PROTEINS: Structure, Function, and Genetics 33:74–87 (1998)

Screening a Peptidyl Database for Potential Ligands to Proteins With Side-Chain Flexibility

Volker Schnecke,[1] Craig A. Swanson,[2] Elizabeth D. Getzoff,[3] John A. Tainer,[3] and Leslie A. Kuhn[1]*
[1]*Protein Structural Analysis and Design Laboratory, Department of Biochemistry, Michigan State University, East Lansing, Michigan*
[2]*Department of Pathology, Stanford University, School of Medicine, Stanford, California*
[3]*Department of Molecular Biology, The Scripps Research Institute, La Jolla, California*

ABSTRACT **The three key challenges addressed in our development of SPECITOPE, a tool for screening large structural databases for potential ligands to a protein, are to eliminate infeasible candidates early in the search, incorporate ligand and protein side-chain flexibility upon docking, and provide an appropriate rank for potential new ligands. The protein ligand-binding site is modeled by a shell of surface atoms and by hydrogen-bonding template points for the ligand to match, conferring specificity to the interaction. SPECITOPE combinatorially matches all hydrogen-bond donors and acceptors of the screened molecules to the template points. By eliminating molecules that cannot match distance or hydrogen-bond constraints, the transformation of potential docking candidates into the ligand-binding site and the shape and hydrophobic complementarity evaluations are only required for a small subset of the database. SPECITOPE screens 140,000 peptide fragments in about an hour and has identified and docked known inhibitors and potential new ligands to the free structures of four distinct targets: a serine protease, a DNA repair enzyme, an aspartic proteinase, and a glycosyltransferase. For all four, protein side-chain rotations were critical for successful docking, emphasizing the importance of inducible complementarity for accurately modeling ligand interactions. SPECITOPE has a range of potential applications for understanding and engineering protein recognition, from inhibitor and linker design to protein docking and macromolecular assembly. Proteins 33:74–87, 1998.** © 1998 Wiley-Liss, Inc.

Key words: docking; distance geometry; drug design; peptidyl inhibitors; protein-peptide interactions; inducible complementarity; aspartic proteinase; glycosyltransferase; serine protease; DNA repair enzyme

INTRODUCTION

Many protein recognition processes involve the binding of peptides and other small ligands, and peptides act as in vivo inhibitors and agonists of proteins as diverse as serine proteases[1] and hormone receptors.[2,3] Computational approaches to understanding and predicting such interactions are therefore of significant interest. With the increase in computational power and availability of structural information for proteins and small molecules, computer-based drug design has become a competitive methodology to identify new inhibitors,[4–7] and there are several practical applications of extending this methodology to screen for protein-peptide interactions, such as protein folding and docking, and inhibitor, agonist, and linker design. Although computational methods do not replace the in vitro tests during drug development and protein engineering, they can rule out many possibilities and propose unexpected new leads, and thus accelerate the early design stages. Our work on screening for peptidyl ligands to proteins complements the successes of others, who have shown it is possible to computationally design a peptide sequence that inhibits a protein[8] and accurately dock peptides to proteins.[9–13,61] Our goal in this work is to develop methodology that can effectively evaluate a large number of peptide sequences and their known structures for complementarity to a protein target.

A primary challenge for the computational discovery of inhibitors is to solve the docking problem, i.e., to predict the binding mode of a small ligand molecule in the active or ligand-binding site of a protein.

Grant sponsor: American Cancer Society, California Division; Grant number: S-65-92; Grant sponsor: National Science Foundation; Grant numbers: BIR 9631436 and BIR 9600831; Grant sponsor: Deutsche Forschungsgemeinschaft; Grant number: SCHN 576/1-1; Grant sponsor: MSU Research Excellence Funds for Academic Computing and Protein Structure, Function, and Design.

Volker Schnecke and Craig A. Swanson contributed equally to the research.

*Correspondence to: Leslie A. Kuhn, Protein Structural Analysis and Design Laboratory, Department of Biochemistry, Michigan State University, East Lansing, MI 48824-1319. E-mail: kuhn@agua.bch.msu.edu; WWW: http://www.bch.msu.edu/labs/kuhn

Received 20 January 1998; Accepted 1 June 1998

A more general task is to screen a database of small molecules for potential ligands to a given binding site. While the docking problem must be solved during this screening process, too, the approaches for docking in the context of screening are subject to a crucial restriction, the computation time per molecule. Docking a small flexible molecule with high accuracy takes at least several minutes on a desktop workstation for the fastest of the recent algorithms.[14–19] If only three minutes were spent per ligand candidate when screening a database of 100,000 structures, the resulting computation time would be more than six months. To allow screening within a reasonable time frame, some approximations are usually made in modeling the ways the protein and ligand can interact. For single-ligand ("fine") docking approaches, full flexibility of the ligand[14,16–18,20] and sometimes limited flexibility of the target protein[21–23,62] are considered. For screening, conformational modeling must be at least partially abandoned when evaluating tens of thousands of molecules for docking. The key to effective screening is to efficiently rule out infeasible candidates without losing the most promising ones, since in reducing 100,000 molecules to ~100 potential ligands, most of the time is spent eliminating poor candidates. The output of screening should ideally be a ranked list of 10–100 compounds for further inspection, including fine docking.

Existing docking algorithms can be classified into descriptor-based methods, grid-search or kinetic techniques, and fragment-based or incremental docking approaches.[6] All but the grid-search techniques employ a template that characterizes the binding site of the target protein. This template consists of points above the protein's solvent-accessible surface to be matched by ligand atoms. Spheres can be used for defining a negative image of the binding pocket, as in the DOCK tool[24] developed by Kuntz et al., which has been extended to consider chemical complementarity[25] and include hydrogen-bonding interaction centers[16] in addition to the shape template. Other current approaches specify a set of interaction points ("hot spots") defining favorable positions for polar interactions to the target protein. The docking tool Hammerhead[26] by Welch et al. docks ligands based on automatically generated probe points on the surface of the target protein.[27] The docking tool FLEXX by Rarey et al. is also based on discrete interaction points.[17,18] Whereas these approaches consider several different interactions, the use of hydrogen-bond donors or acceptors alone seems to give sufficient ligand specificity. Hydrogen bonds between protein and ligand are assumed to provide somewhat less of an energetic contribution than hydrophobic interactions to the stability of a protein-ligand complex, yet are essential for specificity.[28,29] The GOLD method of Jones et al. samples different conformers of an organic ligand and matchings of its donors and acceptors onto a hydrogen-bonding template, and transforms it based on the corresponding least-squares fit into the binding site.[22,23] Meyer et al. use potential hydrogen bonding positions between two molecules in protein-protein docking as starting points for a finer rotational and translational search of the rigid molecules.[30] ADAM by Mizutani et al. docks different conformations of ligands based on hydrogen bonds, then does a conformational search for the flexible parts not involved in the hydrogen-bonding pattern.[31]

Docking approaches based on techniques from distance geometry[32] have also been described. These techniques can be used to certify the feasibility of a mapping of ligand atoms onto template points in the binding site without explicitly computing the rigid-body transformation of the molecule to dock it. Flexibility is incorporated by considering lower and upper distance bounds for the ligand atoms and/or template points. Kuhl et al. propose a combinatorial algorithm to identify subsets of matching distances between two sets of points, based on graph theory.[33] Smellie et al. expand upon this approach and use it to efficiently generate mappings of different ligand conformations that are docked into the binding site and checked for steric fit.[34] Ghose and Crippen describe an approach to generate geometrically feasible binding modes for flexible ligands and propose a method to consider chirality.[35]

Most results presented for database screening have been obtained by docking tools originally developed for fine docking of single ligands. Sheridan and Venkataraghavan use an approach based on DOCK to search a database of 5,000 organic compounds for nicotinic agonists,[36] deriving a template by superimposing known agonists. DesJarlais et al. use DOCK to screen a subset of 2,700 randomly chosen structures from the Cambridge Structural Database.[37] Lawrence and Davis describe a tool CLIX, which maps ligand atoms to prespecified grid points in the target binding site.[38] They screen a subset of 30,000 structures from the Cambridge Structural Database in 33 hours to identify ligands to a mutant influenza-virus hemagglutinin. Shoichet et al. describe a successful application of DOCK for screening 55,000 compounds of the Fine Chemicals Directory[39] (FCD) for inhibitors to thymidylate synthase.[40] More recently, Gschwend et al. published the results of using DOCK for screening about 53,000 FCD entries within two weeks to identify selective inhibitors of fungal DHFR over human DHFR.[41] Makino and Kuntz have developed an extension of DOCK for flexibly docking ligand fragments, followed by energy minimization.[42] They screen a subset of 17,000 drug-like molecules from the Available Chemicals Directory (ACD, successor to the FCD) within a few days to identify potential dihydrofolate reductase (DHFR) ligands. Böhm describes the application of his fragment-docking tool LUDI[43,44] for screening a set of

30,000 small rigid molecules from the FCD and reports runtimes between 60 and 300 minutes when screening for potential ligands to four different proteins. Hammerhead by Welch et al.[26] incrementally constructs the ligand in the active site by docking and linking different conformations of fragments, with refinement based on a continuous and differentiable scoring function.[45] 80,000 entries of the ACD were screened within a few days, and full flexibility of the ligands was considered.

In light of the impressive recent results in this field, our goals with SPECITOPE (our algorithm for identifying a "Specific Epitope" that binds to a protein) were to increase the efficiency of eliminating poor candidates, incorporate side-chain flexibility for both the ligand and protein during screening and docking, and provide appropriate complementarity scores for potential ligands. In the results presented here, ligand-free structures were used for the target proteins to avoid bias during docking, and side-chain flexibility was considered for both the potential ligands and the target. This modeling of inducible complementarity was made possible by the efficient distance geometry steps used to select viable ligand candidates.

METHODS

SPECITOPE automatically screens a database of all peptidyl fragments from the PDB-Select list of dissimilar (<25% sequence-identical) protein chains[46] from the Brookhaven Protein Data Bank (PDB).[47] This database was chosen for three reasons: peptides are easily-synthesized, important leads for protein inhibitors and agonists; peptidyl interactions form the basis for recognition and docking between proteins; and the available peptide-bound and free structures for a number of proteins provide a basis for validating SPECITOPE'S results under the stringent criterion of docking to the free structure, which may require some conformational change.

The only human interaction in SPECITOPE is the design of the template, which specifies locations for polar ligand atoms and consists of up to five points from which hydrogen bonds can be formed to atoms in the target protein. For designing templates when only the structure of a single complex was available (as for two of our applications, uracil-DNA glycosylase and cyclodextrin glycosyltransferase), the complex was superimposed onto the ligand-free protein structure, and the template was based on the positions of polar atoms in the ligand that could form hydrogen bonds to atoms in the free protein. When several complexes with different ligands were available (as for aspartic proteinase), the complexes were superimposed onto the free target, and the average positions of polar ligand atoms involved in hydrogen bonds to target atoms were taken as the template points. For subtilisin, the template points were defined by the positions of five water molecules in the free target structure that were displaced by polar ligand atoms in a complex with eglin-c.

SPECITOPE identifies hydrogen-bond donors and acceptors of the ligand candidates that match the template points geometrically and chemically, and uses the orientation specified by the matched atoms for docking the molecules into the ligand-binding site. This hydrogen-bond template approach is equally applicable to organic molecule databases such as the Cambridge Structural Database and 3-dimensional structures derived from the Available Chemicals Database, and can be generalized to include hydrophobic and electrostatic interactions. For each molecule in our screening database, up to ten distance geometry, hydrogen-bond, steric, and hydrophobic complementarity checks are executed, as outlined below. In each step, molecules that do not meet a particular threshold are ruled out.

For each potential ligand in the database:

1. Compare the longest distance between all *m* polar (potentially hydrogen-bonding) atoms in the ligand with the longest distance between the *n* template points; if the longest intra-template distance exceeds the longest distance between polar ligand atoms by 2.0 Å, discard the ligand, since it cannot match the template.

For each set of n polar atoms in all ligands passing step 1 (computed as all subsets of n polar atoms from the m polar atoms in each ligand):

2. Check whether the number of hydrogen-bond donors and acceptors in each ligand set matches the number of donor and acceptor points in the template. If so, proceed.
3. Compare the shortest and longest distances between the polar atoms in the ligand set with those in the template; if the distances match within 1.4 Å, proceed. If they do not match, this ligand set exceeds the hydrogen-bond distance bounds and is excluded.
4. Compute the root-mean-square deviation between corresponding elements in the sorted lists of distances between atoms within the ligand set and between points within the template ($RMSD_{list}$, defined below). If the $RMSD_{list}$ value is below 0.7 Å, proceed.

For each possible matching between polar atoms in the ligand set and points in the template (computed as all possible one-to-one correspondences (permutations) between the n polar atoms and the n-point template):

5. Check whether the hydrogen-bonding activities are compatible for this matching, i.e., donors

are matched to donor template points and likewise for acceptors. If so, proceed.

6. Compute the distance matrix error (*DME*) for this matching (as detailed below); this gives a computationally inexpensive estimate of the superpositional RMSD of the polar ligand atoms onto the template. If the *DME* is less than 0.7 Å, proceed with this matching.

For the matching with minimal *DME* between the polar ligand atoms set and the template:

7. Transform the ligand onto the template using least-squares superposition of the *n* polar ligand atoms, given their one-to-one correspondence. If the RMSD of the *n* atoms in this superposition is below 1.0 Å, proceed.

For the transformed ligand:

8. Check for overlaps of ligand and target main-chain atoms, and determine if they can be resolved by iterative translations of the ligand as a rigid body.
9. After resolving any overlaps between main-chain atoms, check for overlaps between ligand side-chain atoms and target atoms, and resolve them, if possible, by minimally rotating side chains (ligand side chains first, then protein side chains).
10. For ligands with no remaining inter- or intramolecular overlaps, evaluate chemical complementarity using a scoring function based on hydrophobic contacts and the total number of hydrogen bonds (with favorable bond lengths and angles) between the docked ligand and protein, and reject all ligand orientations with fewer than two hydrogen bonds to the protein.

Note that steps 1 to 4 do not require a one-to-one correspondence of polar ligand atoms with template points, only the matching of their interatomic distances and hydrogen-bond activity (donor/acceptor) with the template points. Steps 5 and 6 combinatorially check all possible matchings of a set of *n* ligand atoms onto the *n* template points. Through step 7, only the polar ligand atoms are considered, then for steps 8–10, all ligand atoms are evaluated. The checks become computationally more complex by the end, so they are organized to rule out a maximal number of geometrically infeasible ligands in the early stages, before transforming them into the binding site.

Distance Geometry

SPECITOPE uses simple distance geometry[32] techniques to screen out ligands with incompatible geometry relative to a template specifying positions for polar ligand atoms. While the first distance check (step 1 above) considers all *m* polar atoms in the ligand, the remaining distance checks (steps 3, 4, and 6) only deal with a subset of atoms equal in number to the template points. All subsets of *n* polar atoms in the ligand are tested for their ability to match this template via a series of distance and hydrogen-bond complementarity screens. This involves checking all $m(m-1)\ldots(m-n+1)$ possible ways of matching all *n*-atom subsets of the *m* polar atoms in each ligand onto the *n* template points; however, the majority of these matchings can be ruled out based on the incompatibility between interatomic distances in the ligand and inter-template-point distances, as discussed below.

Given a set of *n* polar ligand atoms, a sorted list of their $n \cdot (n-1)/2$ interatomic distances, l_i, is compared to the sorted list, t_i, of distances between template points. With d_i equal to the difference between distances l_i and t_i, the root-mean-square deviation between distances in the two lists, defined as:

$$RMSD_{list} = \sqrt{\frac{2}{n \cdot (n-1)} \sum_{i=1}^{n\cdot(n-1)/2} (d_i)^2}$$

gives a measure for the compatibility of distances between the polar ligand atoms and between points in the template (step 4). A more exact measure for their compatibility is the distance matrix error:

$$DME = \sqrt{\frac{2}{n \cdot (n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (D_{ij})^2}$$

where $D_{ij} = L_{ij} - T_{ij}$ is defined as the matrix of differences between entries in the *L* and *T* matrices containing the distances between atoms and distances between template points, respectively (step 6). The $RMSD_{list}$ can be proven to give a lower bound for the *DME* for any matching of the two sets. Hence, if the $RMSD_{list}$ is above a given threshold for the current set of atoms, this set can be ruled out, since the *DME* for any one-to-one matching of these atoms to the template points can only exceed this value. An advantage of using the $RMSD_{list}$ as a screening criterion before the *DME* check is that the factorial complexity of specifying one-to-one correspondences between ligand and template points can be avoided for the majority of cases. We set the $RMSD_{list}$ and *DME* thresholds to 0.7 Å, because this restricts the search to sets that can match the template with reasonable accuracy (preserving the possibility of hydrogen bonding) and retains known ligands during screening.

In SPECITOPE, flexibility of the molecules is considered by allowing the single bonds in side chains to rotate. A simple heuristic is used to reflect side-chain flexibility during the distance checks without mak-

ing the distance criteria too loose: if one or two side-chain atoms are included in the pair of atoms for which a distance is compared, then the contribution of this distance to the $RMSD_{list}$ and DME (steps 4 and 6, respectively) is half as strong as for main-chain atoms. Thus, the term d_i in the formula for the $RMSD_{list}$ is actually calculated as:

$$d_i = \begin{cases} l_i - t_i & \text{if distance } l_i \text{ is between two main-chain ligand atoms} \\ 0.5 \cdot (l_i - t_i) & \text{otherwise} \end{cases}$$

and the flexibility adjustment for computing D_{ij} in the DME is:

$$D_{ij} = \begin{cases} L_{ij} - T_{ij} & \text{if distance } L_{ij} \text{ is between two main-chain ligand atoms} \\ 0.5 \cdot (L_{ij} - T_{ij}) & \text{otherwise} \end{cases}$$

Flexibility and Bump Resolving

After a set has passed all complementarity and distance checks (steps 1 through 6), the matched atoms are superimposed onto the template points by a least-squares transformation (step 7). This transformation involves a weighted least-squares fit that accounts for side-chain flexibility by weighting the contributions of side-chain atoms half as much as those of main-chain atoms. If the root-mean-square deviation of this superposition is above 1.0 Å, this atom set is ruled out; otherwise, the entire peptide is transformed into the binding site based upon the least-squares fit of the matched atoms onto the template points. The docked peptide is then checked for steric fit by computing the van der Waals overlaps of its atoms with protein atoms (steps 8 and 9), using van der Waals radii expanded to reflect the contributions of covalently-bonded hydrogen atoms (see Methods in Reference 48); an overlap tolerance of 0.3 Å is used for main-chain atoms, and 0.4 Å for side-chain atoms. Side-chain flexibility of both the ligand and the target protein are exploited, as described below, to identify an overlap-free orientation of the ligand in the binding site.

If there are, on average, no more than two overlaps per peptide residue between main-chain atoms of the two molecules, they are resolved by a translation of the peptide (step 8) as follows. For each overlapping pair of main-chain (including C_β) atoms, the translational direction for resolving all overlaps is computed by adding the vectors representing the minimal translations required for each atom. The ligand is then translated by this vector to resolve the overlaps. Of course, new main-chain overlaps might result from this translation. If there are still an average of no more than two main-chain ligand atoms per residue overlapping with the protein main chain, the same technique is applied and iterated up to 100 times. If main-chain overlaps remain, this particular matching of the ligand is rejected.

The side chains are considered to be the flexible parts in both molecules. Each side-chain single bond is taken as rotatable, and ligand side chains are tried first for resolving overlaps. Each ligand side chain is checked for overlaps to any protein atom, and overlaps are cleared by rotating this side chain through the minimal angle that resolves them (step 9). The single bond closest to the bumping atoms in the side chain is used first to resolve the overlap. If a bump-free conformation cannot be generated with this rotation, the next rotatable bond closer to the ligand backbone is rotated. The aim of this step in SPECITOPE is not to predict the optimal side-chain conformation, but to ensure that a bump-free conformation exists in the given ligand orientation. If it is not possible to resolve an overlap by rotating a ligand side chain, the same approach is applied to the protein side chain involved in the collision. If an intermolecular collision remains, despite testing all side-chain single-bond rotations, this ligand matching to the template is deemed too close and rejected. For ligand matchings in which all intermolecular overlaps have been resolved, both molecules are checked for intramolecular collisions. If a rotation has caused an internal clash, then the side chain is rotated back to its original conformation, and the next single bond closer to the backbone is rotated. This procedure is followed by rechecking for inter- and intramolecular collisions, until either a collision-free conformation is found, or all possibilities have been exhausted and this ligand matching is excluded.

Scoring

In SPECITOPE, complementarity evaluation of a complex is done only for the ~100 peptides passing the previous checks. A molecule that has passed all checks is considered a potential ligand based on its hydrogen bonds and overlap-free steric fit to the protein, but it cannot be assumed that all aspects of the ligand conformation and orientation are optimal. Thus, the scoring function of SPECITOPE (step 10) is mainly used to recognize molecules that lack chemical complementarity and to emphasize those molecules that fit well in the given binding mode.

The scoring function is based on two terms, the number of hydrogen bonds between protein and ligand, and the hydrophobic complementarity between the two molecules. Because hydrogen atom positions are not present in most PDB structures, SPECITOPE computes the optimal position of the shared hydrogen to identify intermolecular hydrogen bonds with good geometry, for donors and acceptors separated by 2.8 to 3.5 Å. The hydrogens of the N-terminus and lysine side-chain amino groups and the hydrogens of the serine, threonine, and tyrosine hydroxyl groups are assumed to be free to rotate on a

circle (defined by the D–H bond length and X–D–H angle, where X is the non-hydrogen atom covalently bonded to the donor) and are directed to the nearest acceptor.[49] For all other donors, the hydrogen position is unambiguous, and donation to multiple acceptors is considered if the angular constraints are fulfilled. A distance of 1.0 Å is used between the donor and the hydrogen atom, and a range of 140° to 180° is accepted for the D–H · · · A angle.[50] All hydrogen bonds are considered as giving equivalent contributions to the overall complementarity.

The hydrophobicity measure is based on a statistical survey of atomic hydration in 56 protein structures.[48] The contribution of a single ligand atom is based on the comparison of its hydrophobicity value with the average hydrophobicity of the surrounding protein surface atoms. Given the hydrophobicity $h(a)$ of an atom a, with $h(a) \in [0 \ldots 635]$ calculated as the average number of hydrations per 1000 occurrences of that atom type (Table II in Reference 48), a value of 0 represents a maximally hydrophobic atom, 635 is maximally hydrophilic, and 317 is intermediate. The hydrophobic complementarity of the contact surface between protein P and ligand L is computed as:

$$\mathrm{HPHOB(P, L)} = \sum_{\substack{l_i \in L \\ \#P_i > 0}} \frac{\mathrm{avg}\,\{h'(l_i),\ \overline{h}(P_i)\}}{\max\{\mathrm{abs}\,(h'(l_i) - \overline{h}(P_i)),\ 10\}}$$

where

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

considers only the hydrophobic contribution of ligand atoms l_i. The hydrophobicity $\overline{h}(P_i)$ of the neighboring protein atoms P_i for a single ligand atom l_i is defined as the average hydrophobic contribution of all protein atoms p_j within a distance of 4.0 Å of l_i:

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

Note that for computing the average hydrophobicity for the protein neighborhood of a ligand atom, the hydrophilic atoms are also considered, since the maximum is taken after computing the average hydrophobicity of the protein atoms. This results in a lower $\overline{h}(P_i)$ score for a neighborhood containing hydrophilic atoms, since this term is designed to measure favorable hydrophobic-hydrophobic contacts; favorable hydrophilic interactions are taken into account separately by the hydrogen-bond term (described below). The denominator in each term of the sum describing the hydrophobic score (HPHOB(P, L)) is always greater than or equal to 10, which is 3% of the maximum score for a single ligand atom. This ensures that the overall HPHOB(P, L) score is not dominated by a few contacts with very small differences between protein and ligand hydrophobicity.

The overall complementarity of a protein-ligand complex is given by a weighted sum of the number of hydrogen bonds and the hydrophobic complementarity:

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

Based on the functions of Böhm[51] and Jain,[45] a ratio of 1:1.2 is assumed for the relative contributions of the hydrogen bond and hydrophobic interaction terms to the overall stability of the protein-ligand complex. The weights A and B have been empirically tuned using 30 protein complexes with small peptidyl ligands from the PDB. The average number of intermolecular hydrogen bonds in these complexes is 6.3, and the average value of HPHOB(P, L) is 49.7, which gives weights of 158.0 and 24.2 for A and B, respectively, yielding a ratio of 1:1.2 for the hydrogen-bond and hydrophobic terms.

RESULTS

A database of ~140,000 peptides was screened by SPECITOPE to identify potential ligands to the active sites of four different enzymes. The database included all overlapping peptides of fixed length from structures of diverse chains in the PDB.

First, the SPECITOPE scoring function was validated by comparing the scores for protein-peptide complementarity in 30 known complexes with those for the complementarity of 34 buried and 34 surface peptides in proteins with their surroundings (Fig. 1). In the latter two cases, the main-chain atoms preceding and following the selected peptide were removed, and the peptide's complementarity to the remaining protein structure was checked in its original position. From the statistics (Table I), it is apparent that the complementarity for buried peptides is one-and-a-half times that of surface peptides and peptidyl ligands, and the best peptidyl ligands rival the complementarity of buried peptides (Fig. 1). These scores provide baselines for interpreting the results of SPECITOPE screening, since an optimal peptidyl ligand essentially becomes part of the protein.

SPECITOPE has been used to identify potential ligands to four protein targets, subtilisin, uracil-DNA glycosylase, aspartic proteinase, and cyclodextrin glycosyltransferase, which interact with natural ligands ranging from peptides to DNA to oligosaccharides. They were chosen because a ligand-free structure exists for each, providing a greater challenge for docking, since side-chain motion may be required upon ligand binding. There were some significant differences in the active-site conformations for the free protein structures versus their crystallographic complexes (Table II). Also, for three of the four cases

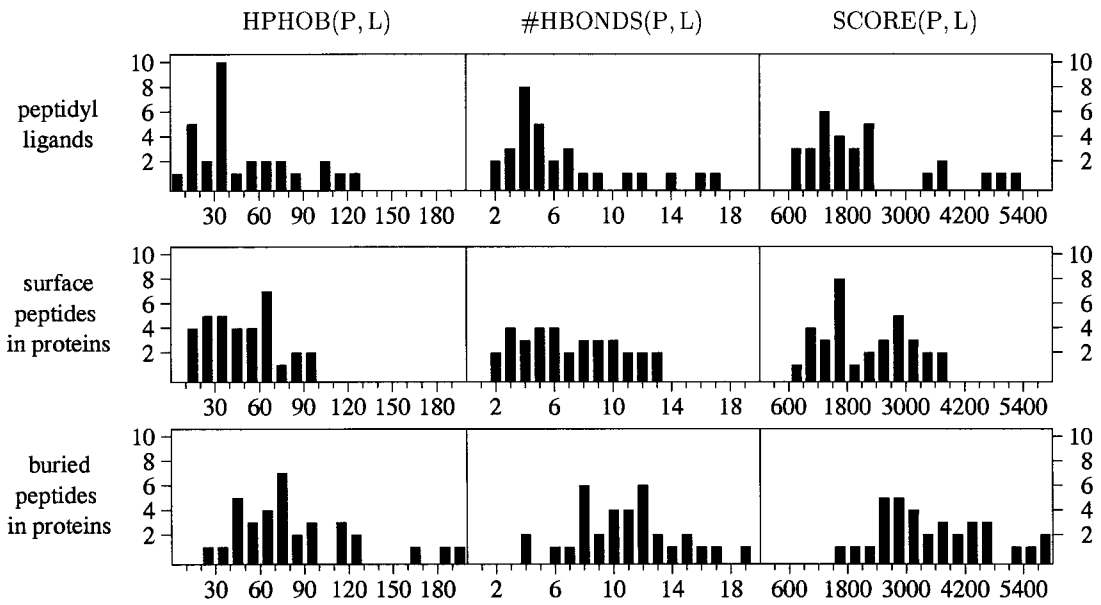

Fig. 1. The distribution of values for HPHOB and #HBONDS terms and overall SPECITOPE scores for a series of known interactions between peptides (L) and proteins (P), with one vertical unit per peptide complex. Results for 30 protein-peptide complexes are given (first row), followed by the scores for 34 surface and 34 buried peptides within protein structures (second and third rows, respectively), evaluated for complementarity with their protein surroundings.

TABLE I. The Average and Standard Deviation, σ, of the Complementarity Scores for Three Test Sets†

	Peptidyl ligands		Surface peptides		Buried peptides	
	Average	σ	Average	σ	Average	σ
Number of atoms	44.67	21.89	52.74	15.00	53.71	19.87
HPHOB(P, L)	49.68	32.00	49.37	23.54	84.61	41.34
# HBONDS(P, L)	6.33	3.96	7.03	3.29	10.74	3.41
SCORE(P, L)	2102.78	1217.56	2196.65	877.79	3573.94	1094.61

†30 peptidyl ligands from known complexes with proteins, and 34 surface and 34 buried peptides within proteins, each evaluated for complementarity with their protein surroundings.

there are structures of complexes with peptidyl or semi-peptidyl ligands, providing a convenient way of validating potential ligands identified by SPECITOPE from screening the peptidyl database. The fourth case, cyclodextrin glycosyltransferase, was chosen to test whether SPECITOPE can identify peptidyl mimics with similar shape and binding chemistry to other kinds of ligands (e.g., carbohydrates), despite their different molecular backbones. Several approaches were used to design the four- or five-point templates (Table III) for ligand binding to our targets. In each case, known complexes were used to identify favorable binding points in the active site. To avoid bias, these template points were specified in the context of the ligand-free structure.

Subtilisin

Subtilisin is a bacterial serine protease that cleaves its peptidyl substrate using a catalytic triad of Asp-Ser-His residues, but is otherwise divergent from the mammalian serine proteases. One of its naturally occurring inhibitors is a small protein, eglin-c, with a surface loop that binds to the active site of subtilisin.[1]

To generate the template for subtilisin, the subtilisin-eglin-c complex (PDB entry 2sec) was superimposed onto the free subtilisin structure (PDB 1s01). The interaction template points were chosen from the positions of five water molecules (HOH 413, HOH 463, HOH 493, HOH 495, and HOH 498) in 1s01 that are displaced by atoms in eglin-c upon complex formation. During screening, these sites specified the positions of one hydrogen-bond donor and four acceptors to be matched by the ligands.

SPECITOPE screened a data set of 139,253 pentapeptides and identified 23 potential ligands with appropriate steric fit and a range of complementarity scores, within a runtime of 65 minutes. All timing data are CPU times on a Sun Ultra1 140 MHz workstation (SPECint95, 4.66, SPECfp95, 7.90;

TABLE II. The Crystallographically Observed Conformational Change of the Active Site When Binding the Known Ligand†

PDB code (free/complex)	All atoms	Main chain
1s01/2sec		
RMSD	0.87	0.49
Maximum displacement	3.40	0.89
Number of atoms	72	40
HUDG/HUDG-UGI		
RMSD	1.00	0.46
Maximum displacement	3.64	1.11
Number of atoms	63	28
2apr/3apr		
RMSD	0.32	0.31
Maximum displacement	1.08	1.08
Number of atoms	111	56
1cgt/1cgu		
RMSD	0.89	0.46
Maximum displacement	2.96	1.02
Number of atoms	93	36

†The active site for each enzyme was defined as all residues having atoms within 3.5 Å of any ligand atom in the crystal structure. RMSD and displacement values (in Å) are based on protein C_α superposition between the complex and free structures.

8 Mb RAM were used by SPECITOPE). The epitope from the known inhibitor (residues I56 to I60 in complex 2sec) obtained the top rank (Table IV). While this structure was included in the screened database for verification, SPECITOPE also found and assigned the third-highest score to the same epitope from a different eglin-c structure, one of the 553 non-homologous PDB chains used to construct the peptidyl database. The top-scoring known ligand docked by SPECITOPE closely matched the orientation of the eglin-c epitope from complex 2sec superimposed onto the free structure (Fig. 2). The main-chain RMSD of the peptide docked by SPECITOPE in comparison to the epitope in the complex was 0.93 Å (based on protein superposition only), resulting from a slightly different ligand placement (maximum main-chain displacement: 1.23 Å). Four side chains in subtilisin and one in eglin-c were rotated during docking, and the large conformational change of the subtilisin tyrosine side chain upon eglin-c docking also occurs in the experimentally defined complex. Although the overall conformational difference in the active site of the free subtilisin structure (1s01) is rather small (main-chain RMSD 0.49 Å) compared to the corresponding complex (2sec), the maximal side-chain atom displacement is 3.40 Å (Table II).

The average number of predicted intermolecular hydrogen bonds for the 23 peptides docked to subtilisin was 2.2, and the top ligand made 4 hydrogen bonds (Table IV). In fact, all of the hydrogen bonds predicted by SPECITOPE docking of known ligands for subtilisin, uracil-DNA glycosylase, and aspartic proteinase were also observed in their crystallographic complexes. Although the potential ligands identified by SPECITOPE have donors or acceptors matching the template, this generates a preference rather than guarantee of forming hydrogen bonds at these points, since the hydrogen-bond geometry can only be checked once the ligand has been transformed into the active site (in step 7). The SPECITOPE-identified hydrogen bonds are likely to be a subset of the number that may be attained by flexibly fine-docking the same ligand. All intermolecular hydrogen bonds, including template-based ones, were counted in SPECITOPE'S scoring function upon docking the ligand.

Uracil-DNA Glycosylase

Uracil-DNA glycosylase recognizes uracil that has been incorporated inappropriately into DNA and initiates base excision repair by hydrolyzing the bond linking the uracil base to deoxyribose. A native inhibitor of uracil DNA glycosylase is the 82-residue protein UGI (for "uracil-DNA glycosylase inhibitor"), which mimics DNA binding.[52]

When designing the template for uracil-DNA glycosylase, only the structures of the free enzyme (HUDG) and the complex with UGI were used.[52,53] The complex was superimposed onto the free structure, and a five-residue linear epitope in the inhibitor was chosen for defining the template. The positions of one donor and four acceptors that are involved in intermolecular hydrogen bonds in the complex and are also within hydrogen-bonding distance to the corresponding atom positions in the free target structure were taken as template points.

SPECITOPE identified 14 potential ligands for HUDG by screening 139,331 pentapeptides (runtime 38 min). The linear epitope from the known inhibitor, UGI, ranked fifth (Table V) and was docked very similarly to the orientation of UGI in the crystallographic complex; the main-chain RMSD (based on protein superposition) for the ligand in SPECITOPE'S orientation versus that in the complex, was 0.28 Å, even though the active-site conformation in the free structure used for docking had main-chain displacements of up to 1.1 Å relative to the complex (Table II). The known and top-ranked ligands had the same backbone conformation and similar side-chain shape and chemistry (Fig. 3), despite these parameters not being specified by the template. The potential ligands generally had the sequence pattern (polar)(negatively charged)(X)(X)(hydrophobic), with some preference for polar side chains in the X positions.

Aspartic Proteinase

Rhizopuspepsin, an aspartic proteinase, is a homolog of medically important inhibitor design targets including renin, which is active in the vasoconstriction pathway associated with high blood pressure, and HIV protease, which is essential for processing the gag and gag-pol polyproteins to produce infec-

TABLE III. The Template Characteristics for the Four Target Proteins†

Target protein	PDB code	Number of template points	Shortest distance	Longest distance	Average distance	Standard deviation	Peptide length
Subtilisin	1s01	5 (1D, 4A)	2.4	11.5	7.1	3.3	5
Uracil-DNA Glycosylase	HUDG	5 (1D, 4A)	3.0	14.0	8.7	3.2	5
Rhizopuspepsin	2apr	5 (4D, 1A)	2.2	11.3	6.3	2.9	6
Glycosyltransferase	1cgt	4 (2D, 4A)	5.6	9.5	7.1	1.6	2

†The numbers of donor (D) and acceptor (A) atoms are the maximal number required to match the template; the two donor points and two donor or acceptor points in cyclodextrin glycosyltransferase could be matched by as many as four acceptors or as few as two acceptors plus two donors in the ligand. Distances between donor/acceptor points are in Å and peptide lengths are in residues.

TABLE IV. The Top-Five Potential Ligands for Subtilisin (PDB 1s01) Identified by Specitope†

Rank	PDB	Residues	Sequence	SCORE(P, L)	# HBONDS(P, L)
1	**2sec**	**i56-i60**	PVTLD	**1186.4**	**4**
2	1ctn	208-212	QFSGE*	1091.2	2
3	**1cse**	**i42-i46**	PVTLD	**953.1**	**3**
4	1nif	276-280	TEQDL*	833.1	2
5	1tht	a265-a269	DGGSL	818.8	2
1–23	Average (standard deviation)			639.2 (216.5)	2.2 (0.5)

†Sequences marked with * have been reversed, because the corresponding ligand was bound in an orientation opposite to that of the known ligand. The 2sec and 1cse matches represent the same epitope from different structures of the known subtilisin inhibitor, eglin-c, ranking first and third in the screening.

TABLE V. The Top-Five Potential Ligands for Human Uracil-DNA Glycosylase†

Rank	PDB	Residues	Sequence	SCORE(P, L)	# HBONDS(P, L)
1	1tss	a16-a20	GSDTF	972.2	2
2	1mxa	170-174	DDYQF*	927.4	2
3	1dih	255-259	SEKGS*	883.8	2
4	1bgl	a293-a297	NEVNL*	881.8	2
5	**ugi**	**19-23**	QESIL	**867.8**	**2**
1–14	Average (standard deviation)			704.6 (189.9)	2.1 (0.3)

†The linear epitope from the known inhibitor, UGI, ranked fifth. Sequences marked * are given in reverse order to reflect their orientation relative to the known inhibitor.

tious virions.[54] The template for rhizopuspepsin was designed by superimposing three complexes of this protein with pepstatin (PDB 6apr) or pepstatin-like renin inhibitors (4apr, 5apr) onto the free structure (2apr). The average positions of four hydrogen-bond donors and one acceptor in the three inhibitors were selected as the template points. The conformation of the ligand-free active site is highly conserved relative to the 3apr complex (all-atom RMSD: 0.32 Å, Table II), with a maximum atomic displacement of 1.08 Å.

SPECITOPE identified 53 potential ligands (Table VI shows the top five) out of a set of 138,710 hexapeptides in 153 minutes. Since no rhizopuspepsin inhibitors were in the set of non-homologous protein chains screened for ligands, for verification we included a known peptidyl ligand, chain I in PDB entry 3apr, in the screening database. This ligand was not used in designing the template; however, it obtained the top rank, with a score of 2891, which is comparable to the values for a series of known protein-peptide interactions (Table I). The scores for the top rhizopuspepsin ligands were higher than for subtilisin and uracil-DNA glycosylase, including their known ligands, because the binding pocket of this target is a narrow cleft, yielding a larger interface between the molecules. For subtilisin and uracil-DNA glycosylase, the known ligands were continuous epitopes forming a critical part of a larger interface. The orientation of the known peptidyl inhibitor of rhizopuspepsin proposed by SPECITOPE was very similar to the orientation of this ligand superimposed from the corresponding complex (Fig. 4). Side-chain rotations upon docking were visible both for the ligand and target side chains; although the conformational changes of the target side chains were small, they were necessary to generate an overlap-free orientation. The main-chain RMSD for

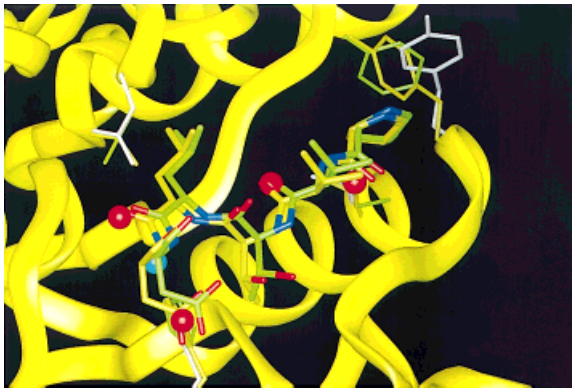

Fig. 2. The top-ranked SPECITOPE ligand for subtilisin is the epitope from the known ligand, eglin-c. The main-chain ribbon and orientations of key side chains are shown in yellow for the ligand-free subtilisin structure (PDB 1s01), with the SPECITOPE-identified and docked peptide from eglin-c shown in green at center (carbon atoms: green, nitrogen: blue, oxygen: red). Subtilisin side chains reoriented by SPECITOPE are also shown in green, whereas the orientations of these side chains in the known crystallographic complex (PDB 2sec) are shown by white tubes. The eglin-c epitope from this complex is shown in yellow tubes, positioned by superimposing the complex with the ligand-free main chain. Donor and acceptor template points are shown as blue and red spheres, respectively.

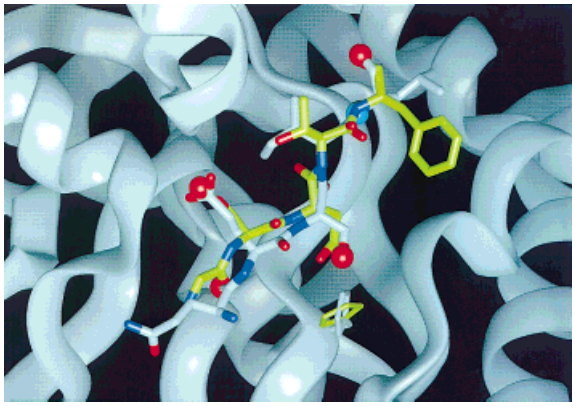

Fig. 3. The epitope from the known inhibitor, UGI, is shown superimposed from the complex with human uracil-DNA glycosylase (blue) together with SPECITOPE'S top-ranked ligand, sequence GSDTF (green) from PDB entry 1tss. Note the similarity between atom positions and chemistry for the UGI and 1tss ligands. The five template points are indicated by positions of the hydrogen-bond donor (blue sphere) and acceptors (red spheres). The histidine side chain in uracil-DNA glycosylase that was rotated by SPECITOPE upon binding the 1tss epitope is also shown beneath the ligand (blue: His from ligand-free structure; green: His from complex with 1tss peptide).

the docked and crystallographic ligand orientations was 0.97 Å, based on protein superposition only.

Cyclodextrin Glycosyltransferase

Cyclodextrin glycosyltransferase catalyzes the degradation of starches and starch-like compounds by partially converting them into cyclodextrins, which

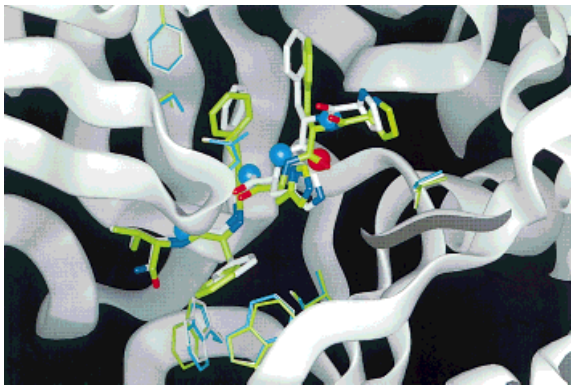

Fig. 4. The known ligand received the top rank upon screening peptides for complementarity to the active site of rhizopuspepsin (peptidyl ligand from PDB 3apr complex and main chain from ligand-free 2apr are shown in grey). SPECITOPE'S docking of the peptidyl ligand is shown by green tubes, and all side chains that were rotated upon ligand binding are shown in their native conformation in the free structure (white), after SPECITOPE rotation (green), and superimposed from the known complex with this ligand (blue). The four donor and one acceptor template points are shown as blue and red spheres, respectively.

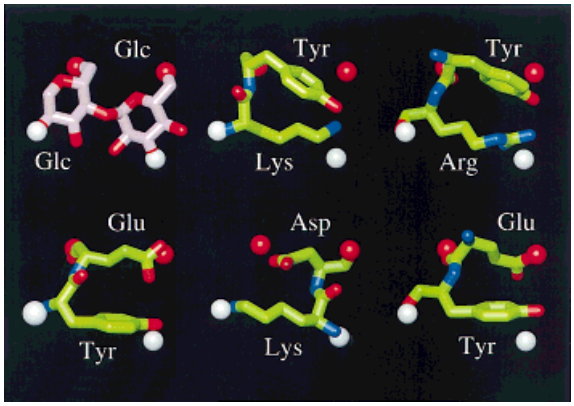

Fig. 5. The known ligand for cyclodextrin glycosyltransferase (PDB 1cgt), a disaccharide consisting of two glucose residues (pink tubes, upper left, from the complex, PDB 1cgu), is shown together with the top five dipeptides SPECITOPE identified. All ligands are shown in the docked orientation relative to the four template points (red spheres were matched by hydrogen-bond acceptors, and white spheres were matched by either donor or acceptor). Note the similar main-chain and side-chain conformations of four of the five SPECITOPE ligands, whereas the fifth (second row, center) effectively mirrors the others.

can be imported into cells and metabolized.[55] SPECITOPE screened 140,885 dipeptides (being similar in size to the disaccharide bound in the known complex) to match a four-point template for cyclodextrin glycosyltransferase. This template was designed by superimposing the glycosyltransferase complex with two glucose residues (1cgu) onto the free target structure (1cgt) and identifying key hydrogen-bond donor and acceptor positions. The SPECITOPE runtime was 2 minutes, and 13 potential ligands were identified (Table VII). The structures of the five top-ranked

TABLE VI. The Five Top-Scoring Ligand Candidates for the Aspartic Proteinase, Rhizopuspepsin (PDB 2apr)†

Rank	PDB	Residues	Sequence	SCORE(P, L)	# HBONDS(P, L)
1	**3apr**	**i2-i7**	PFHFFV	**2891.3**	**4**
2	1sbp	92-97	PYTSTI	2541.9	4
3	1ypt	b311-b316	ITVESK	2489.0	4
4	2por	4-9	LSGDAR	2161.1	3
5	1prt	b110-b115	RLLSST	1916.5	2
1–53	Average (standard deviation)			1307.0 (515.8)	3.3 (1.0)

†The known inhibitor, Pro-Phe-His-Frd-Phe-Val, obtained the top rank.

TABLE VII. The Potential Ligands SPECITOPE Identified for Glycosyltransferase (PDB 1cgt)†

Rank	PDB	Residues	Sequence	SCORE(P, L)	# HBONDS(P, L)
1	1rva	a229-a230	KY	1870.8	2
2	2dkb	173-174	RY*	1762.3	3
3	5rub	a72-a73	YE	1542.0	3
4	3cla	78-79	KD	1206.2	2
5	1ubs	a202-a203	YE*	1205.0	2
1–13	Average (standard deviation)			998.7 (491.5)	2.3 (0.5)

†The sequence order has been reversed for peptides (*) oriented oppositely to the top-ranked ligand.

TABLE VIII. The Number of Peptides Checked by SPECITOPE in Different Stages of the Screening Process for the Four Proteins, and Their Runtimes on a Sun SPARC Ultra1 Workstation

Number of peptides	HUDG	1s01	2apr	1cgt	Average	Time
Total (database size)	139,331	139,253	138,710	140,885	139,545 (100.00%)	0.00%
After step 6 (DG checks)	10,192	52,166	102,195	1,855	41,602 (29.81%)	91.82%
After step 7 (RMSD check)	920	12,025	53,264	361	16,643 (11.93%)	92.69%
After step 8 (main chain)	768	1,364	2,079	328	1,135 (0.81%)	99.25%
After step 9 (side chain)	96	173	76	20	91 (0.07%)	99.86%
After step 10 (scoring)	14	23	53	13	26 (0.02%)	100.00%
Screening time (CPU min.)	38	65	153	2	65	

dipeptides showed interesting similarities to the structure of the glucose residues (Fig. 5) from PDB entry 1cgu, all shown in their orientation relative to the template. A horseshoe conformation and similar side-chain chemistry were adopted by all dipeptides mimicking the disaccharide. The (K,R)(Y) ligands were selected by SPECITOPE from more than 600 occurrences of this sequence pattern (enumerated by the PDB sequence-pattern searching software, SEQUERY[56,57]) in the 140,885 dipeptides, and thus reflect the conformational as well as side-chain specificity of the site. 293 occurrences of (Y)(E) and 526 occurrences of (K)(D) were screened.

DISCUSSION

Almost all work published on ligand database screening involves tools that were developed and optimized for fine docking, that is, to accurately predict the binding mode of a single, sometimes flexible, compound. The goal of SPECITOPE is the rapid screening and identification of potential ligands out of a large set of compounds that have a known favorable conformation (e.g., from crystallographic structure determination).

SPECITOPE's strength is its ability to rule out most infeasible candidates in the early steps, based only on distance geometry and hydrogen-bond complementarity checks. Applications of distance geometry to docking have been reported,[33-35] but interatomic distance correspondence alone proved insufficient for defining a feasible binding mode. However, because the matching of intramolecular distances (equivalently, shapes) for the protein and ligand is a necessary criterion during docking, distance geometry can be used to exclude poor candidates. On average, more than 70% of the ligand candidates were ruled out (reduction between lines 1 and 2 in Table VIII) by the initial checks, taking about 90% of the overall screening and docking time. The number of peptides to be checked for main-chain overlaps with the target was then significantly reduced, by 60% on average (from the number of peptides passing distance checks; reduction from Line 2 to Line 3 in Table VIII) by ruling out all peptides with a

root-mean-square deviation greater than 1.0 Å from superimposing the polar ligand atoms onto the template. The screening time was mainly influenced by the size of the molecules screened, since the most complex step is the enumeration of all possible matchings of polar ligand atoms to template points. Another factor is the shape of the template, which ruled out most molecules during the distance checks for the compact glycosyltransferase template, but let many molecules pass for aspartic proteinase.

The conformational search for each ligand assesses whether collisions can be resolved, resulting in a feasible conformation and orientation. Leach and Kuntz[16] use a similar approach to resolve side-chain overlaps and maximize hydrogen-bond interactions during the conformational search when fine-docking a flexible ligand. In contrast to their approach, where all possible conformers for the side chain are tested and the lowest-energy one is chosen, SPECITOPE proceeds along the rotatable bonds closest to the bumping atoms to resolve overlaps in that side chain. However, as in their approach, the number and quality of side-chain hydrogen bonds could be optimized when selecting a bump-free conformation. Because of the time savings provided by the distance geometry checks, more computationally intensive strategies can be used at this stage of SPECITOPE, such as fine docking to optimize the binding mode for the top candidates. While the current version of SPECITOPE assumes a rigid ligand backbone—a reasonable assumption for peptidyl epitopes that are part of a larger inhibitor, and for polycyclic organic structures—future extensions to the method will include backbone flexibility via single bond rotations between rigid substructures.

An important difference between our approach and other published screening results is that we allow conformational changes in the target protein by searching side-chain conformers explicitly once the ligand has been docked. Known ligands can be identified by other screening methods in part because they screen against the active-site structure from the complex, a simplified case in which the necessary side-chain conformational changes have already been made. Importantly, for all our test cases, active-site side-chain conformational change relative to the free structure is required for interaction between the known ligand and the target protein, both in SPECITOPE docking and in the crystallographic complex (Table II). In other protein-peptide complexes, inducible side-chain conformational changes are known to be important for docking and complex formation.[13]

The quality of any docking tool depends on the accuracy of its scoring function. Even if it were possible to determine the binding free energy exactly, the remaining work of identifying the ligand binding mode providing maximal affinity would mean searching an intractable number of orientations and conformations. Empirically tuned scoring functions to estimate the binding energy of a protein-ligand complex have been proposed by Böhm[51] and Jain,[45] and in the newest version of the tool AutoDock,[58] such a function has replaced the original, forcefield-based scoring function.[20] These semi-empirical functions consider hydrogen bonds, ionic interactions, the hydrophobic character of the interface, and the loss of entropy from binding a free, flexible ligand. SPECITOPE's scoring function considers the hydrogen-bond and hydrophobic complementarity of the protein-ligand interface, since these are the dominant terms. The function has been validated by comparing the scores of known peptidyl ligands with those for buried and surface peptides within proteins, and yielded high ranks for the three known ligands in our test cases. The scoring function could also distinguish between the best (including known) and average ligand candidates; for the four proteins we analyzed, the best peptide had a score 1.4 to 2.2 times as high as the average score for sterically feasible ligands. Adding a term for the van der Waals packing and contact area between the molecules in the scoring function would improve the ligand selectivity for proteins with open active sites, as would including a term to reflect the favorability of having hydrophilic, rather than hydrophobic, atoms in the solvent-exposed portion of the ligand.

It is difficult to compare SPECITOPE runtimes directly with those of other screening approaches, because the other methods assume a rigid protein and differ in the degree of ligand flexibility, and their runtimes do not include scoring done outside the screening program, such as molecular graphics assessment of the complexes. Accounting for the different database sizes, runtimes for recent methods assuming rigid ligands are a few hours, roughly comparable to those of SPECITOPE (a few minutes to a few hours, with side-chain flexibility), whereas methods modeling flexible ligands take a few days. Another consideration is that SPECITOPE requires no precomputation of partial charges or interaction grids and is fully automated, aside from the specification of a four- or five-point template representing favorable sites for polar ligand atoms.

For designing the ligand template, we employed three strategies based on the structures of known complexes, as described in Methods. However, hydrogen-bonding templates can be rationally designed using bound water positions when structural data for binding to other ligands is unavailable. A tool, *Consolv*, has been developed in our laboratory to predict the conservation or displacement of protein-bound water molecules upon ligand binding, based on the favorability of their environments.[59] *Consolv* can be used to identify water molecules that are likely to be displaced by polar ligand atoms, and thus provide a rational basis for template design. Furthermore, hydrophobic and electrostatic interactions can be incorporated into future SPECITOPE templates.

The success in identifying key epitopes from known peptidyl inhibitors for subtilisin, uracil-DNA glycosylase, and aspartic proteinase suggests SPECITOPE can be used as a protein-protein docking algorithm, as well as a way to identify inhibitor and agonist leads and peptidyl linkers. Furthermore, SPECITOPE can provide structural insight into the binding modes of ligands identified by in vitro peptide library screening.[57,60] Conversely, SPECITOPE can be applied to screen for organic mimics of peptidyl ligands, which is now being evaluated by creating an interface to the Cambridge Structural Database of small organic structures. Since organic molecules are generally less polar than peptides, screening based only on a hydrogen-bond template will no longer suffice; Jones et al. have noted this when applying their tool GOLD, which is also based on matching polar ligand atoms to a hydrogen-bonding template, to ligand complexes in which hydrophobic interactions are dominant.[23] In the extension of SPECITOPE, hydrophobic interaction sites will be included in the template in addition to the hydrogen-bonding pattern. For screening organic molecules, van der Waals overlaps can be handled by defining rigid units (e.g., cyclic structures, analogous to the peptide backbone) and flexible substituents (analogous to side chains) and using the current methodology of directed, minimal translations and rotations.

CONCLUSION

SPECITOPE is currently able to screen over 100,000 potential ligands and identify and dock a small set (usually tens) of high-scoring ligand candidates in less than two hours. This speed results from the powerful and computationally efficient distance geometry checks, which rule out a majority of infeasible ligand candidates before transforming them into the active site. Our implementation of both protein and ligand side-chain flexibility and the use of ligand-free protein structures as targets allows more realistic docking while screening. Furthermore, side-chain inducible complementarity was crucial for the identification and docking of the known ligands for three protein targets. Potential ligands are scored by SPECITOPE based on the number of intermolecular hydrogen bonds and the hydrophobic complementarity for the protein-ligand interface. For two of the three proteins with known peptidyl ligands, the known ligand received the top SPECITOPE score, and in all three cases, it ranked within the top five of the 140,000 molecules screened.

ACKNOWLEDGMENTS

We thank Garry Gippert and Michael Pique for their ideas on applying distance geometry to this problem, and Michael Pique and David Goodsell for their critical feedback on the manuscript. We also acknowledge the American Cancer Society, California Division (senior postdoctoral fellowship S-65-92 to L.K.) and the National Science Foundation (grant BIR 9631436 to E.D.G. and J.A.T.) for supporting the early work at Scripps, and the Deutsche Forschungsgemeinschaft (postdoctoral fellowship SCHN 576/1-1 to V.S.), the MSU Research Excellence Funds for Academic Computing and Protein Structure, Function, and Design, and the National Science Foundation (grant BIR 9600831 to L.K.) for supporting our ongoing work on this project at Michigan State University.

REFERENCES

1. McPhalen, C.A., James, M.N.G. Structural comparison of two serine proteinase-protein inhibitor complexes: Eglin-c-subtilisin Carlsberg and CI-2-subtilisin novo. Biochemistry 27:6582–6598, 1988.
2. Livnah, O., Stura, E.A., Johnson, D.L., Middleton, S.A. et al. Functional mimicry of a protein hormone by a peptide agonist: The EPO receptor complex at 2.8 Å. Science 273:464–471, 1996.
3. Hruby, V.J., Al-Obeidi, F., Kazmierski, W. Emerging approaches in the molecular design of receptor-selective peptide ligands: Conformational, topographical, and dynamic considerations. Biochem. J. 268:249–262, 1990.
4. Verlinde, C.L., Hol, W.G. Structure-based drug design: Progress, results and challenges. Structure 2:577–587, 1994.
5. Kuntz, I.D. Structure-based strategies for drug design and discovery. Science 257:1078–1081, 1992.
6. Kuntz, I.D., Meng, E.C., Shoichet, B.K. Structure-based molecular design. Accounts of Chem. Res. 27:117–123, 1994.
7. Gschwend, D.A., Good, A.C., Kuntz, I.D. Molecular docking towards drug discovery. J. Mol. Recognition 9:175–186, 1996.
8. Moon, J.B., Howe, W.J. Computer design of bioactive molecules: A method for receptor-based de novo ligand design. Proteins 11:314–328, 1991.
9. Vakser, I.A. Evaluation of GRAMM low-resolution docking methodology on the hemagglutinin-antibody complex. Proteins Suppl. 1:226–230, 1997.
10. Gulukota, K., Vajda, S., DeLisi, C. Peptide docking using dynamic programming. J. Computat. Chem. 17:418–428, 1996.
11. Wallqvist, A., Covell, D.G. Docking enzyme-inhibitor complexes using a preference-based free-energy surface. Proteins 25:403–419, 1996.
12. Fischer, D., Lin, S.L., Wolfson, H.L., Nussinov, R. A geometry-based suite of molecular docking processes. J. Mol. Biol. 248:459–477, 1995.
13. Friedman, A.R., Roberts, V.A., Tainer, J.A. Predicting molecular interactions and inducible complementarity: Fragment docking of Fab-peptide complexes. Proteins 20: 15–24, 1994.
14. Gelhaar, D.K., Verkhivker, G.M., Reijto, P.A. et al. Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: Conformationally flexible docking by evolutionary programming. Chemistry & Biology 2:317–324, 1995.
15. Knegtel, R.M.A., Kuntz, I.D., Oshiro, C.M. Molecular docking to ensembles of protein structures. J. Mol. Biol. 266:424–440, 1997.
16. Leach, A.R., Kuntz, I.D. Conformational analysis of flexible ligands in macromolecular receptor sites. J. Computat. Chem. 13:730–748, 1992.
17. Rarey, M., Wefing, S., Lengauer, T. Placement of medium-sized molecular fragments into active sites of proteins. J. Comput. Aided Mol. Des. 10:41–54, 1996.
18. Rarey, M., Kramer, B., Lengauer, T., Klebe, G. A fast flexible docking method using an incremental construction algorithm. J. Mol. Biol. 261:470–489, 1996.
19. Strynadka, N.C.J., Eisenstein, M., Katchalski-Katzir, E. et al. Molecular docking programs successfully predict the

binding of a β-lactamase inhibitory protein to TEM-1 β-lactamase. Nat. Struct. Biol. 3:233–239, 1996.
20. Morris, G.M., Goodsell, D.S., Huey, R., Olson, A.J. Distributed automated docking of flexible ligands to proteins: Parallel applications of AutoDock 2.4. J. Comput. Aided Mol. Des. 10:293–304, 1996.
21. Leach, A.R. Ligand docking to proteins with discrete side-chain flexibility. J. Mol. Biol. 235:345–356, 1994.
22. Jones, G., Willett, P., Glen, R.C. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. J. Mol. Biol. 245:43–53, 1995.
23. Jones, G., Willett, P., Glen, R.C., Leach, A.R., Taylor, R. Development and validation of a genetic algorithm for flexible docking. J. Mol. Biol. 267:727–748, 1997.
24. Kuntz, I.D., Blaney, J.M., Oatley, S.J., Langridge, R., Ferrin, T.E. A geometric approach to macromolecule-ligand interactions. J. Mol. Biol. 161:269–288, 1982.
25. Shoichet, B.K., Kuntz, I.D. Matching chemistry and shape in molecular docking. Protein Eng. 6:723–732, 1993.
26. Welch, W., Ruppert, J., Jain, A.N. Hammerhead: fast, fully automated docking of flexible ligands to protein binding sites. Chemistry & Biology 3:449–462, 1996.
27. Ruppert, J., Welch, W., Jain, A.N. Automatic identification and representation of protein binding sites for molecular docking. Protein Sci. 6:524–533, 1997.
28. Fersht, A.R. The hydrogen bond in molecular recognition. Trends in Biochem. Sci. 12:301–304, 1987.
29. Janin, J., Chothia, C. The structure of protein-protein recognition sites. J. Biol. Chem. 265:16027–16030, 1990.
30. Meyer, M., Wilson, P., Schomburg, D. Hydrogen bonding and molecular surface shape complementary as a basis for protein docking. J. Mol. Biol. 264:199–210, 1996.
31. Mizutani, M.Y., Tomioka, N., Itai, A. Rational automatic search method for stable docking models of protein and ligand. J. Mol. Biol. 243:310–326, 1994.
32. Crippen, G.M., Havel, T.F. "Distance Geometry and Molecular Conformation." New York: John Wiley & Sons, 1988.
33. Kuhl, F.S., Crippen, G.M., Friesen, D.K. A combinatorial algorithm for calculating ligand binding. J. Comp. Chem. 5(1):24–34, 1984.
34. Smellie, A.S., Crippen, G.M., Richards, W.G. Fast drug-receptor mapping by site-directed distances: A novel method for predicting new pharmacological leads. J. Chem. Inf. Comput. Sci. 31:386–392, 1991.
35. Ghose, A.K., Crippen, G.M. Geometrically feasible binding modes of a flexible ligand molecule at the receptor site. J. Computat. Chem. 6:350–359, 1985.
36. Sheridan, R.P., Venkataraghavan, R. Designing novel nicotinic agonists by searching a database of molecular shapes. J. Comput. Aided Mol. Des. 1:243–256, 1987.
37. DesJarlais, R.L., Sheridan, R.P., Seibel, G.L., Dixon, J.S., Kuntz, I.D. Using shape complementarity as an initial screen in designing ligands for a receptor binding site of known three-dimensional structure. J. Med. Chem. 31:722–729, 1988.
38. Lawrence, M.C., Davis, P.C. CLIX: A search algorithm for finding novel ligands capable of binding proteins of known three-dimensional structure. Proteins 12:31–41, 1992.
39. Gunner, O.F., Hughes, D.W., Dumont, L.M. An integrated approach to three-dimensional information management with MACCS-3D. J. Chem. Inf. Comput. Sci. 31:408–414, 1991.
40. Shoichet, B.K., Stroud, R.M., Santi, D.V., Kuntz, I.D., Perry, K.M. Structure-based discovery of inhibitors of thymidylate synthase. Science 259:1445–1450, 1993.
41. Gschwend, D.A., Sirawaraporn, W., Santi, D.V., Kuntz, I.D. Specificity in structure-based drug design: Identification of a novel, selective inhibitor of Pneumocystis carinii dihydrofolate reductase. Proteins 29:59–67, 1997.
42. Makino, S., Kuntz, I.D. Automated flexible ligand docking method and its application for database search. J. Computat. Chem. 18(14):1812–1825, 1997.
43. Böhm, H.-J. On the use of LUDI to search the Fine Chemicals Directory for ligands of proteins of known three-dimensional structure. J. Comput. Aided Mol. Des. 8:623–632, 1994.
44. Böhm, H.-J. The computer program LUDI: A new method for the de novo design of enzyme inhibitors. J. Comput. Aided Mol. Des. 6:61–78, 1992.
45. Jain, A.N. Scoring noncovalent protein-ligand interactions: A continuous differentiable function tuned to compute binding affinities. J. Comput. Aided Mol. Des. 10:427–440, 1996.
46. Hobohm, U., Sander, C. Enlarged representative set of protein structures. Protein Sci. 3:522–524, 1994.
47. Abola, E.E., Bernstein, F.C., Bryant, S.H., Koetzle, T.F., Weng, J. Protein data bank. In: "Crystallographic Databases—Information Content, Software Systems, Scientific Applications," Allen, F.H., Bergerhoff, G., Sievers, R. (eds.). Bonn/Cambridge/Chester: Data Commission of the International Union of Crystallography. 107–132, 1987.
48. Kuhn, L.A., Swanson, C.A., Pique, M.E., Tainer, J.A., Getzoff, E.D. Atomic and residue hydrophilicity in the context of folded protein structures. Proteins 23:536–547, 1995.
49. Hooft, R.W.W., Sander, C., Vriend, G. Positioning hydrogen atoms by optimizing hydrogen-bond networks in protein structures. Proteins 26:363–376, 1996.
50. Habermann, S.M., Murphy, K.P. Energetics of hydrogen bonding in proteins: A model compound study. Protein Sci. 5:1229–1239, 1996.
51. Böhm, H.-J. The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. J. Comput. Aided Mol. Des. 8:243–256, 1994.
52. Mol, C.D., Arvai, A.S., Sanderson, R.J. et al. Crystal structure of human uracil-DNA glycosylase in complex with a protein inhibitor: Protein mimicry of DNA. Cell 82:701–708, 1995.
53. Mol, C.D., Arvai, A.S., Slupphaug, G. et al. Crystal structure and mutational analysis of human uracil-DNA glycosylase: Structural basis for specificity and catalysis. Cell 80:869–878, 1995.
54. Lunney, E.A. Structure-based design and two aspartic proteases. Network Science 1:http://www.netsci.org/Science/Cheminform/feature01.html, 1995.
55. Hofmann, B.E., Bender, H., Schulz, G.E. Three-dimensional structure of cyclodextrin glycosyltransferase from Bacillus circulans at 3.4 Å resolution. J. Mol. Biol. 209:793–800, 1989.
56. Collawn, J.F., Kuhn, L.A., Liu, L.-F.S., Tainer, J.A., Trowbridge, I.S. Transplanted LDL and mannose-6-phosphate receptor internalization signals promote high-efficiency endocytosis of the transferrin receptor. EMBO J. 10:3247–3253, 1991.
57. Craig, L., Sanschagrin, P.C., Rozek, A., Lackie, S., Kuhn, L.A., Scott, J.K. The role of structure in antibody cross-reactivity between peptides and folded proteins. J. Mol. Biol., 281:in press.
58. Morris, G.M., Goodsell, D.S., Halliday, R.S. et al. AutoDock 3.0: Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. J. Computat. Chem., in press.
59. Raymer, M.L., Sanschagrin, P.C., Punch, W.F., Venkataraman, S., Goodman, E.D., Kuhn, L.A. Predicting conserved water-mediated and polar ligand interactions in proteins using a k-nearest-neighbors genetic algorithm. J. Mol. Biol. 265:445–464, 1997.
60. Edmundson, A.B., Harris, D.L., Fan, Z.-C. et al. Principles and pitfalls in designing site-directed peptide ligands. Proteins 16:246–267, 1993.
61. Abagyan, R., Totrov, M., Kuznetsov, D. ICM- A new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation. J. Computat. Chem. 15:488–506, 1994.
62. Abagyan, R., Batalov, S., Cardozo, T., Totrov, M., Webber, J., Zhou, Y. Homology modeling with internal coordinate mechanics: Deformation zone mapping and improvements of models via conformational search. Proteins Suppl. 1:29–37, 1997.