addition, serine protease inhibition by molecules that include these frameworks has been reported.[64–67] For our computational study, each scaffold was allowed three sites of diversity, and the set of 24 substituents was used at each site (Fig. 1). Hence, the virtual libraries each consist of 13,284 molecules covering identical side chain "chemical" space.

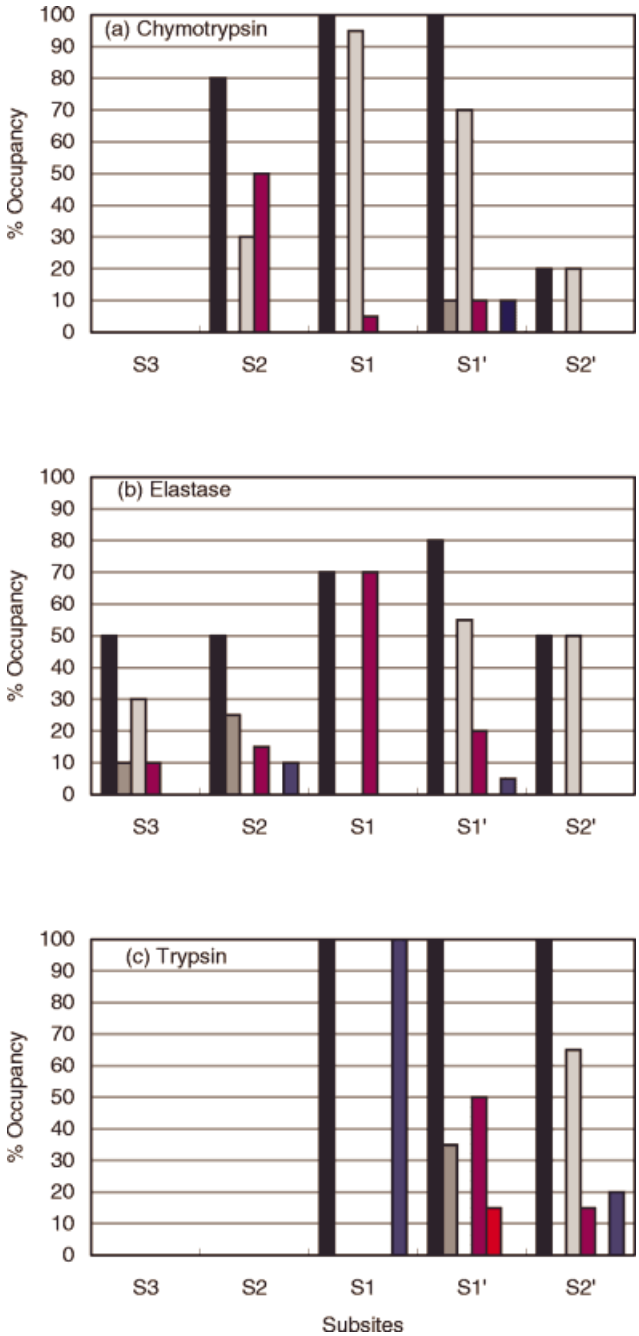

Fig. 12. Pocket characterization by substituent class for (a) chymotrypsin, (b) elastase, and (c) trypsin. Overall site occupancy (black), aliphatic (dark gray), aromatic (light gray), polar (purple), acidic (red), or basic (blue). See Figure 6 for classification of amino acids.

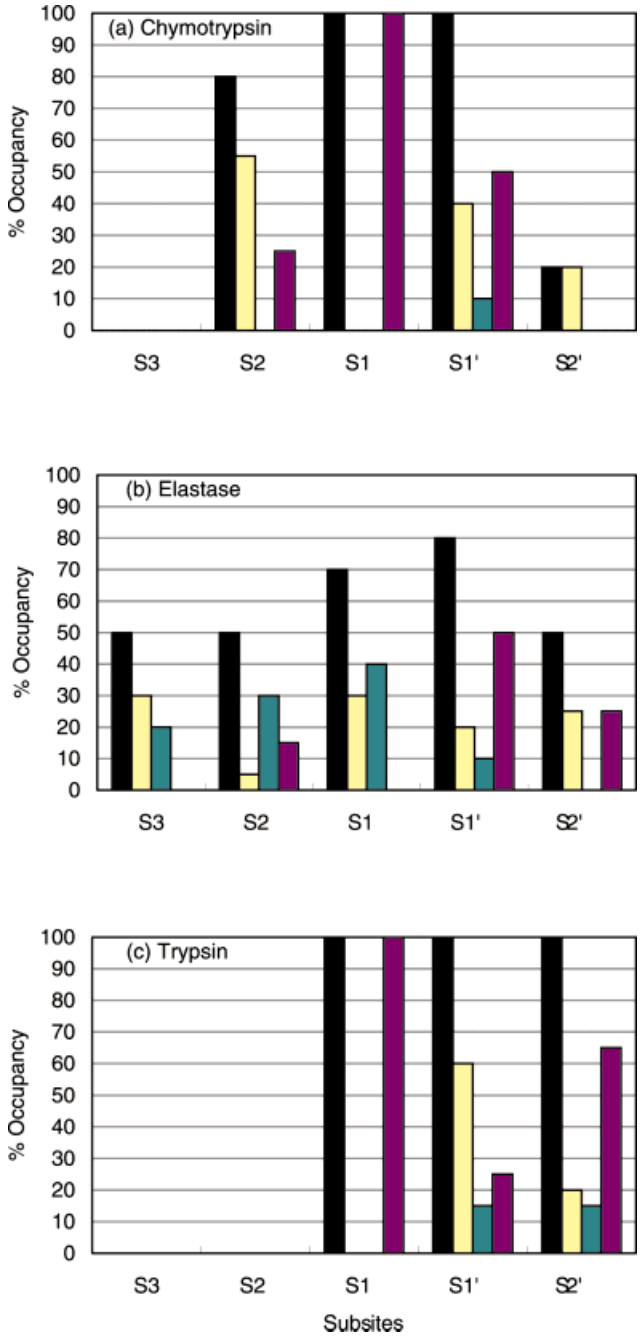

Fig. 13. Pocket characterization by substituent size, classified by the number of heavy atoms in each substituent, for (a) chymotrypsin, (b) elastase, and (c) trypsin. Overall site occupancy (black), 0–3 heavy atoms (yellow), 4–6 heavy atoms (green), 7–10 heavy atoms (purple).

Using the same scaffold docking procedure as described for the peptide libraries, we selected the 20 best, diversity-weighted THQ and BZP scaffold configurations for each of the three proteases. As with the peptides, these scaffold poses delivered diverse catalytic registry from S_3 to S_2' (Table V). In the case of trypsin, only two conformations of the docked BZP scaffold survived the Scaffold Select and RMSD clustering filter; most orientations could not place the methyl probes simultaneously into three protease pockets. Use of larger substituent probes on the benzodiazepine scaffold might result in more scaffold configurations.

A clear preference for arginine or lysine in the trypsin S_1 site is observed for both BZP and THQ scaffolds. The BZP and THQ libraries also yield large aromatic side chains for the S_1 site in chymotrypsin; however, despite our vector look-ahead method, the best-scoring libraries place a

TABLE IV. Timings for Each Phase of the Docking Protocol for Peptide Libraries and Chymotrypsin[†]

Phase	Phase	Time, s
Scaffold docking (overhead cost)[b]	Conformation generation and docking (PPT)	15,166 (4 h)
	Scaffold Select and RMSD clustering	600
Substituent selection[c] (one pose)	Enumerate independent substituent sublibraries (SYBYL)	15
	DOCK growth and force field scoring	497
	Free energy rescoring	8,123
Full molecules (one pose)	Enumerate $5 \times 5 \times 5$ sub-library (SYBYL)	15
	Intramolecular clash check and minimization, DOCK force field	100
	Minimize molecules with free energy scoring	3,277
Total time per pose		12,027 (3.3 h)
Total time for 20 poses[d]		66
Total time for the PPT library[e]		71

[†]All times based on a single processor, 195 MHz, R10000 machine.
[b]The peptide scaffold is highly flexible, with 10 rotatable bonds and represents an upper limit for the time allotted to scaffold docking.
[c]Evaluation of 24 substituents at three diversity sites per scaffold.
[d]For fewer poses, the time required is decreased proportionately (see text).
[e]PPT library: evaluation based on 20 scaffold poses, including the overhead costs for scaffold docking.

portion of the scaffold into the large hydrophobic subsite. For elastase, small portions of many substituents, as well as the scaffold itself, may be found in S_1. With its small binding pocket, it is reasonable that non-specific interactions in other regions of the active site may outweigh the S_1 pocket contributions for elastase.

We asked if our approach could identify the scaffold of choice for these proteins. Given that the peptide is the "natural" scaffold, we used peptide library results as the standard for comparison with the BZP and THQ library results. The scores for the best member of each library are shown in Figure 14. Over the set of targets, each library has a distinct profile. The peptide library exhibited breadth; it scored well against all three targets. The tetrahydroisoquinolinone library exhibited a more narrow profile. It scored better than the peptide against elastase (in one configuration) but significantly worse against the other two targets. The benzodiazepine library did not score well against any of the targets.

In the analysis of the docking results against each target, we again focus on the top 20 unique sequences (molecules). For each scaffold-target pair, we analyze the best scores, the range of scores, and the average score for this set of molecules. This information quantifies how well a library performs against a given target and can be used to compare scaffolds against a single target (these metrics are given in Supplement II). Average scores for the PPT libraries are always more favorable than those for BZP or THQ libraries, although the best-scoring THQ molecule for elastase is competitive with the best-scoring peptide. THQ libraries score better on average than BZP libraries against each of the three targets as well. The range of scores is largest for trypsin, reflecting the electrostatic effects of differences in positioning of basic substituents in S_1.

As shown above, several libraries may be directly compared against one target, as an extension of traditional structure-based database screening. With our present data, comparison of the scores for one library (or molecule) across several targets is more complicated. Although the results (see Supplement II) might suggest that optimal peptides will bind better to chymotrypsin than elastase, for example, a proper calculation would require that all details of target preparation were uniform and appropriate across the protein family. Primarily, the net charge on each protease must adequately represent the constant pH conditions of the typical inhibitor-binding experiment and be distributed correctly among the ionizable residues throughout the structure.

Putting aside the difficulties of explicit score comparisons, we can combine the best molecule lists for each scaffold versus the three targets and ask: how many molecules hit two or more targets? If we examine the best 20 molecules from each scaffold, there are no duplications across the targets. Of the best 50 molecules, the peptides remain specific, whereas the Gln-Trp-Met sequence for benzodiazepine scores favorably for chymotrypsin (-9.5 kcal/mol) and elastase (-5.2 kcal/mol), and the THQ sequences of Trp-His-Phe, Trp-Tyr-Ahp, Trp-Tyr-Phe, Trp-Phe-Trp) hit both targets as well. Thus, it appears that the peptide scaffold is able to deliver side chains in a target-specific way, whereas the non-peptide scaffolds are more generic in their best-scoring molecules. It might be possible, then, to develop a serine protease family-specific agent based on these scaffolds.

Our docked library results provide some structural insight into the observed preference of the PPT scaffold for the target proteins. In the case of PPT library, the best scoring candidates tend to correlate with favorable hydrogen bond interactions between the scaffold and the catalytic residues from the oxyanion hole, which are important for recognition of the tetrahedral intermediate of the proteolytic hydrolysis reaction (Table II). However, these oxyanion-like interactions are often missing in the case of best scoring candidates from THQ and BZP libraries (Table V). The structural rigidity of THQ and BZP scaffolds locks the potential acceptor (carbonyl oxygen) into a fixed distance from the substituent attachment sites. This constraint forces a trade-off between optimizing the side chains into the catalytic pockets and maintaining the oxyanion pocket interactions. Our results show that the THQ scaffold does better than BZP in reaching the oxyanion

TABLE V. Characteristics of the Best-Scoring Molecules from the Top 5 Poses for the Non-Peptide Libraries[a]

	Molecule			Scaffold	ΔG,	Registration					Oxyanion
Library rank	R_1	R_2	R_3	select rank	kcal/mol	S_3	S_2	S_1	S_1'	S_2'	C═O
CHYM-THQ											
1	Q	S	F	1	−17.1			Z/Q	F		
2	5	Y	W	13	−15.6		Y	Z	W		
3	H	N	F	14	−14.0	Z	F	H			
4	W	S	Y	19	−13.7	Z	W	Y			
5	M	N	W	6	−13.7	Z		M/N	W		
CHYM-BZP											
1	Y	W	G	6	−14.6		W	Z			yes
2	3	N	W	9	−11.8		Z	N	3		
3	W	S	5	4	−11.3		W	5			
4	W	Y	N	19	−11.0		W	Y			
5	W	Y	hS	1	−10.2		Y	Z	hS		yes
PPE-THQ											
1	W	Y	W	4	−13.7	W	Y	Z	W		
2	F	S	Q	5	−10.9			Q	S	F	
3	S	Y	W	6	−9.9		Y	S	Z	W	yes
4	R	S	T	7	−9.5		S	T			
5	W	S	I	16	−9.4			I			
PPE-BZP											
1	N	W	H	7	−7.1		H	N			
2	Y	W	W	14	−6.7	W	Z	W			
3	Q	W	M	17	−5.2	M	Z	Q			
4	W	H	T	16	−4.2	W	T	H			
5	5	5	R	13	−4.2		Z	5			
TRYP-THQ											
1	R	F	Q	2	−7.8		F	R			yes
2	Y	N	K	3	−4.5	Y	N	K			
3	Y	D	R	5	−3.9			R	D		
4	Q	W	R	1	1.8			R	Q		
5	T	H	5	15	2.6	5	H		T		
TRYP-BZP[b]											
1	T	R	H	1	−2.5		T	R			yes
2	H	T	S	2	10.7	H					

[a]Only best five poses are shown. The notation Z indicates that a portion of the scaffold, rather than a substituent, was observed within the subsite. Other abbreviations are as noted in Table I.
[b]Only two poses resulted from the Scaffold Select and RMSD clustering phase of the docking protocol.

site while simultaneously presenting substituents to the pockets well.

Finally, we return to the question of library design and summarize our results for side-chain selection for each scaffold. Starting with the peptide library and trypsin, there is one dominant pose for the 20 best-scoring sequences that is in a canonical registration placing Arg at S_1 (Table VIa). For chymotrypsin (Table VIb), there are two canonical poses represented in the 20 best sequences. The major one places Tyr or Phe at R_2; the other has Tyr or Phe at R_1. A library design that distinguishes between these two registrations is indicated in the table.

Elastase presents a more complicated challenge. The top 20 sequences are divided among four poses. The second-ranked pose library does not interact with the oxyanion hole and places the acetyl termination in S_1. A separate library that includes variation of the capping group should be developed to test this pose. Otherwise, the primary pose places R_1 into S_1, whereas the next best configurations place R_3 into S_1. A composite design to explore these alternatives is given in Table VIc.

For the benzodiazepine scaffold, five poses supply the top 20 molecules for chymotrypsin and elastase (75% of the molecules come from two poses each); only one pose is dominant for trypsin, placing R_2 into S_1. The THQ scaffold makes use of four poses for chymotrypsin and three for trypsin, but all of the best molecules for elastase are derived from one pose that places the scaffold in S_1. The orientation of the best-scoring molecule (Trp-Tyr-Trp) from this pose is illustrated in Figure 15. Composite designs for these scaffolds are also presented in Table VI.

In summary, we have shown that our approach to library design is able to select the preferred side chains and identify the scaffold of choice across a protein family. Our approach may be easily adapted to other protein target families and libraries.

DISCUSSION

Historically, structure-based drug design efforts were concerned with retrieval of compounds from an existing chemical database (database mining), lead optimization and de novo design. The criteria for success were finding

TABLE VIa. Composite Library Designs Based on the Top 20 Molecules for Each Target: Trypsin[†]

Scaffold	R_1	R_2	R_3
PPT	R	hS G H D A	R F W Y Q
BZP	F Y W S T	R 4	F T H Y W
THQ	R (F Y)	F W Y (N D)	F W Y (K R)

VIb. Composite Library Designs Based on the Top 20 Molecules for Each Target: Chymotrypsin[†]

Scaffold	R_1	R_2	R_3
PPT	W H T (Y F)	Y F (H W)	W (Y H 5 R)
BZP	W (F H Y M 3)	W F Q (S N Y)	G (W 5 hS N)
THQ	Q 4 (5 3 H)	S (N A G W)	W (Y F R)

VIc. Composite Library Designs Based on the Top 20 Molecules for Each Target: Elastase[†]

Scaffold	R_1	R_2	R_3
PPT	N S (W F T)	Y F (Q M H)	W (F Y N hS)
BZP	F (N H Y Q 5)	W (H F Y hS 5)	W (H F N 4 T R)
THQ	W F Y H	R H F Y W	W F H Y

[†]Sidechains ordered by frequency of occurrence. Sidechains in parentheses arise from alternate scaffold orientations.

new "hits" and/or improving the affinity of existing compounds. False positives were to be expected, and false negatives were rarely detected. With the advent of combinatorial chemistry, it became feasible to design and computationally screen large virtual libraries of compounds to yield small libraries that could be synthesized easily. Because it was straightforward to include simple filters for "drug-like" properties,[68] these libraries were judged on the fraction of compounds that interacted with the target and that were active in cells.[8,9] The design strategy of "scaffold + substituents" has proven successful in several laboratories.[8–10,69] In this article, we explore multiple scaffolds, drawing on the same substituent list for purposes of easy comparison. We direct these virtual libraries against a set of related targets. Questions for discussion include: How well does the library design protocol work? What are the major areas that need to be considered for improved protocols? and What are the most useful metrics for comparing libraries?

1. How well does the protocol work? Specifically, we explore two issues: Are reasonable scaffold and side-chain choices made, and what are the computational resources needed to carry out these calculations?

We can make a direct assessment of the scoring and flexible docking features by using the macromolecular P_1 inhibitor data from laboratories of Laskowski and Otlewski. Generally, the conformations of the side chains and the free energy of binding contributions (including solvation effects) are very well described for the nonpolar side chains and slightly overestimated for the polar, uncharged side chains providing that no rearrangements of the protein active site are required. Semiquantitative data can be generated for the protonatable side chains if the correct charge state is used; thus, the charged state of the ligand can be inferred for these systems. Fujinaga et al.[70] offer an interesting alternative analysis of the binding of these inhibitors. Their conclusion is that the parameters they explored are not generalizable to additional proteases. Thus far, our solvation model parameters[2] appear to be reasonably transferable.

Turning to the designed libraries, the known preferences for substituents in the S_1 specificity pockets are retrieved. We find this result in all cases (3 scaffolds $\times$ 3 targets). The library docking protocol contains more degrees of freedom than we explored in our macromolecular inhibitor-based calculations. This freedom of motion (scaffold conformations and orientations) leads to three interesting results: alternative geometric presentations of the scaffold (registrations) are nearly isoenergetic and must be explored as part of the library substituent choices. Second, small displacements of the scaffold are sufficient to overcome some of the problems associated with a rigid protein target; tyrosines, for example, are now correctly chosen as chymotrypsin S_1 substituents. Third, a good scaffold configuration can score well with many possible sequences, but no sequence (within the substituent set explored here) can rescue a poor scaffold configuration. We make clear predictions for occupancy preferences at other subsites, although the preponderance of the interaction for trypsin and chymotrypsin seems to reside at S_1. There are currently no comparable experimental binding data for these extended subsites, although Lu and coworkers[71] have reported progress in this arena. Although the fact that the protocol selects the expected fragments at S_1 is encouraging, experimental testing of these library designs is needed to establish the accuracy of the method.[8,9] We also recognize that peptides discussed here would be substrates, not inhibitors. In sum, the use of a solvation scoring procedure, coupled to a standard docking algorithm, seems sufficient to select plausible side chains, scaffolds, and docking geometries.

The computational issues include the time required for this protocol, processor and memory concerns, and the scalability of the algorithm. Sample timings for each step of the protocol for chymotrypsin and the flexible peptide scaffold were presented in Table IV. Approximately 10% of the total time is spent in scaffold docking, 65% in spent in side chain selection, and 25% in evaluation of the best final molecules. Memory requirements are that of a normal docking run, because the fully enumerated library is never considered. The tactics that we used to improve efficiency include preliminary screening of scaffold positions as well as an initial filtering of the side chains. A variety of other positional screening tools are available, including the well-known CAVEAT program[72] and work from Stahl and Bohm.[73] Our vector method is specifically tailored to evaluate the position of a scaffold within an active site. We also find that we can expand the side-chain lists substantially without loss of overall efficiency. The present protocol scales linearly in the number of libraries and linearly in the number of targets.

2. What needs to be done to further develop this approach? An efficient protocol to examine large numbers of libraries and large numbers of targets will have to improve upon the linear dependencies discussed above. We could,

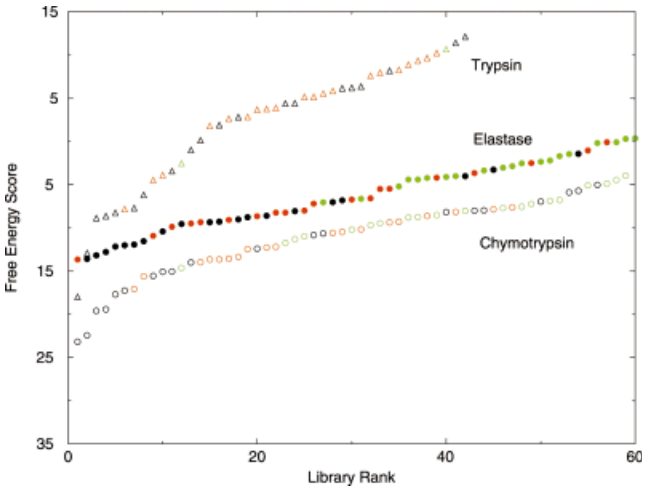

Fig. 14. Best-scoring molecules from each pose for each library versus each target. Scaffolds/libraries: PPT (black), THQ (red), and BZP (green). Targets: chymotrypsin (open circles), elastase (filled circles), and trypsin (open triangles).

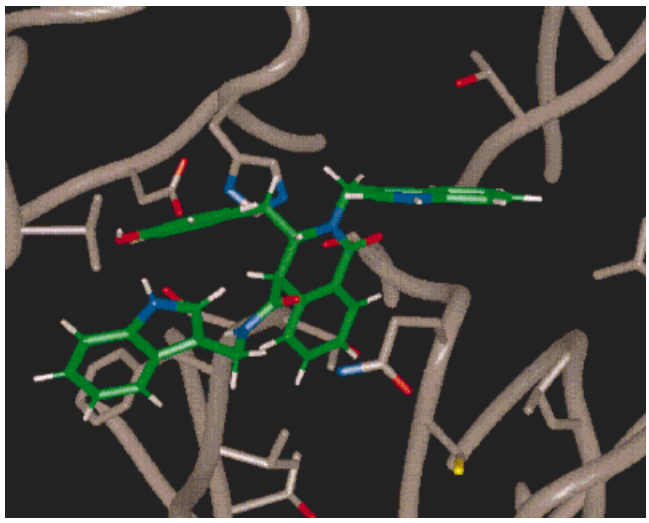

Fig. 15. Orientation of the (a) best-scoring peptide (yellow) versus elastase and (b) best-scoring THQ molecule (green) versus elastase. (Figure was created with MidasPlus.[44])

for example, align the targets based on known family relationships or active-site geometries to avoid evaluating libraries against similar targets in early stages of the discovery process. Furthermore, we note that we obtained the same rank ordering of scaffolds regardless of protease: peptide > THQ > BZP. If this observation could be generalized across a family, we could reduce testing the full scaffold × target array to, in the limit, evaluating a single column (target). Thus, we could rapidly identify the superior scaffolds for a target class and discard scaffolds that failed the initial screening. More promising scaffolds could then be analyzed in greater detail.

We have shown that the efficiency and speed of this protocol can be improved (twofold) by reducing the number of scaffold poses used for library generation. We could also improve the triage of scaffolds by using the faster CombiDOCK[1] procedure as an initial screen. A restricted set of size-ordered substituents could be used with either procedure to reduce the number of repetitive searches that end in rejection of similar side chains for a specific target.

By using the present algorithm and current generation of processors, the screening of all libraries published in 1999 against a large number of targets as described in the Introduction has been reduced by $\sim 10^4$. This factor arises primarily from the screening of side chains in a linear fashion before the combinatorial calculation of individual molecules. A similar factor needs to be found from the improvements above *and* the use of multiple processors to make the overall goal feasible.

Inclusion of Target Family-Specific Information

In the interests of developing a general method, we have avoided incorporation of much protease-specific information to direct our docking. However, for a typical library design project, one could easily restrict sampling when many details of intermolecular recognition are known in advance, with an associated potential gain in speed and accuracy of results. For example, with the PPT libraries, the peptide backbone could have been fixed to a crystal structure configuration (and possibly allowed to minimize away from the starting pose). Alternatively, an explicit covalent bond could be formed between a scaffold and the catalytic serine. A protease structure in complex with a ligand would be required for each added target and would bias scaffold selection to those that closely mimic portions of the known ligand. The side chains of Arg or Lys could also have been used as "probes" for scaffold docking, resulting in scaffold positions for trypsin that primarily direct the basic group into S_1. Furthermore, asymmetrically substituted libraries were not explored here. We could have looked at the range of substituent scores, observed that against trypsin, Arg and Lys score significantly better than other side chains at certain positions, and carried only these two substituents forward to full molecules at those positions. Of course, biasing libraries in this way virtually guarantees that no novel S_1-binding substituents will be found and would be more suited to an optimization rather than discovery phase of design.

3. Library comparison metrics. The simplest basis for comparing multiple libraries against a single target is to ask which library generated the best-scoring compound. Although this criterion is certainly useful, there is clearly much more information contained in the virtual screening computations (or experimental screening efforts). For example, we previously suggested that the limiting slope on a histogram of scores might indicate how hard it would be to find more tightly binding compounds.[8] The distribution of scores over different geometric presentations provides a basis for designing class-specific or target-specific inhibitors. A number of approximately equivalent scaffold poses can warn of a complicated SAR. The distribution of scores over a variety of scaffolds suggests whether different chemical approaches are promising. Libraries can also be compared on the basis of any easily calculable physical

property that might serve as a surrogate for "drug-like" behavior.[68] Libraries could also be graded on how selective they are for a specific target or how broadly they hit within a class. As families of structures become available, it will be useful to develop libraries that exhibit the breadth necessary, as the peptide library in this case, to permit parallel inhibitor development across a family of targets. Ongoing collaborations will provide us with sufficient multiple library, multiple target binding affinities to explore, and refine these metrics in more detail.

CONCLUSIONS

The computational time demands of a multiple-library, multiple-target endeavor are too large to use "undirected" docking procedures. Look-ahead and pruning algorithms are essential to provide efficient searching of chemical, configurational, and conformational space. We find that simple docking of a "bare" scaffold is not a good method of placing the scaffold into productive geometries. A site-generated evaluation filter, such as Scaffold Select, has proved valuable. We also have found that free energy-based scoring functions are required to evaluate a diverse population of polar, non-polar, and charged fragments. A desolvation term is of particular importance when comparing charged and uncharged moieties and should be incorporated at an early stage of screening. Finally, the comparison of libraries by the simple, structure-based score of the best-scoring molecule has provided reasonable predictions for scaffold and S_1-specificity preferences for the serine proteases. Future work will explore the general applicability of these docking methods to additional libraries across a variety of target families.

ACKNOWLEDGMENTS

We thank Kenneth A. Brameld, Jennifer L. Harris, Connie M. Oshiro, Judith Hempel, and Michael N.G. James for helpful discussions. Support for this project was provided by grants from the National Institutes of Health (GM31497 to IDK, GM56531 to P. Ortiz de Montellano, Principal Investigator, and CA 72006 to M. Shuman, Principal Investigator) and from Daiichi, Isis, and Pfizer. Tripos, Incorporated (St. Louis, MO) and MDL Information Systems, Incorporated (San Leandro, CA) each made software and databases available.

REFERENCES

1. Sun Y, Ewing TJA, Skillman AG, Kuntz ID. CombiDOCK: structure-based combinatorial docking and library design. J Comp Aided Mol Design 1998;12:597–604.
2. Zou X, Sun Y, Kuntz ID. Inclusion of solvation in ligand binding free energy calculations using the generalized-born model. J Am Chem Soc 1999;121:8033–8043.
3. Gallop MA, Barrett RW, Dower WJ, Fodor SP, Gordon EM. Applications of combinatorial technologies to drug discovery. 1. Background and peptide combinatorial libraries. J Med Chem 1994;37:1233–1251.
4. Gordon EM, Barrett RW, Dower WJ, Fodor SP, Gallop MA. Applications of combinatorial technologies to drug discovery. 2. Combinatorial organic synthesis, library screening strategies, and future directions. J Med Chem 1994;37:1385–1401.
5. Warr WA. Combinatorial chemistry and molecular diversity. An overview. J Chem Information Computer Sci 1997;37:134–140.
6. Amzel LM. Structure-based drug design. Curr Opin Biotechnol 1998;9:366–369.
7. Marrone TJ, Briggs JM, McCammon JA. Structure-based drug design: computational advances. Annu Rev Pharmacol Toxicol 1997;37:71–90.
8. Kick EK, Roe DC, Skillman AG, et al. Structure-based design and combinatorial chemistry yield low nanomolar inhibitors of cathepsin D. Chem Biol 1997;4:297–307.
9. Haque TS, Skillman AG, Lee CE, Habashita H, Gluzman TY, Ewing TJA, Goldberg DE, Kuntz ID, Ellman JA. Potent, low-molecular-weight non-peptide inhibitors of malarial aspartyl protease plasmepsin II. J Med Chem 1999;42:1428–1440.
10. Murray CW, Clark DE, Auton TR, Firth MA, Li J, Sykes RA, Waszkowycz B, Westhead DR, Young SC. PRO-SELECT: combining structure-based drug design and combinatorial chemistry for rapid lead discovery. 1. Technology. J Comp Aided Mol Design 1997;11:193–207.
11. Balkenhohl F, von dem Bussche-Hunnefeld C, Lansky A, Zechel C. Combinatorial synthesis of small organic molecules. Angew Chem Int Ed Engl 1996;35:2288–2337.
12. Walters WP, Stahl MT, Murcko MA. Virtual screening—an overview. Drug Discovery Today 1998;3:160–178.
13. Gray NS, Wodicka L, Thunnissen A-MWH, Norman TC, Kwon S, Espinoza FH, Morgan DO, Barnes G, LeClerc S, Meijer L, et al. Exploiting chemical libraries, structure, and genomics in the search for kinase inhibitors. Science 1998;281:533–538.
14. Carroll CD, Patel H, Johnson TO, Guo T, Orlowski M, He Z-M, Cavallaro CL, Guo J, Oksman A, Gluzman IY, et al. Identification of potent inhibitors of *Plasmodium Falciparum* plasmepsin II from an encoded statine combinatorial library. Bioorg Med Chem Lett 1998;8:2315–2320.
15. Szardenings AK, Harris D, Lam S, Shi L, Tien D, Wang Y, Patel DV, Navre M, Campbell, DA. Rational design and combinatorial evaluation of enzyme inhibitor scaffolds: identification of novel inhibitors of matrix metalloproteinases. J Med Chem 1998;41:2194–2200.
16. Frye SV. Structure-activity relationship homology (SARAH): a conceptual framework for drug discovery in the genomic era. Chem Biol 1999;6:R3–R7.
17. Kauvar LM, Higgins DL, Villar HO, Sportsman JR, Engqvist-Goldstein A, Bukar R, Bauer KE, Dilley H, Rocke DM. Predicting ligand binding to proteins by affinity fingerprinting. Chem Biol 1995;2:107–118.
18. Kauvar LM, Villar HO, Sportsman JR, Higgins DL, Schmidt DE Jr. Protein affinity map of chemical space. J Chromatogr B 1998;715:93–102.
19. Briem H, Kuntz ID. Molecular similarity based on DOCK-generated fingerprints. J Med Chem 1996;39:3401–3408.
20. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne, PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
21. Sali A. 100,000 protein structures for the biologist. NSB 1998;5:1029–1032.
22. Charifson PS, Corkery JJ, Murcko MA, Walters WP. Consensus scoring: a method for obtaining improved hit rates from docking databases of three-dimensional strcutures into proteins. J Med Chem 1999;42:5100–5109.
23. Still WC, Tempczyk A, Hawley RC, Hendrickson T. Semianalytical treatment of solvation for molecular mechanics and dynamics. J Am Chem Soc 1990;112:6127–6129.
24. Whittaker M. Discovery of protease inhibitors using targeted libraries. Curr Opin Chem Biol 1998;2:386–396.
25. Dolle RE. Comprehensive survey of combinatorial library synthesis: 1999. J Comb Chem 2000;2:383–433.
26. Babine RE, Bender SC. Molecular recognition of protein-ligand complexes: applications to drug design. Chem Rev 1997;97:1359–1472.
27. Schechter I, Berger A. On the size of the active site in proteases. I. Papain. Biochem Biophys Res Commun 1967;27:157–162.
28. Lu W, Apostol I, Qasim MA, Warne N, Wynn R, Zhang WL, Anderson S, Chiang YW, Ogin E, Rothberg I, et al. Binding of amino acid side-chains to S_1 cavities of serine proteinases. J Mol Biol 1997;266:441–461.
29. Qasim MA, Ganz PJ, Saunders CW, Bateman KS, James MNG, Laskowski M Jr. Interscaffolding additivity: association of P_1 variants of eglin c and of Turkey ovomucoid third domain with serine proteinases. Biochemistry 1997;36:1598–1607.

30. Qasim MA, Lu SM, Ding J, Bateman KS, James MNG, Anderson S, Song J, Markley JL, Ganz PJ, Saunders CW, et al. Thermodynamic criterion for the conformation of P_1 residues of substrates and of inhibitors in complexes with serine proteinases. Biochemistry 1999;38:7142–7150.
31. Krowarsch D, Dadlez M, Buczek O, Krokoszynska I, Smalas AO, Otlewski J. Interscaffolding additivity: binding of P_1 variants of bovine pancreatic trypsin inhibitor to four serine proteases. J Mol Biol 1999;289:175–186.
32. Ewing TJA, Kuntz ID. Critical evaluation of search algorithms for automated molecular docking and database screening. J Comput Chem 1997;18:1175–1189.
33. Ewing TJA. Automated molecular docking: Development and evaluation & new search methods. [Ph.D. dissertation]: University of California, San Francisco; 1998.
34. SYBYL. 6.5. 1699 South Hanley Road, St. Louis, Missouri, 63144, USA: Tripos Inc.
35. Daylight Chemical Information Software. Santa Fe, NM: Daylight Chemical Information Systems, Inc.; info@daylight.com.
36. Scheidig AJ, Hynes TR, Pelletier LA, Wells JA, Kossiakoff AA. Crystal structures of bovine chymotrypsin and trypsin complexed to the inhibitor domain of Alzheimer's amyloid B-protein precursor (APPI) and basic pancreatic trypsin inhibitor (BPTI): engineering of inhibitors with altered specificities. Protein Sci 1997;6:1806–1824.
37. Tsunemi M, Matsuura Y, Sakakibara S, Katsube Y. Crystal structure of an elastase-specific inhibitor elafin complexed with porcine pancreatic elastase determined at 1.9 angstroms resolution. Biochemistry 1996;35:11570–11576.
38. Lee AY, Smitka TA, Bonjouklian R, Clardy J. Atomic structure of the trypsin-A90720A complex: a unified approach to structure and function. Chem Biol 1994;1:113–117.
39. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Jr., Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. J Am Chem Soc 1995;117:5179–5197.
40. Meng EC, Shoichet BK, Kuntz ID. Automated docking with grid-based energy evaluation. J Comput Chem 1992;13:505–524.
41. Richards FM. Areas, volumes, packing, and protein structure. Annu Rev Biophys Bioeng 1977;6:151–176.
42. Connolly ML. Solvent-accessible surfaces of proteins and nucleic acids. Science 1983;221:709–713.
43. Connolly ML. Analytical molecular surface calculation. J Appl Crystallogr 1983;16:548–558.
44. Ferrin TE, Huang CC, Jarvis LE, Langridge R. The MIDAS display system. J Mol Graphics 1988;6:13–27.
45. Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Ferrin TE. Geometric approach to macromolecule-ligand interactions. J Mol Biol 1982;161:269–288.
46. DesJarlais RL, Sheridan RP, Seibel GL, Dixon S, Kuntz ID. Using shape complementarity as an initial screen in designing ligands for a receptor binding site of known three-dimensional structure. J Med Chem 1988;31:722–729.
47. DOCK Manual. 4.0. San Francisco, CA: University of California, San Francisco; 1998.
48. Pearlman RS. Rapid generation of high quality approximate 3-dimension molecular structures. Chem Des Auto News 1987;2:1.
49. CONCORD Manual. St. Louis, MO: Tripos, Inc.; 1995.
50. Schnecke V, Swanson CA, Getzoff ED, Tainer JA, Kuhn LA. Screening a peptidyl database for potential ligands to proteins with side-chain flexibility. Proteins 1998;33:74–87.
51. Fujinaga M, Sielecki AR, Read RJ, Ardelt W, Laskowski M Jr, James MNG. Crystal and molecular structures of the complex of α-chymotrypsin with its inhibitor Turkey ovomucoid third domain at 1.8 A resolution. J Mol Biol 1987;195:397–418.
52. Capasso C, Rizzi M, Menegatti E, Ascenzi P, Bolognesi M. Crystal structure of the bovine α-chymotrypsin: kunitz inhibitor complex. An example of multiple protein:protein recognition sites. J Mol Recogn 1997;10:26–35.
53. Huang K, Strynadka NCJ, Bernard VD, Peanasky RJ, James MNG. The molecular structure of the complex of ascaris chymotrypsin/elastase inhibitor with porcine elastase. Structure 1994;2:679–689.
54. Marquart M, Walter J, Deisenhofer J, Bode W, Huber R. The geometry of the reactive site and of the peptide groups in trypsin, trypsinogen and its complexes with inhibitors. Acta Crystallogr 1983;B39:480–490.
55. Kuntz ID, Chen K, Sharp KA, Kollman PA. The maximal affinity of ligands. Proc Natl Acad Sci 1999;96:9997–10002.
56. Beckmann J, Mehlich A, Schroder W, Wenzel HR, Tschesche H. Preparation of chemically mutated aprotinin homologues by semi-synthesis: P1 substitutions change inhibitory specificity. Eur J Biochem 1998;176:675–682.
57. Bundi A, Wuhtrich K. Proton NMR parameters of the common amino acid residues measured in aqueous solutions of the linear tetrapeptides H-Gly-Gly-X-L-Ala-OH. Biopolymers 1979;18:285–297.
58. Matthew JB, Gurd FRN, Garcia-Moreno B, Flanagan MA, March KL, Shire SJ. pH-dependent processes in proteins. CRC Crit Rev Biochem 1985;18:91–197.
59. Carlson HA, McCammon JA. Accomodating protein flexibility in computational drug design. Mol Pharmacol 2000;57:213–218.
60. Mattos C, Rasmussen B, Ding X, Petsko GA, Ringe D. Analogous inhibitors of elastase do not always bind analogously. Struct Biol 1994;1:55–58.
61. Wallace RA, Kurtz AN, Niemann C. Interaction of aromatic compounds with alpha-chymotrypsin. Biochemistry 1963;2:824–836.
62. Ellman JA. Design, synthesis and evaluation of small-molecule combinatorial libraries. Acc Chem Res 1996;29:132–143.
63. Griffith MC, Dooley CT, Houghten RA, Kiely JS. Solid-phase synthesis, characterization, and screening of a 43,000-compound tetrahydroisoquinoline combinatorial library. In: Chaiken IM, Janda KD, editors. Molecular diversity and combinatorial chemistry: libraries and drug discovery. Washington, D.C.: American Chemical Society; 1996. p 50–57.
64. Scarborough RM; COR Therapeutics, Inc., assignee. Preparation of selective factor Xa inhibitors containing a fused diazepinone structure. WO patent 9907730. 1999 February 18.
65. Bihovsky R, Wells GJ, Tao M; Cephalon, Inc., assignee. Benzothiazo and related heterocyclic group-containing cysteine and serine protease inhibitors. U.S. patent 5952328. 1999 September 14.
66. Spruce L, Gyorkos A; Ono Pharmaceuticals, Co. Osaka, Japan, assignee. Indole and tetrahydroisoquinoline containing alpha-keto oxadiazoles as serine protease inhibitors. WO patent 9962514. 1999 December 9.
67. Gyorkos A, Spruce LW; Cortech, Inc., Bedminster, NJ, assignee. Serine protease inhibitors-cycloheptane derivatives. U.S. patent 6015791. 2000 January 18.
68. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Delivery Rev 1997;23:3–25.
69. Graybill TL, Agrafiotis DK, Bone R, Illig CR, Jaeger EP, Locke KT, Lu T, Salvino JM, Soll RM, Spurlino JC, et al. Enhancing the drug discovery process by integration of high-throughput chemistry and structure-based drug design. In: Chaiken IM, Janda KD, editors. Molecular diversity and combinatorial chemistry: libraries and drug discovery. Washington, D.C.: American Chemical Society; 1996. p 17–27.
70. Fujinaga M, Huang K, Bateman KS, James MNG. Computational analysis of the binding of P1 variants of domain 3 of Turkey ovomucoid inhibitor to streptomyces griseus protease B. J Mol Biol 1998;284:1683–1694.
71. Lu SM, Lu W, Qasim MA, Ranjbar MR, Anderson S, Laskowski M Jr. Contributions of each contact residue in a standard mechanism canonical inhibitor. Protein Soc Symp 1999; Boston, MA.
72. Lauri G, Bartlett PA. CAVEAT: a program to facilitate the design of organic molecules. J Comp Aided Mol Design 1994;8:51–66.
73. Stahl M, Bohm H-J. Development of filter functions for protein-ligand docking. J Mol Graphics Mod 1998;16:121–132.