In "Rigidity in Theory and Applications"
M. F. Thorpe and P. M. Duxbury (eds.)
Plenum Publishing, New York, 1999, 385–400

FLEXIBLY SCREENING FOR MOLECULES INTERACTING WITH PROTEINS

Volker Schnecke and Leslie A. Kuhn

Protein Structural Analysis and Design Laboratory
Department of Biochemistry
Michigan State University
East Lansing, MI 48824-1319, U.S.A.
http://www.bch.msu.edu/labs/kuhn

INTRODUCTION

Flexibility in Proteins and Their Interactions

The flexibility of proteins and their *ligands* (molecules specifically bound by proteins) has a major influence on the ways they interact. Generally, protein molecules are thought of as primarily rigid structures, with chemically specific and somewhat flexible side chains attached to a main chain of fixed structure. The tendency to think of proteins as rigid is reinforced by the fact that X-ray crystallography, the most widely-used technique for analyzing protein structures at atomic resolution, traps the copies of a protein molecule in the crystalline lattice into a single state. However, as seen in Figure 1, rotatable single bonds in the main as well as side chains of a protein provide significant potential for flexibility (*conformational change*). For a number of proteins, such as the HIV protease, lysine-arginine-ornithine binding protein, and adenylate kinase, the protein is known to undergo significant conformational change upon binding its natural ligand or drugs designed to inhibit its activity. Flexibility is thus a biologically essential feature of proteins.

Despite the importance and widespread interest in characterizing protein flexibility, this remains a challenge both experimentally and computationally (see papers by David Case, Ruben Abagyan, and Mark Gerstein in this volume). Our laboratory's goal has been to develop computational methods that incorporate realistic modeling of protein flexibility into the design of new ligands for proteins. In collaboration with Jacobs and Thorpe (see accompanying paper), we have shown that graph-theoretic analysis of the covalent and hydrogen-bond networks in proteins using the FIRST algorithm provides an extremely fast way of assessing large-scale flexibility in proteins, e.g., when large, independently folded regions of the protein (*domains*) are attached by hinge joints, resulting in clamshell-like motion. In the present paper, we review the state of the art in template-based algorithms for analyzing protein–

ligand interactions, which reduce the orientational search for the optimal placement of the ligand relative to the protein, and discuss how ligand flexibility is modeled in some of these methods. Finally, we address how side-chain motion in the protein, coupled with flexibility in the ligand, can be modeled in an algorithm that allows fast and effective computational screening of hundreds of thousands of compounds for ligands. This method, Specitope, is the first to model protein, as well as ligand, flexibility, in the context of screening. The identification of favorable ligand candidates is a first and crucial step in modern drug design, with major potential to develop new therapeutics for AIDS, cancer, bacterial infection, arthritis, and other diseases.

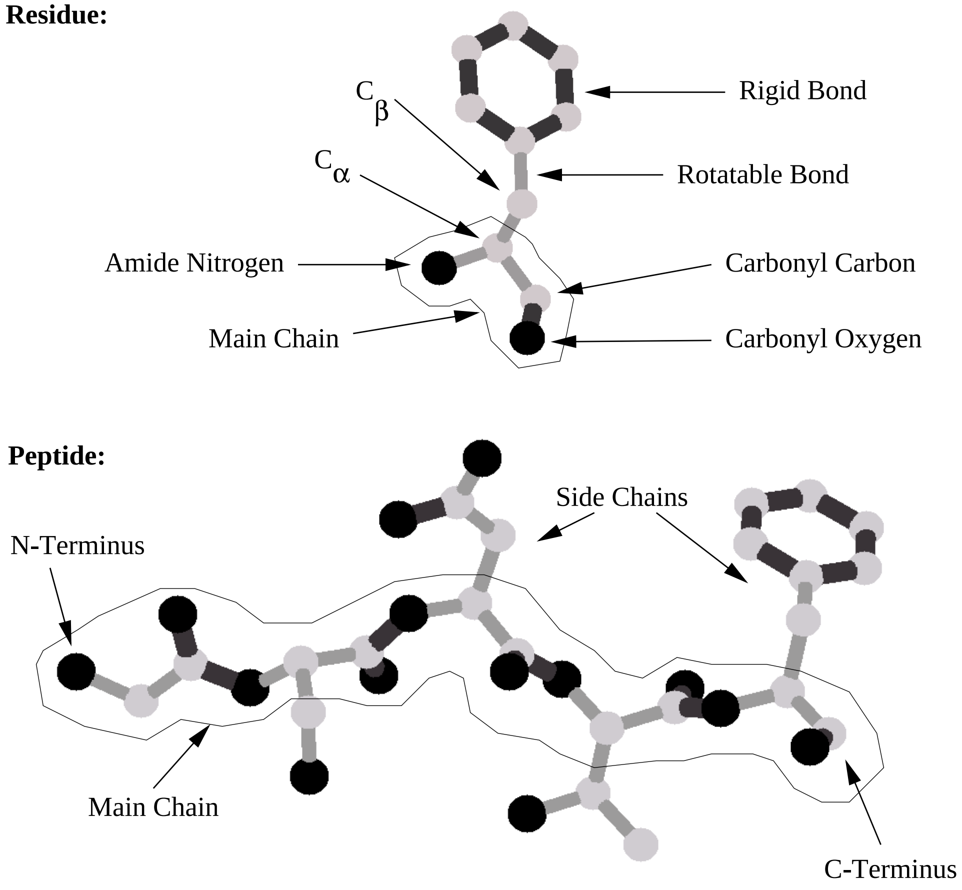

Figure 1. *Protein structure and intrinsic flexibility.* The basic repeating unit within a protein, a *residue*, is shown at top. The side chain of the residue in this case (phenylalanine) consists of a phenyl ring (at top) connected to the protein main chain (polymer *backbone*) at the alpha-carbon. Rigid, unrotatable bonds with partial or full double-bond character are shown by black tubes, and rotatable single bonds are shown as thin grey tubes; carbon atoms appear in light grey and polar (oxygen or nitrogen) atoms in black. In the bottom panel, a *peptide* (fragment of protein) is shown, consisting of several linked residues with various side chains. The main chain of the polymer, formed by the repeating (amide nitrogen)–(alpha-carbon)–(carbonyl carbon and oxygen) motif, is outlined. The N-terminus is the start of the peptide chain, and the C-terminus is its end. In reality, not all of the flexibility implied by the single bonds is accessible, since van der Waals, hydrogen-bond, and electrostatic interactions typically lock the protein chain into a unique *conformation* (molecular shape) or set of conformations dependent on its sequence of amino acid residues.

Ligand Docking and Database Screening

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* (ligands that block the function of proteins)[1–4]. Although computational methods do not replace the *in vitro* and *in vivo* tests during drug development, they can be very efficient in identifying and optimizing the structures of ligand candidates as *lead* compounds for further development, and thus accelerate the early design stages. Generally, there are two tasks involved in identifying leads in computational drug design, *screening* the structures of compounds for potential ligands, and *docking*, or optimally fitting, these potential inhibitors into the binding site of the target (typically a protein). Screening is important for reducing the vast number of potential ligands to an experimentally testable number. Furthermore, developing methods to predict protein–ligand interactions – in particular, predicting ligands and their *binding modes*, or specific orientation and conformation upon interaction with the protein – provides significant insights into the way proteins work, which is always by interacting with other molecules. Docking combined with modeling structural modifications of the protein or ligand also provides a valuable design tool for developing new ligands with greater specificity for a given protein and *vice versa*.

Computational screening is often closely associated with docking, as the final evaluation of ligand candidates requires a detailed evaluation of how well the ligand and protein fit together. Docking approaches can be classified based on how they characterize the ligand-binding site of the protein. *Grid-search techniques* fill the space around the binding site with a three-dimensional grid, precompute the potentials (van der Waals, electrostatic, etc.) at each grid point without the ligand, then sample different ligand conformations and orientations on the grid and compute the resulting binding free energy. An example for this approach is AutoDock, which used simulated annealing in its previous releases[5,6], but now applies a hybrid genetic algorithm to sample over the feasible binding modes of the ligand relative to the protein[7,8]. The advantages of grid-based docking are that a template of favorable interactions in the ligand-binding site does not need to be defined, reducing bias in modeling the protein–ligand interactions, and evaluation of binding modes is made more efficient by precomputing protein potentials on the grid. However, the accuracy and timing of this approach depends on the grid fineness, making this approach too computationally intensive for database screening, in which thousands of molecules (as well as ligand orientations and shapes, or *conformers*) need to be tested. Furthermore, precomputation of the protein grid potentials limits this approach to rigid binding sites.

When the ligand-binding site in the protein is known, this can be utilized by constructing a *template* for ligand binding based upon favorable interaction points in the binding site. During the search for a favorable ligand-binding mode, different conformations of the ligand can be generated and subsets of its atoms matched to complementary template points, as a basis for docking the ligand into the binding site. Advantages to this template-based approach are that it can incorporate known features of ligand binding (for example, conserved interactions observed experimentally for known ligands), and it reduces the docking search space to matching N ligand atoms onto N template points, rather than the 6-dimensional orientational search space (3 degrees of rotational freedom and 3 degrees of translational freedom) required in other approaches for sampling and evaluating ligand binding.

In the well-known docking tool DOCK[9], the template typically consists of up to 100 spheres that generate a negative image of the binding site. During the search, subsets of ligand atoms are matched to spheres, based on the distances between ligand atoms. DOCK has been extended to consider chemistry[10] and include hydrogen-bonding interaction centers[11] in addition to the shape template. Other current template approaches specify a set of interaction points defining favorable positions for placing polar ligand atoms or *hydrophobic* (nonpolar) centers, e.g., aromatic rings. Such a template can be generated automatically, e.g., by placing probe points on the solvent accessible surface of the binding site[12], or interactively by superimposing known protein–ligand complexes to identify favorable interaction points based on observed binding modes for known ligands. FlexX[13,14] uses a template of 400 to 800 points when docking drug-size molecules (up to 40 atoms, not including hydrogens) to define positions for favorable interactions of hydrogen-bond donors and acceptors, metal ions, aromatic rings, and methyl groups. The ligand is fragmented and incrementally constructed in the binding site and matched to template points based on geometric *hashing* (indexing) techniques, bond torsional flexibility is modeled discretely, and a tree-search algorithm is used to keep the most promising partially constructed ligand conformations during the search. Hammerhead[15] uses up to 300 hydrogen-bond donor and acceptor and *steric* (van der Waals interaction) points to define the template, and the ligand is incrementally constructed, as in FlexX. A fragment is docked based on matching ligand atoms and template points with compatible internal distances, similar to the matching algorithm used in DOCK. If a new fragment is positioned closely enough to the partially constructed ligand, the two parts are merged, and the most promising placements kept. Other successful docking approaches, such as GOLD[16,17] and the method of Oshiro *et al.*[18], use genetic algorithms to sample over possible matchings of conformationally flexible ligands to the template. However, a drawback of genetic algorithm approaches, including AutoDock, is the high computation time, especially in comparison to fragment-based docking approaches.

When screening databases of more than 10^5 compounds to identify potential ligands, the computational efficiency of the search process becomes a significant concern. Docking a small flexible molecule with high accuracy takes at least several minutes on a desktop workstation for the fastest of the recent algorithms[11,13,14,19–21]. Spending only one minute to dock each molecule when screening a dataset of 100,000 compounds results in a computation time on the order of two months, which is unacceptably slow, particularly when improving and validating the method. Recent screening tools can identify potential ligands from up to 150,000 compounds within a few days when ligand flexibility is modeled[15,22–26]; however, none of these methods models protein conformational change upon ligand binding (also called *induced complementarity*).

For a number of structurally characterized protein–ligand complexes, induced complementarity of protein side chains is known to be important for ligand binding[27]. Thus, for the development of a new screening procedure, SPECITOPE, our goal has been to model protein side-chain flexibility as well as ligand flexibility when evaluating their interaction. SPECITOPE narrows down the vast number of ligand candidates to several dozen molecules with good shape and chemical complementarity to a protein ligand-binding site within 3 hours on a typical desktop workstation[28]. It is difficult to make direct comparisons between the timing of SPECITOPE and other screening algorithms, because the other methods assume the protein is rigid, differ in their modeling of ligand flexibility, and some require manual scoring (molecular graphics assessment by a structural biologist) outside the algorithm. However, methods screening rigid ligands typically take several hours, whereas methods modeling full ligand flexibility typically take several days, plus time spent for external scoring.

SPECITOPE's relative speed results from adapting distance geometry techniques[29] (see also the paper by Havel in this volume) to perform quick feasibility checks on each ligand based on comparing its interatomic distances and number of hydrogen-bond donors and acceptors with those in a template representing the binding site. We have shown that protein and ligand side-chain flexibility can be modeled while screening a large database of peptide structures for inhibitors to three diverse proteins, an aspartyl proteinase, a serine protease, and a DNA repair enzyme[28]. This approach was successful in identifying the known peptidyl inhibitors within the top five of 140,000 ligand candidates screened, and for two of the three proteins, the known ligand received the top score, based on shape complementarity and favorable hydrophobic and hydrogen-bond interactions with the protein. In each case, pro-

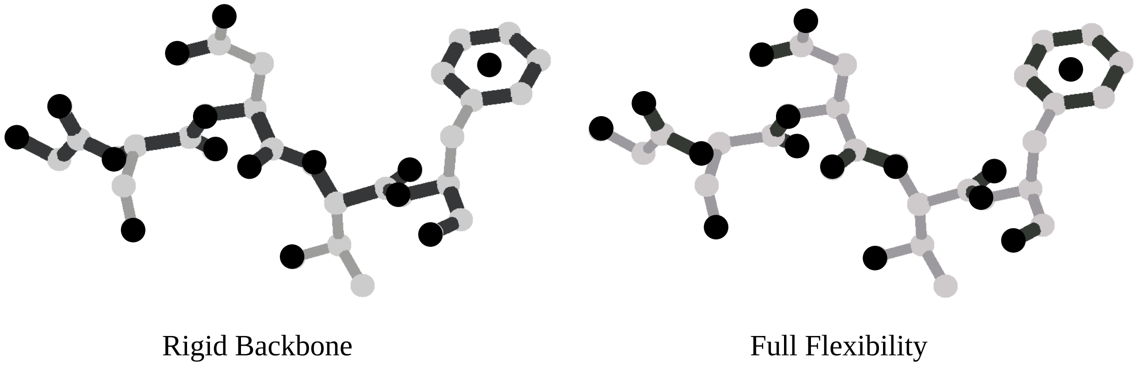

Figure 2. *Side-chain versus full flexibility in a peptide.* For the same peptide shown in Figure 1, two representations of peptide flexibility used by screening algorithms are shown: a model in which the backbone (main chain) is a rigid unit and the side chains are free to rotate, and a fully flexible model, where the backbone and side-chain dihedral angles are free to rotate. Rigid bonds are shown as black tubes and flexible bonds as grey tubes; in the full-flexibility model, rigid bonds arise from the partial-double and double bonds inherent in the structure. Black spheres indicate polar (nitrogen, oxygen) atoms to be compared with a hydrogen-bonding template representing the ligand-binding site, whereas the black sphere in the ring center (top right) indicates a hydrophobic ring center to be matched with hydrophobic centers in the template. Grey spheres indicate carbon atoms.

tein side-chain motion was known to be important for ligand binding and was appropriately modeled by SPECITOPE during docking.

Here we present and validate improvements of SPECITOPE to probe the ligand-binding site more thoroughly; this approach will also enable screening of fully flexible ligands in the near future. Geometric hashing techniques are employed so that the exhaustive checking of different matchings of the ligand to the protein's ligand-binding site is reduced to checking only those matchings that are feasible, based on distance and chemistry indices stored for the template in a look-up table. This provides time savings and linear scaleability, allowing sampling over more template points within the binding site and modeling of full flexibility for peptidyl and small organic ligand candidates. Hydrophobic interaction centers are now considered, in addition to hydrogen-bonding centers, in the ligand and in the binding-site template. Figure 2 compares the difference between peptidyl ligand flexibility, as previously modeled by SPECITOPE, where side chains were flexible but main-chain dihedral angles were held fixed, with the degree of flexibility that will be enabled by the hashing approach, where both side-chain and main-chain dihedrals are rotatable (within the limitations imposed by van der Waals contacts). Many organic compounds are not polymeric and thus do not have a clear main chain/side chain distinction, though they often have rigid frameworks provided by ring systems. To illustrate how the same concepts applied here to peptides can be applied to more general organic compounds, a goal for our future ligand design work, Figure 3 shows the interaction centers and flexible bonds for two compounds arbitrarily chosen from crystallographic structures in the Cambridge Structural Database (http://www.ccdc.cam.ac.uk).

METHODS

SPECITOPE shares with other current docking and screening approaches the use of a binding site template to limit the orientational search for each prospective ligand, and differs in the use of distance geometry techniques to avoid the computationally intensive fitting of infeasible ligands into the binding site. The speed gained by distance geometry allows the sec-

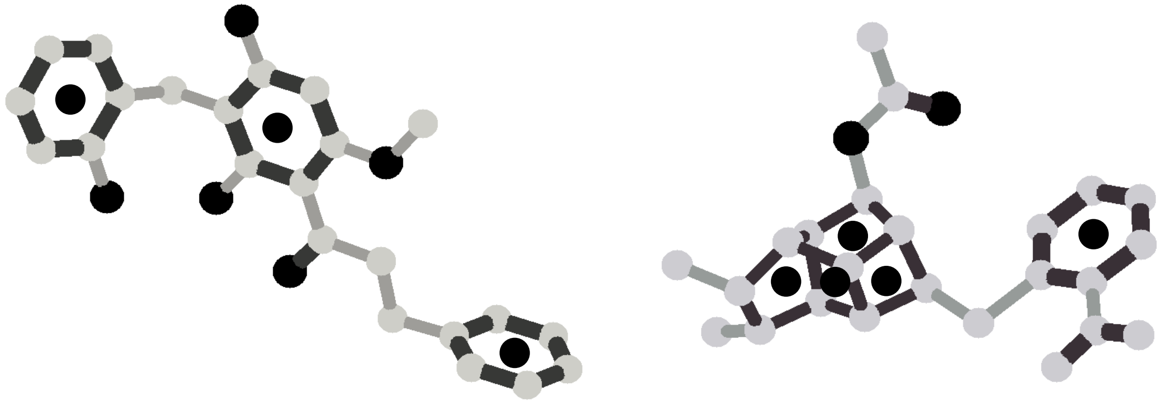

Figure 3. *Flexibility in organic compounds.* Two compounds from the Cambridge Structural Database are shown, indicating how the representation of rigid and flexible bonds and polar and hydrophobic (nonpolar) atom centers carries from peptides to more general compounds (bonds and atoms colored as in previous figure). Notice the similar degree of flexibility of the compound at left to the fully flexible peptide (previous figure), whereas the compound at right has significantly more rigid and hydrophobic clusters than are found in peptides.

ond advantage over other screening methods, modeling protein side-chain flexibility during docking. Here we overview the algorithm and present the use of hashing techniques adapted from fragment-based docking to allow sampling over more interaction sites as well as future modeling of full ligand flexibility. We now include hydrophobic centers for template matching, in addition to the hydrogen-bonding centers previously used. While hydrogen bonds are important for providing specificity to most protein–ligand interactions, hydrophobic interactions are especially significant for some organic ligands, and contribute favorably to the binding free energy[17,30,31]. The main steps of SPECITOPE, as described below, are: template design; distance geometry and hashing steps for screening out geometrically infeasible ligand candidates and efficiently matching ligands to the template; rigid-body translations of the ligand coupled with ligand and protein side-chain flexibility to resolve steric overlaps in the complex; and ligand scoring based on interactions with the protein.

Template Design

For SPECITOPE, the template consists of key interaction points (hydrogen-bond donors, hydrogen-bond acceptors, and hydrophobic centers) where ligand atoms with matching character can make favorable interactions with the protein. Design of the template can be based on observed interactions in the structure of a known ligand in complex with the protein; for two of our recent applications, uracil-DNA glycosylase and cyclodextrin glycosyltransferase, templates were based on the positions of polar atoms in the ligand that formed hydrogen bonds to protein atoms[28] (see Figure 4). When structures are available for the protein in complex with several different ligands, as for aspartyl proteinase, interactions shared by the ligands can be used to develop a consensus template. When ligand-bound protein structures are unavailable, templates can be based on the positions of crystallographically observed water molecules bound in the ligand-binding site of the protein, representing favorable sites for making hydrogen bonds to the protein[32]; this approach was taken for screening subtilisin inhibitors[28]. Template-design methods from other docking and screening approaches (discussed in the Introduction) may also be employed.

In the prior version of SPECITOPE, the template was limited to about five interaction points, due to the combinatorics of sampling every possible matching of five ligand atoms

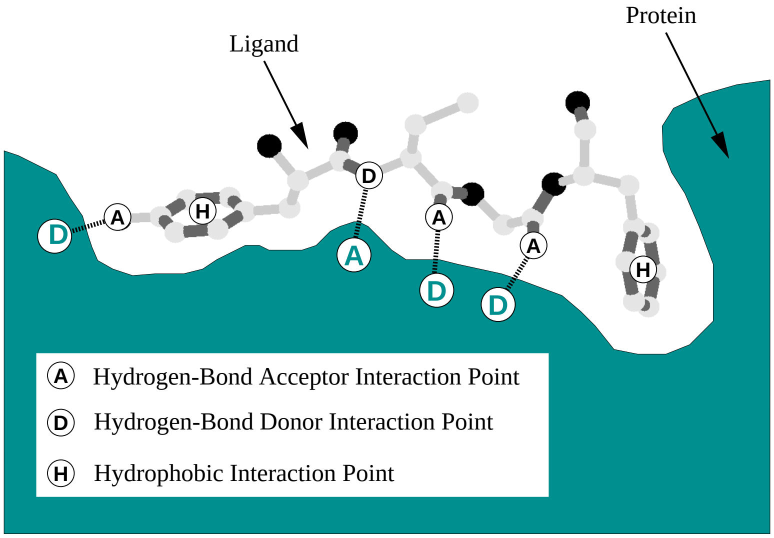
Protein
Ligand
D
A
H
D
A
A
D
D
A
H
A Hydrogen-Bond Acceptor Interaction Point
D Hydrogen-Bond Donor Interaction Point
H Hydrophobic Interaction Point

Figure 4. *Example of a binding site template.* The template is comprised of sites above the protein circled in black and labeled as hydrophobic or hydrogen-bond interaction points. This template represents sites where similarly-labeled ligand atoms can make favorable interactions with the protein (whose hydrogen-bond acceptor and donor sites are indicated below by grey letters).

(out of a larger number, typically ~15 for five-residue peptides) onto five template points; the computational complexity of this step is factorial, due to enumerating all permutations (matchings) of the ligand atoms onto template points (see equation below). In the present version, we avoid this complexity by considering 3-point subsets of larger (10 or more point) templates and using *triangle hashing* to evaluate only the geometrically feasible subsets of these triangles. As seen in Table 1, the number of potential matchings of 3-point ligand atom subsets onto a template with T points scales linearly as T increases, whereas the previously-used *complete enumeration*, in which all N-point subsets of the ligand are permuted onto the T-point template, scales poorly in both N and T. The scaling is linear in practice because the number of feasible triangles in each bin in the hash table (see Distance Geometry, Hashing, and Docking section below) has been found experimentally to scale linearly in T, effectively reducing the first term in the product below from the binomial coefficient, T-choose-N, to T.

$$\left(\begin{array}{c}\text{\# of potential}\\ \text{matchings}\end{array}\right) = \left(\begin{array}{c}\text{\# of ways of}\\ \text{choosing}\\ N \text{ template pts}\\ \text{from } T \text{ pts}\end{array}\right) \cdot \left(\begin{array}{c}\text{\# of ways of}\\ \text{choosing}\\ N \text{ ligand pts}\\ \text{from } L \text{ pts}\end{array}\right) \cdot \left(\begin{array}{c}\text{\# of ways of}\\ \text{matching}\\ N \text{ ligand pts to}\\ N \text{ template pts}\end{array}\right)$$

$$= \left(\frac{T!}{N!(T-N)!}\right) \cdot \left(\frac{L!}{N!(L-N)!}\right) \cdot (N!)$$

Matching triangles of ligand interaction centers onto triangular subsets of the template makes sense for two other reasons. The previous matching of four or five interaction centers in a ligand to the same number of template points in the binding site effectively rigidified the ligand, since most bonds between those atoms in the ligand could not be rotated without

Table 1. *Combinatorics of complete enumeration versus hashing approaches for matching N ligand points to T template points.*

	T	N	Number of Potential Matchings
Complete Enumeration	5	5	360,360
	10	5	90,810,720
	10	3	327,600
	20	3	3,112,200
Triangle Hashing	5	3	13,650
	10	3	27,300
	20	3	54,600
	50	3	136,500

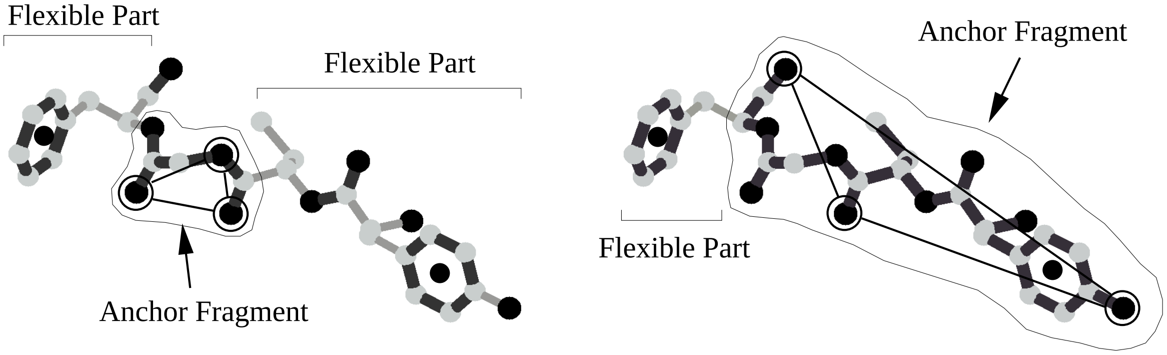

Figure 5. *The rigid and flexible regions in a ligand, using triangle-based docking.* During the template matching described below, every possible triangle of hydrogen-bonding and hydrophobic interaction points in the ligand is matched to every possible triangle of template points; however, the hashing procedure focuses directly on those matchings with feasible geometry and chemistry. The resulting triangle-based docking essentially rigidifies the triangular *anchor fragments*, which may be small (as in the left panel) or large (right panel), while maintaining the flexibility of the other parts of the ligand. Docking is based on matching the ligand triangle to the template and adjusting the flexible parts of the ligand and/or protein to remove intra- or intermolecular van der Waals collisions (overlaps between atoms). Then, all non-colliding ligand dockings are scored according to their hydrogen-bond and hydrophobic complementarity with the protein.

disrupting the template match. When, alternatively, all chemically and geometrically feasible ligand triangles are tested for docking to the protein via hashing, much smaller rigid fragments of the ligand are also tested (Figure 5). Furthermore, organic compounds, a major future application, tend to have fewer hydrogen-bonding atoms than peptides and are more chemically diverse. Thus, it is useful to be able to screen and dock these compounds based on fewer interaction points (3, which still uniquely define an orientation with respect to the protein), include hydrophobic as well as hydrogen-bond interaction centers, and consider shape complementarity as another major contributor to specificity. For compounds with more than 3 interaction points, each triangle can be docked independently, by optimizing the rotations in the linkages between them. An interface to read organic compounds and identify their flexible bonds, hydrophobic centers (interpreted as carbon-ring centers), and hydrogen-bond donors and acceptors has been developed in our lab. This interface is based on the generic mol2 molecular data file (of the Tripos Sybyl software) frequently used with other databases and modeling tools, making Specitope portable and compatible with other systems.

Distance Geometry, Hashing, and Docking

SPECITOPE first uses simple distance geometry[29] techniques to screen out ligands with incompatible geometry relative to a template specifying the positions of hydrophobic and hydrogen-bond interaction centers. Many of the ligand–template matchings can be ruled out by distance geometry alone, based on the incompatibility between ligand interatomic distances and inter-template-point distances. Given a set of N (in this case, 3) ligand interaction centers, a sorted list of the $N \cdot (N-1)/2$ distances, l_i, between ligand interaction centers is compared to the sorted list, t_i, of distances between the N 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, is defined as:

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

The $RMSD_{list}$ gives a measure for the compatibility of distances between the ligand atoms and between points in the template. 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 matched distances in the L and T matrices containing the distances between ligand interaction centers and distances between template points, respectively. 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 interaction centers and template points, this set can be ruled out, since the DME for any one-to-one matching of these centers 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.

Aside from the major time savings from comparing intra-template and intra-ligand distances via sorted distance lists rather than by docking the molecules together, SPECITOPE screens out infeasible ligands by a series of quick, initial distance and chemistry checks: Does the longest distance between interaction centers in the ligand significantly exceed the longest distance in the template? If so, they cannot match. Are there enough hydrogen-bond and hydrophobic interaction points in the ligand to match the template? If not, they do not match. Overall, these distance geometry and chemistry checks, through the DME step, typically rule out 70% of the infeasible ligand candidates before the time-intensive docking steps[28].

Figure 6 compares the current, hashing-based strategy used by SPECITOPE with the steps used in the previous complete-enumeration screening[28]. The distance geometry and docking (least-squares fit) steps remain the same, with the major difference being that triangles of ligand interaction centers are matched by hashing to the triangles of template points. The use of triangles, rather than more complex geometric objects, provides a convenient basis for screening ligand–template matches based on simple chemical and geometric characteristics of the triangles. This is done in three stages of hashing, or indexing, and is extremely efficient because a ligand triangle is only compared to template triangles with similar characteristics. As shown in Figure 7, the hash table allows direct access to all template triangles having

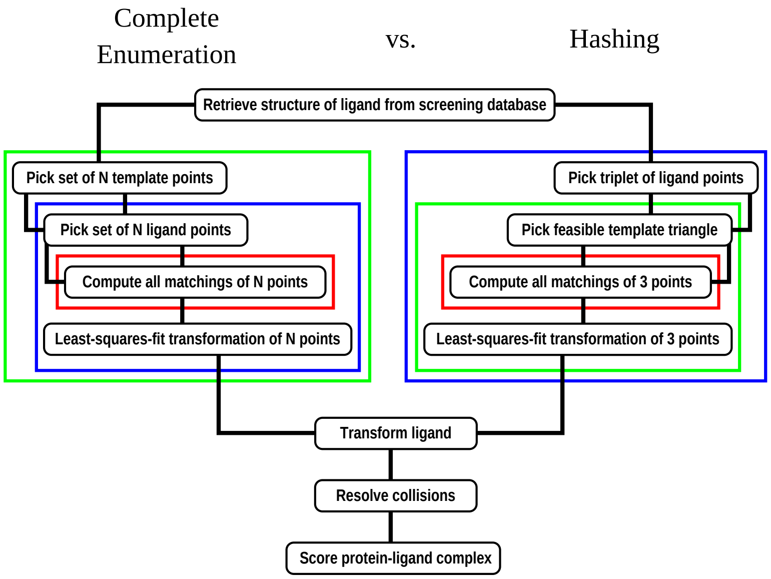
Complete Enumeration
vs.
Hashing
Retrieve structure of ligand from screening database
Pick set of N template points
Pick set of N ligand points
Compute all matchings of N points
Least-squares-fit transformation of N points
Pick triplet of ligand points
Pick feasible template triangle
Compute all matchings of 3 points
Least-squares-fit transformation of 3 points
Transform ligand
Resolve collisions
Score protein-ligand complex

Figure 6. *Comparison of the steps in hashing versus complete enumeration screening of ligands by* SPECITOPE.

the same number and type of interaction points as the ligand triangle (e.g., one hydrogen-bond acceptor and two donors, or ADD), a similar perimeter, and a similar length for the longest triangle side. The distance geometry and hashing steps ensure that the relatively time-intensive DME and docking (least-squares fit) calculations are only done for feasible template triangles. The $RMSD_{list}$ and DME are then calculated for each matching of triangles, using a cutoff of 0.2 Å, reflecting a much closer match than could be required for matching a larger number of points. This close match to the template also preserves the possibility of hydrogen bonding and was found to retain known ligands during screening. Docking involves taking the minimal-DME matching between the current ligand interaction centers and the template, then transforming the ligand into the protein's ligand-binding site based on a least-squares fit of these matched points. If the root-mean-square deviation of this transformation is below a fixed threshold (0.2 Å), the entire ligand is transformed into the ligand-binding site.

Van der Waals Collisions and Flexibility

This ligand transformation into the protein's ligand-binding site results in a close fit of the ligand interaction points with the template, but also may result in van der Waals collisions between ligand and protein atoms. Because of the flexibility of the protein and ligand, many such collisions are ultimately resolvable. In its current incarnation, halfway to full ligand flexibility, SPECITOPE employs triangle hashing for matching the ligand and template, while modeling the peptidyl ligand backbone as rigid and the ligand and protein side chains as flexible. The following discussion also covers the more general case in which the ligand backbone is flexible, which triangle hashing will make possible. If the collisions are between the rigid part of the ligand (backbone or anchor fragment) and the rigid backbone of the protein, the minimal translation vector for resolving these collisions is calculated, and the ligand is translated accordingly. If new rigid-body collisions result, this procedure is iterated up to 100 times, effectively shaking the ligand inside the ligand-binding site. If these collisions cannot be resolved, this ligand matching to the template is discarded. When the collisions can be resolved, then van der Waals overlaps involving flexible parts of the ligand and the

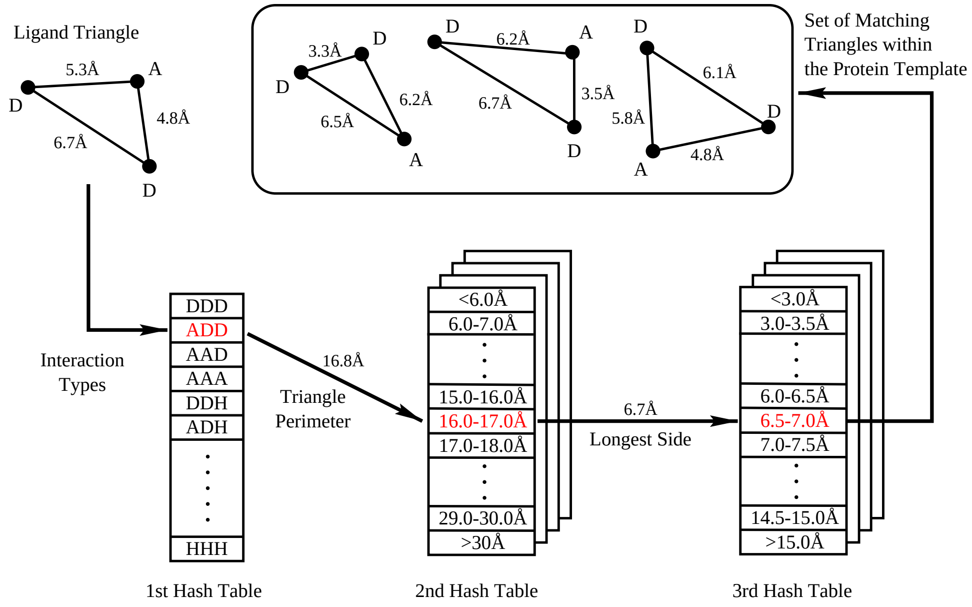

Figure 7. *Three-stage hashing applied to ligand–template matching.* Hashing enables direct access to those sets of template triangles (shown at top center, with length and interaction-type data precomputed and exhaustively listed in a table) that match a given set of three ligand points (upper left). The index into the first table is based on the interaction types of the ligand atoms (in this case, two hydrogen-bond donors and one acceptor), which points to the subset of template triangles with the same label (ADD). The second index, the perimeter of the ligand triangle, locates those template triangles that are labeled ADD and have a similar perimeter. The third index, the length of the longest side of the triangle, points to those ADD template triangles with similar perimeter and longest side. This results in more efficient, but still exhaustive, checking of ligand–template matches.

protein are addressed.

Each flexible part in the ligand is checked for overlaps with protein atoms, which are cleared by rotating this part of the ligand through the minimal angle that resolves the overlaps. The single bond closest to the colliding atoms in the ligand is used first to resolve the overlap. If a collision-free conformation cannot be generated with this rotation, the next rotatable bond closer to the rigid part of the ligand is rotated. When it is not possible to resolve an overlap by rotations within the ligand, the same approach is applied to the protein side chain involved in the collision. If an intermolecular collision remains, despite testing all protein side-chain and ligand 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 flexible group is rotated back to its original conformation, and the next single bond closer to the backbone or anchor fragment 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. The aim of this step in SPECITOPE is not to predict the optimal ligand conformation, but to ensure that a collision-free conformation of the molecules exists for this matching.

Scoring

A scoring function is then used to rank the relative complementarity of the ∼100 ligand candidates (passing the previous screening steps) to the protein's ligand-binding site. Because the conformation and orientation of each ligand candidate could likely be optimized by fine docking, scoring is mainly intended to recognize molecules that lack chemical complementarity to the protein and to emphasize molecules that fit well in the given binding mode. The scoring function weighs the dominant factors, the number of hydrogen bonds between the protein and ligand and their hydrophobic complementarity. For hydrogen-bond donors and acceptors separated by 2.8 to 3.5 Å, SPECITOPE computes the optimal position of the shared hydrogen atom (because most X-ray structures do not provide hydrogen atom positions) to identify intermolecular hydrogen bonds with good geometry. The hydrophobicity measure is based on a statistical survey of atomic hydration in 56 protein structures[33] and compares the hydrophobicity value of each ligand atom to the average hydrophobicity value of nearby protein atoms. The overall complementarity of the protein-ligand complex, SCORE (protein,ligand), is given by a weighted sum of the number of hydrogen bonds and the hydrophobic complementarity:

$$\mathrm{SCORE}(\mathrm{protein},\mathrm{ligand}) = \mathrm{A}\cdot\mathrm{HBONDS}(\mathrm{protein},\mathrm{ligand}) + \mathrm{B}\cdot\mathrm{HPHOB}(\mathrm{protein},\mathrm{ligand})$$

Based on the functions of Böhm[34] and Jain[35], 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, with the weights A and B tuned accordingly, based on the values of HBOND and HPHOB from 30 structures of protein complexes with small peptidyl ligands[28] from the Protein Data Bank (PDB)[36]. These structures have 6.3 hydrogen bonds between protein and ligand, on average.

RESULTS

Given the new implementation of triangle hashing, our goal here is to assess whether hashing identifies ligands with similar or better protein complementarity relative to those identified by complete enumeration.

Screening for Aspartic Proteinase Ligands

To compare ligands from the hashing and complete enumeration approaches, SPECITOPE was used to screen for peptidyl ligands of rhizopuspepsin, an aspartic proteinase. Rhizopuspepsin is a relative of medically important inhibitor design targets including renin, which regulates blood pressure, and HIV protease, which is essential to the virus life cycle and a major target for AIDS drug design[37]. The template for rhizopuspepsin was designed by superimposing three complexes of this protein with pepstatin (PDB structure 6apr) or pepstatin-like renin inhibitors (4apr, 5apr) onto the ligand-free structure of rhizopuspepsin (2apr). Figure 8 shows the structure of rhizopuspepsin bound to a known peptidyl ligand.

For complete enumeration screening, the average positions of four hydrogen-bond donors and one acceptor in the three inhibitors were selected as the template points. For hashing, 9 such template points were chosen, and the characteristics of each 3-point subset of these template points were listed in the hashing tables (see Figure 7). In both cases, all 5-residue peptides were screened from 140,000 peptides occurring in known protein structures with low similarity (<25% sequence identity)[38].

Complete enumeration identified 117 possible peptidyl ligands for rhizopuspepsin in 96 minutes on a desktop workstation (Sun SPARC Ultra 140), whereas hashing identified 357 possible ligands in 75 minutes. For this case, the 10-fold or more computational savings possible from hashing were not realized because more peptides passed the distance geometry steps; however, for another protein tested, human uracil-DNA glycosylase, screening was 3 times as fast. More importantly, the peptidyl ligands identified by hashing tended to have higher complementarity to the protein than the ligands identified by complete enumeration (Table 2). The feasible ligands from triangle hashing had an average complementarity score

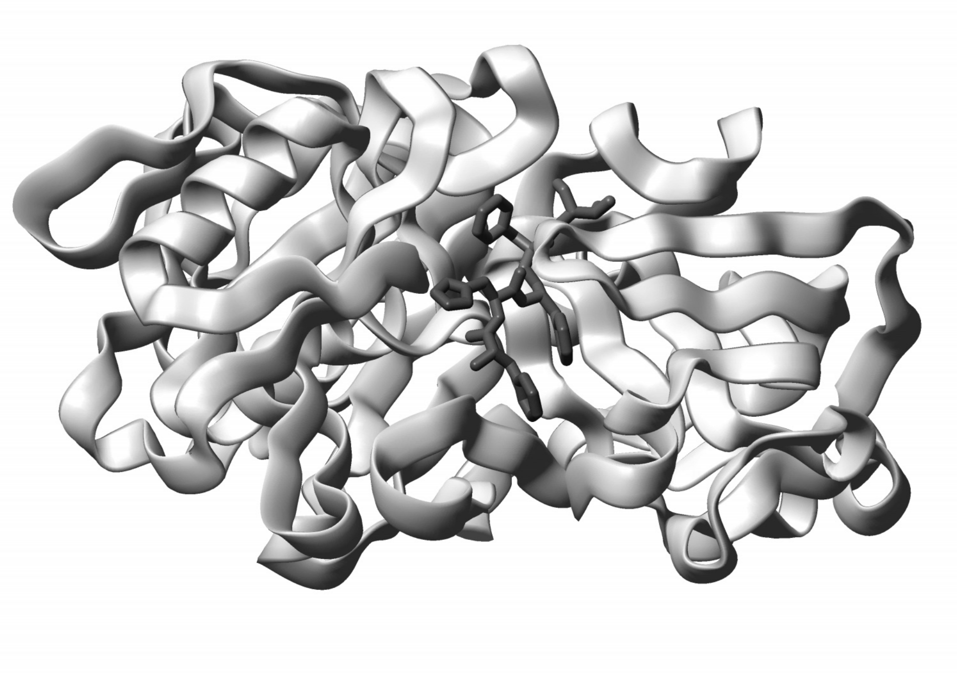

Figure 8. *Rhizopuspepsin in complex with a peptidyl ligand.* The rhizopuspepsin backbone is shown as a grey ribbon, with its ligand-binding site forming the vertical cleft at center. Bound in this site is the known peptidyl ligand, FHFFV, shown in black tubes. The crystallographic structure of the protein-ligand complex is from PDB entry 3apr.

of 1369.8 and an average number of 3.5 hydrogen bonds between protein and ligand, whereas complete enumeration resulted in ligands with an average score of 816.8 and an average of 2.2 protein-ligand hydrogen bonds. The top-five ligands found by hashing had scores 200-300 points higher and typically had one more hydrogen bond than those found by complete enumeration. In both cases, a known rhizopuspepsin ligand, the peptide with amino-acid sequence FHFFV (PheHisPhePheVal), was identified as the top-scoring ligand, with similar scores for complete enumeration and hashing; this reflects similar dockings despite the different number of ligand atoms matched to template points, 5 for complete enumeration and 3 for hashing. The docking based on hashing also resulted in a very similar ligand-binding mode to that observed in the crystallographic structure of the rhizopuspepsin-peptide complex (Figure 9). The backbones of the ligand from the crystal structure and as docked by SPECITOPE (backbones running horizontally in Figure 9) are essentially superimposed, with some side-chain reorientation (in the rings at top).

CONCLUSIONS AND FUTURE DIRECTIONS

Triangle-based hashing has been implemented in our ligand screening algorithm, SPECITOPE, and provided time savings in ruling out infeasible ligand candidates, as well as more thoroughly sampling the ligand binding site. Screening for rhizopuspepsin ligands showed a 28% speed increase using hashing as compared with the previous complete enumeration approach, while finding ligands with higher complementarity to the protein. Our goal, nearing completion, is to screen fully flexible peptidyl and small organic ligands against proteins with side-chain flexibility. Hashing enables this by sampling over all possible rigid fragments of ligands during screening and docking. Recent construction of a structural database

Table 2. *Summary of the top-five rhizopuspepsin ligands identified by* Specitope *using complete enumeration and triangle hashing.*

Complete Enumeration (5 template points)				**Triangle Hashing (9 template points)**			
Rank	Sequence	H-Bonds	Score	Rank	Sequence	H-Bonds	Score
1	`FHFFV`	5	3178.8	1	`FHFFV`	5	3324.1
2	`KTVTD`	2	3150.6	2	`YYTAL`	4	3280.3
3	`ETTSF`	2	3067.2	3	`NLKFG`	3	2982.4
4	`LWCNG`	3	2573.0	4	`LYIDS`	3	2797.8
5	`YGLSV`	3	2385.0	5	`GYYTA`	4	2770.4
1—117	Average	2.2	816.8	1—357	Average	3.5	1369.8
	Std.Dev.	1.5	769.2		Std.Dev.	1.2	509.7

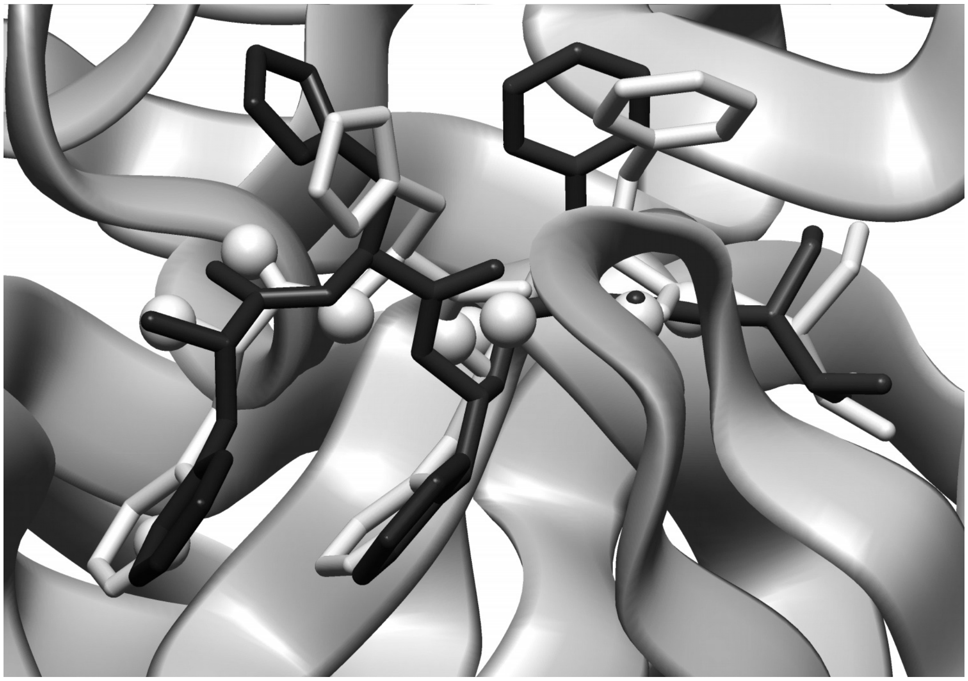

Figure 9. *Close-up view of the known ligand-binding mode for rhizopuspepsin and that predicted by* Specitope *based on triangle hashing.* The protein backbone is shown in grey ribbons, with the peptidyl ligand FHFFV, from its crystal structure in complex with the protein (PDB 3apr), shown in black tubes. The similar binding mode for this peptide predicted by Specitope is shown in grey tubes, and the template points (to which triangles were matched) are shown as grey spheres.

interface to SPECITOPE, including flexible and rigid bond information as well as hydrogen-bond donors and acceptors and hydrophobic interaction sites for each ligand candidate, will facilitate screening of organic compounds as well as peptides. Ultimately, we anticipate a merging of the capabilities of SPECITOPE and FIRST, which can predict regions of protein backbone flexibility (see companion paper by Jacobs, Kuhn, and Thorpe), to model the interactions between fully-flexible ligands *and* fully-flexible proteins. This problem of modeling the induced shape complementarity between proteins and ligands upon binding remains one of the most difficult and important problems in structural biology.

ACKNOWLEDGMENTS

We thank Mike Thorpe and Phil Duxbury for organizing and involving us in this exciting workshop, and the Deutsche Forschungsgemeinschaft (postdoctoral fellowship SCHN 576/1-1 to V.S.) and the MSU Vice Provost for Libraries, Computing, and Technology for supporting this research.

REFERENCES

1. I. D. Kuntz. *Science* 257:1078–1081 (1992).
2. C. L. Verlinde and W. G. Hol. *Structure* 2(7):577–587 (1994).
3. I. D. Kuntz, E. C. Meng, and B. K. Shoichet. *Acc. Chem. Res.* 27(5):117–123 (1994).
4. D. A. Gschwend, A. C. Good, and I. D. Kuntz. *J. Mol. Recog.* 9:175–186 (1996).
5. D. S. Goodsell and A. J. Olson. *Proteins* 8:195–202 (1990).
6. G. M. Morris, D. S. Goodsell, R. Huey, and A. J. Olson. *J. Comput. Aided Mol. Des.* 10:293–304 (1996).
7. C. D. Rosin, R. S. Halliday, W. E. Hart, and R. K. Belew. In *Proc. 7th Int. Conf. on Genetic Algorithms (ICGA)*, T. Bäck, editor, 221–228 (Morgan Kaufmann Publishers, Inc., San Francisco, CA, 1997).
8. G. M. Morris, D. S. Goodsell, R. S. Halliday, R. S. Huey, W. E. Hart, R. K. Belew, and A. J. Olson. *J. Comput. Chem., in press* (1998).
9. I. D. Kuntz, J. M. Blaney, S. J. Oatley, R. Langridge, and T. E. Ferrin. *J. Mol. Biol.* 161:269–288 (1982).
10. B. K. Shoichet and I. D. Kuntz. *Protein Eng.* 6(7):723–732 (1993).
11. A. R. Leach and I. D. Kuntz. *J. Comput. Chem.* 13(6):730–748 (1992).
12. J. Ruppert, W. Welch, and A. N. Jain. *Protein Science* 6:524–533 (1997).
13. M. Rarey, S. Wefing, and T. Lengauer. *J. Comput. Aided Mol. Des.* 10:41–54 (1996).
14. M. Rarey, B. Kramer, T. Lengauer, and G. Klebe. *J. Mol. Biol.* 261:470–489 (1996).
15. W. Welch, J. Ruppert, and A. N. Jain. *Chem. Biol.* 3:449–462 June (1996).
16. G. Jones, P. Willett, and R. C. Glen. *J. Mol. Biol.* 245:43–53 (1995).
17. G. Jones, P. Willett, R. C. Glen, A. R. Leach, and R. Taylor. *J. Mol. Biol.* 267:727–748 (1997).
18. C. M. Oshiro, I. D. Kuntz, and J. Scott Dixon. *J. Comput. Aided Mol. Des.* 9:113–130 (1995).
19. D. K. Gelhaar, G. M. Verkhivker, P. A. Reijto, C. J. Sherman, D. B. Fogel, L. J. Fogel, and S. T. Freer. *Chem. Biol.* 2:317–324 (1995).
20. R. M. A. Knegtel, I. D. Kuntz, and C. M. Oshiro. *J. Mol. Biol.* 266:424–440 (1997).
21. N. C. J. Strynadka, M. Eisenstein, E. Katchalski-Katzir, B. K. Shoichet, I. D. Kuntz, R. Abagyan, M. Totrov, J. Janin, J. Cherfils, F. Zimmerman, A. Olson, B. Duncan, M. Rao, R. Jackson, M. Sternberg, and M. N. G. James. *Nature Struct. Biol.* 3(3):233–239 (1996).
22. M. C. Lawrence and P. C. Davis. *Proteins* 12:31–41 (1992).
23. H.-J. Böhm. *J. Comput. Aided Mol. Des.* 8:623–632 (1994).
24. B. K. Shoichet, R. M. Stroud, D. V. Santi, I. D. Kuntz, and K. M. Perry. *Science* 259:1445–1450 (1993).
25. D. A. Gschwend, W. Sirawaraporn, D. V. Santi, and I. D. Kuntz. *Proteins* 29:59–67 (1997).
26. D. M. Lorber and B. K. Shoichet. *Protein Science* 7(4):938–950 (1998).
27. A. R. Friedman, V. A. Roberts, and J. A. Tainer. *Proteins* 20:15–24 (1994).
28. V. Schnecke, C. A. Swanson, E. D. Getzoff, J. A. Tainer, and L. A. Kuhn. *Proteins* 33(1):74–87 (1998).
29. G. M. Crippen and T. F. Havel. *Distance Geometry and Molecular Conformation*. John Wiley & Sons, New York, (1988).
30. A. R. Fersht. *Trends Biochem. Sci.* 12:301–304 (1987).
31. J. Janin and C. Chothia. *J. Biol. Chem.* 265(27):16027–16030 (1990).

32. M. L. Raymer, P. C. Sanschagrin, W. F. Punch, S. Venkataraman, E. D. Goodman, and L. A. Kuhn. *J. Mol. Biol.* 265:445–464 (1997).

33. L. A. Kuhn, C. A. Swanson, M. E. Pique, J. A. Tainer, and E. D. Getzoff. *Proteins* 23:536–547 (1995).

34. H.-J. Böhm. *J. Comput. Aided Mol. Des.* 8:243–256 (1994).

35. A. N. Jain. *J. Comput. Aided Mol. Des.* 10:427–440 (1996).

36. E. E. Abola, F. C. Bernstein, S. H. Bryant, T. F. Koetzle, and J. Weng. In *Crystallographic Databases – Information Content*, *Software Systems*, *Scientific Applications*, F. H. Allen, G. Bergerhoff, and R. Sievers, editors, 107–132. Data Commission of the International Union of Crystallography, Bonn/Cambridge/Chester (1987).

37. E. A. Lunney. *Network Science* 1:http://www.awod.com/netsci/Issues/Sept95/feature1.html (1995).

38. U. Hobohm and C. Sander. *Protein Science* 3(3):522–524 (1994).