J|A|C|S
ARTICLES
Published on Web 01/25/2002

# TreeDock: A Tool for Protein Docking Based on Minimizing van der Waals Energies†

Amr Fahmy and Gerhard Wagner*

*Contribution from the Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, 240 Longwood Avenue, Boston, Massachusetts 02115*

Received May 21, 2001. Revised Manuscript Received November 29, 2001

***Abstract:*** Predicting protein−protein and protein−ligand docking remains one of the challenging topics of structural biology. The main problems are (i) to reliably estimate the binding free energies of docked states, (ii) to enumerate possible docking orientations at a high resolution, and (iii) to consider mobility of the docking surfaces and structural rearrangements upon interaction. Here we present a novel algorithm, TreeDock, that addresses the enumeration problem in a rigid-body docking search. By representing molecules as multidimensional binary search trees and by exploring a sufficient number of docking orientations such that two chosen atoms, one from each molecule, are always in contact, TreeDock is able to explore all clash-free orientations at very fine resolution in a reasonable amount of time. Due to the speed of the program, many contact pairs can be examined to search partial or complete surface areas. The deterministic systematic search of TreeDock is in contrast to most other docking programs that use stochastic searches such as Monte Carlo or simulated annealing methods. At this point, we have used the Lennard-Jones potential as the only scoring function and show that this can predict the correct docked conformation for a number of protein−protein and protein−ligand complexes. The program is most powerful if some information is known about the location of binding faces from NMR chemical-shift perturbation studies, orientation information from residual dipolar coupling, or mutational screening. The approach has the potential to include docking-site mobility by performing molecular dynamics or other randomization methods of the docking site and docking families to families of structures. The performance of the algorithm is demonstrated by docking three complexes of immunoglobulin superfamily domains, CD2 to CD58, the $V_\alpha$ domain of a T-cell receptor to its $V_\beta$ domain, and a T-cell receptor to a pMHC complex as well as a small molecule inhibitor to a phosphatase.

Association of proteins with other macromolecules or smaller ligands is one of the fundamental events in biology. Much experimental and theoretical work has been dedicated to unraveling the principles of protein interactions.[1] However, predicting correctly associated configurations of protein complexes still remains a challenge. Experimental methods for protein-structure determination have improved dramatically over the past decade, and the number of protein structures determined is ever increasing. However, structures of complexes are still a small minority among the entries in the Protein Data Bank (www.rcsb.org/pdb) although most proteins function in the context of larger complexes. This imbalance reflects the fact that it is more difficult to prepare complexes suitable for structural studies and to determine their structures. In addition, many interactions are weak, and stable complexes that can be studied experimentally may not form. Thus, it would be quite rewarding to have efficient and reliable computational tools available to predict correctly conformations of protein complexes based on experimental structures of the free molecules.

Factors that determine binding affinity and specificity that should ideally be considered when scoring docking conformations include steric complementarity of the shapes of the interaction sites, electrostatic interactions, and hydrogen bonding. Furthermore, exclusion of the solvent from the interface and the associated solvent entropy change play an important role in stabilization of protein interactions. Protein−protein and protein−ligand docking is a dynamic process where sometimes even major conformational changes may take place. The complexity of the dynamic docking problem is comparable to predicting protein-folding. Approaches have been reported to predict docking of flexible ligands to proteins[2−6] or flexible

* Address correspondence to this author at Department of Biological Chemistry and Molecular Pharmacology, Building C1, Room 112, 240 Longwood Avenue, Harvard Medical School, Boston, MA 02115. Telephone: 1-617-432-3213. FAX: 1-617-432-4383. Email: gerhard_wagner@hms.harvard.edu.

† Abbreviations used: L-J, Lennard-Jones; vdw, van der Waals; RMSD, root-mean-square deviation; RMSDM, RMSD of the movable protein; RMSDDS, RMSD of the docking site; NMR, nuclear magnetic resonance; CD, cluster of differentiation; TCR, T-cell receptor; MHC, major histocompatibility complex; PDB, protein data bank.

(1) Jones, S.; Thornton, J. M. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 13−20.
(2) DesJarlais, R. L.; Sheridan, R. P.; Dixon, J. S.; Kuntz, I. D.; Venkataraghavan, R. *J. Med. Chem.* **1986**, *29*, 2149−2153.
(3) Makino, S.; Ewing, T. J.; Kuntz, I. D. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 513−532.
(4) Maurer, M. C.; Trosset, J. Y.; Lester, C. C.; DiBella, E. E.; Scheraga, H. A. *Proteins* **1999**, *34*, 29−48.
(5) Oshiro, C. M.; Kuntz, I. D.; Dixon, J. S. *J. Comput.-Aided. Mol. Des.* **1995**, *9*, 113−130.
(6) Wang, J.; Kollman, P. A.; Kuntz, I. D. *Proteins* **1999**, *36*, 1−19.

proteins to flexible proteins.[7] However, it seems that none of these methods can reliably predict from basic principles the docked conformations of protein−protein complexes from the structures of the free protein components. Rigid docking routines are important components of more-general docking programs, and predicting rigid docking is an important problem.[8] Efficient and correct algorithms for rigid docking may play a key role in predicting flexible docking, as flexible docking may be dissected into a manifold of rigid docking cases.[9] Good rigid-docking programs should be able to reassemble a protein−protein heterodimer after separating the components. Although this seems a simple task, we believe that there are no programs that can achieve this with a satisfactory low root-mean-square deviation (RMSD) to the correct complex structure of less than 1 Å.[10,11]

The existing algorithms for protein docking utilize computational search methods to predict good docking configurations while minimizing a target function. Most published algorithms use stochastic search procedures, such as Monte Carlo or simulated annealing methods. Few programs attempt a complete enumeration of the search space and only at low resolution.[12]

There are many programs that attempt to solve the docking problem, see refs 10, 11, 13, and 14 for reviews. Numerous algorithmic techniques have been employed to solve this problem. These include but are not limited to the following: Searching for negative images of the receptor in a database of ligands is the technique used in DOCK.[15] LUDI [16] uses subgraph isomorphism as another technique to search for ligands in a database. Subgraph isomorphism is also used in programs such as ALLADIN[17] and FOUNDATION.[18] AUTODOCK [19] uses a Lamarckian genetic algorithm for docking. Programs such as GRID,[20] CLIX,[21] and DOCK [15] all use grid-based energy computation. Monte Carlo combined with simulated annealing is the method used in PRODOCK,[22] which is primarily a flexible docking program. A mathematically elegant method [12] based on the fast Fourier transform, has resulted in three different implementations, an original implementation, FTDOCK,[23] and VDW-FFT.[24] The program MCSS[9] uses molecular dynamics to generate multiple functional copies of the docking site and simultaneously docks to each of them. The $A^*$ algorithm was used as a search method to dock proteins with discrete side-chain flexibility.[25] None of these programs claim to work for all molecules.

van der Waals interactions are crucial in defining shape complementarity. However, this energy term is often avoided since the slope of L-J potential is steep and is sensitive to steric clashes. Most docking programs optimize some other function. For example, the docking algorithms that are based on the fast Fourier transform maximize the number of overlapped centers of surface atoms.[12] This approximates the goal of maximizing the contact surface but does not minimize the L-J potential.

The L-J potential between a pair of atoms at distance $r$ is composed of two parts, a repulsive (positive) part proportional to $r^{-12}$ and an attractive (negative) part proportional to $r^{-6}$. The L-J potential is highly sensitive when atoms are near their van der Waals contact distance which is the distance at which the L-J potential is optimal. Closer than this distance, the repulsive part becomes dominant and grows very rapidly with decreasing distance, and above the van der Waals distance, the attractive component is dominant. Between a single pair of atoms the attractive L-J potential is insignificant. However, when numerous pairs of atoms are simultaneously close to their van der Waals distances, the attractive L-J potential becomes significant. This happens only when the shapes of the molecules are complementary.

Note that a molecular configuration with an overall negative L-J potential does not preclude the existence of pairs of atoms with repulsive (but small) L-J potential. Because the L-J potential can be so high when the distances between atoms are too close, the energy due to a single misplaced atom can drown all the other terms in the energy function, including the electrostatic term, low as they may be. This makes the L-J potential a crucial factor when it comes time to optimize the binding energy function which is the main subject of this paper. Minimizing the L-J potential can be expected to maximize the buried surface area, which minimizes the free energy of the solvent.[1]

Search spaces are very large and difficult to cover exhaustively. Only few docking programs attempt to enumerate all of the relative configurations of a pair of molecules. This search space is very large and often cannot be totally covered using a fine enough resolution to satisfy the demands of locating low values of the energy function. To overcome this, large increments are used to get from one configuration to the next thus possibly missing low energy configurations. For example, FTDOCK[23] and vdw-FFT [24] use 20° angular increments in their Euler-angle rotations.

We developed a new program, TreeDock, that enumerates the search space at a user-defined resolution subject to the condition that a pair of atoms, one from each molecule, are always in contact. Three critical design features are the following: (1) TreeDock represents molecules as multidimensional search trees to speed up the computation of a scoring function. (2) While exhaustively (up to the specified resolution) searching all but one dimension of the search space, TreeDock employs a fast analytic algorithm to locate low energy configurations for the innermost dimension. (3) TreeDock allows its user to employ experimental results to make the search space small. So far, TreeDock uses the L-J potential as the only scoring function. Despite this limitation, TreeDock has always found

(7) Sternberg, M. J.; Aloy, P.; Gabb, H. A.; Jackson, R. M.; Moont, G.; Querol, E.; Aviles, F. X. *Ismb* **1998**, *6*, 183−192.
(8) Clore, G. M. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 9021−9025.
(9) Miranker, A.; Karplus, M. *Proteins* **1991**, *11*, 29−34.
(10) Sobolev, V.; Moallem, T. M.; Wade, R. C.; Vriend, G.; Edelman, M. *Proteins* **1997**; *Suppl*. 210−214.
(11) Sternberg, M. J.; Gabb, H. A.; Jackson, R. M. *Curr. Opin. Struct. Biol.* **1998**, *8*, 250−256.
(12) Katchalski-Katzir, E.; Shariv, I.; Eisenstein, M.; Friesem, A. A.; Aflalo, C.; Vakser, I. A. *Proc . Natl. Acad. Sci. U.S.A.* **1992**, *89*, 2195−2199.
(13) Verlinde, C. L.; Hol, W. G. *Structure* **1994**, *2*, 577−587.
(14) Strynadka, N. C.; Eisenstein, M.; Katchalski-Katzir, E.; Shoichet, B. K.; Kuntz, I. D.; Abagyan, R.; Totrov, M.; Janin, J.; Cherfils, J.; Zimmerman, F.; Olson, A.; Duncan, B.; Rao, M.; Jackson, R.; Sternberg, M.; James, M. N. *Nat. Struct. Biol.* **1996**, *3*, 233−239.
(15) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. *J. Mol. Biol.* **1982**, *161*, 269−288.
(16) Böhm, H. *J. Comput.-Aided Mol. Des.* **1992**, *6*, 61−78.
(17) van Drie, J.; Weininger, D.; Martin, Y. *J. Comput.-Aided Mol. Des.* **1989**, *3*, 225−254.
(18) Ho, C.; Marshall, G. *J. Comput.-Aided Mol. Des.* **1993**, *7*, 322.
(19) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. *J. Comput. Chem.* **1998**, *19*, 1639−1662.
(20) Goodford, P. J. *J. Med. Chem.* **1985**, *28*, 849−857.
(21) Lawrence, M. C.; Davis, P. C. *Proteins* **1992**, *12*, 31−41.
(22) Trosset, J. Y.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 8011−8015.
(23) Gabb, H. A.; Jackson, R. M.; Sternberg, M. J. *J. Mol. Biol.* **1997**, *272*, 106−120.
(24) Bliznyuk, A.; Gready, J. *J. Comput. Chem.* **1999**, *20*, 983−988.
(25) Leach, A. *J. Mol. Biol.* **1994**, *235*, 345−356.

the correct solution to be the one with the lowest L-J energy in all rigid docking cases studied.

## Results

**Development of TreeDock.** TreeDock is designed to enumerate efficiently many different orientations at a sufficient resolution in a reasonable amount of time. At the present state, we restrict our objective to solving docking problems under the assumptions that (1) the molecules are rigid bodies, where flexibility can be modeled by generating different rigid bodies with different conformations, and (2) for large molecules, the program user can tell TreeDock an approximate docking site. These restrictive assumptions are meaningful when we target a class of docking situations in which conformational changes are absent or small, a class that may include many cases of docking ligands to proteins. Often the docking site can be approximated by experimental means, for example, by NMR chemical-shift mapping,[26] NOE mapping,[27] or alanine-scanning of protein surfaces.[28,29]

TreeDock uses accurate coordinates of the atoms of the molecules; it does not use grids to compute energy. Rather successive configurations are obtained by using transformations that are adjusted to satisfy a user-defined search resolution parameter. The whole of the molecule is transformed using the same transformation which preserves its shape.

TreeDock minimizes the L-J potential, which is a good indicator of shape fit and often suffices to find low RMSD configurations. It allows configurations of the two molecules in which atoms from different molecules have repulsive (positive) L-J energy; the distance between atoms is allowed to be as small as 70% of the optimal van der Waals distance, a condition present frequently in complexes found in the PDB and commonly used in simulated annealing algorithms for structure determination.[30]

The search space defined in TreeDock is intuitive to the user; it is based on the concept of anchors, the docking of the molecules based on the condition that two chosen atoms (anchors), one from each molecule, must be in contact. The size of the search space in TreeDock is much smaller than in other programs because it uses docking-site information which is supplied by the user. Therefore, the search resolution can be raised which is what is needed to minimize the L-J potential. The smaller size of the search space allows TreeDock to finish the search in a reasonable amount of time while delivering configurations that are very close to the minimum-energy configuration if the choice of anchors and the search resolution permit it. Also, the user will know that the search space has been covered completely (up to the specified search resolution), thus providing information that may help in deciding on their next step.

To test TreeDock, as will be discussed, we selected docked configurations of molecules from the Protein Data Bank (PDB) and went through three steps: (1) take the molecules apart by translating and rotating the coordinates of one molecule of the docked pair; (2) provide TreeDock the coordinates of each molecule separately along with, in most cases, an indication of the approximate docking sites, (3) run TreeDock and determine how closely TreeDock puts the molecules back together in their known proper docking configuration, as measured by the root-mean-square deviation (RMSD) of coordinates of the molecules.

The performance of the algorithm is demonstrated by docking two complexes of immunoglobulin superfamily domains, CD2 to CD58, the $V_\alpha$ domain of a T-cell receptor to its $V_\beta$ domain, and a T-cell receptor to its cognate pMHC complex, as well as a small molecule inhibitor to a phosphatase. Additionally, TreeDock was successfully used as a docking tool in a study of small-molecule inhibitors of the antiapoptotic protein Bcl-xL as described in the accompanying manuscript.[31]

**The Search Space.** In searching for docking configurations, the user provides TreeDock a PDB file that specifies each molecule as a ball-model, i.e., a union of atoms, each defined by the position of its center. The user also provides TreeDock enough information so it can pick a pair of atoms, one from each molecule, called *anchors*. At the cost of increased execution time, the anchors can be more loosely specified and TreeDock can try out various pairs of anchors. If one molecule is small, e.g., a ligand, an anchor does not need to be specified, and TreeDock will try all possible anchor pairs.

Of the two molecules, the larger, called *F*, is fixed in space; the other, called *M*, for "movable", undergoes translations and rotations to generate a search space of configurations. Because in a lowest energy configuration the two molecules must touch, TreeDock can limit its search space: For each pair of anchors that it deals with, TreeDock moves *M* while keeping the *M*-anchor in contact with the *F*-anchor so they meet at some tangent point. The contact constraint leaves five degrees of freedom, four of which are searched exhaustively with respect to a prespecified resolution in a manner to be described shortly; the last degree of freedom is a rotation, analyzed by the *merry-go-round algorithm* which computes the minimum energy configuration, as will be described.

**The Exhaustive Search and Its Resolution.** Each configuration of the two molecules can be reached by translations and rotations of *M* that bring a point on the surface of the *M*-anchor into tangency with a point on the surface of the *F*-anchor, followed by a rotation of *M* around the axis through the centers of the two anchors. Thus the search is over five degrees of freedom: two translational degrees to cover the surface of the *F*-anchor, two rotational degrees to cover the surface of the *M*-anchor, and a third rotational degree about the axis through the centers of the anchors (see Figure 1). Limiting each anchor surface to the part unblocked by atoms of its own molecule—the solvent accessible surface—further shortened the search.

Since the degrees of freedom are continuous, exhaustive search depends on the continuity of the energy as a function of configuration: a sufficiently fine discrete selection of configurations is all that has to be searched. TreeDock selects enough representative points called *contact points* from the solvent accessible surface of the *F*-anchor and the solvent accessible surface of the *M*-anchor to ensure adequate resolution.

(26) Shuker, S. B.; Hajduk, P. J.; Meadows, R. P.; Fesik, S. W. *Science* **1996**, *274*, 1531−1534.
(27) Takahashi, H.; Nakanishi, T.; Kami, K.; Arata, Y.; Shimada, I. *Nat. Struct. Biol.* **2000**, *7*, 220−223.
(28) Bass, S. H.; Mulkerrin, M. G.; Wells, J. A. *Proc. Natl. Acad. Sci. U.S.A.* **1991**, *88*, 4498−4502.
(29) Wells, J. A. *Methods Enzymol.* **1991**, *202*, 390−411.
(30) Brünger, A. T.; Adams, P. D.; Clore, G. M.; DeLano, W. L.; Gros, P.; Grosse-Kunstleve, R. W.; Jiang, J. S.; Kuszewski, J.; Nilges, M.; Pannu, N. S.; Read, R. J.; Rice, L. M.; Simonson, T.; Warren, G. L. *Acta Crystallogr. Sect. D: Biol. Crystallogr.* **1998**, *54*, 905−921.
(31) Lugovskoy, A. A.; Degterev, A. I.; Fahmy, A. F.; Zhou, P.; Gross, J. D.; Yuan, J.; Wagner, G. *J. Am. Chem. Soc.* **2002**, *124*, 1234−1240.

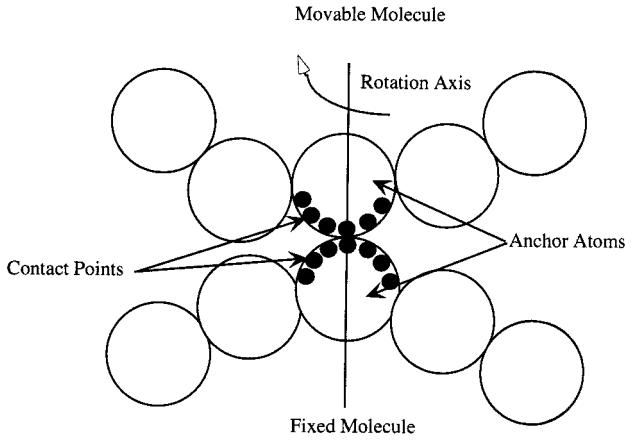


***Figure 1.*** The fixed molecule *F* and the movable molecule *M* touching at contact points on the surface of the anchor atoms. Having fixed an *F*-contact point, choosing between *M*-contact points rotates *M*, and changing *F*-contact points translates *M*. After choosing the *F*-contact point and the *M*-contact point, the only degree of freedom left is a rotation about the axis that connects the centers of the anchors (and the contacts points). The merry-go-round algorithm delimits the clash-free angles (if any) between the two molecules when the *M*-molecule is rotated about this axis.

To find the lowest energy configuration, TreeDock operates nested loops: TreeDock loops through *F*-contact points. For each *F*-contact point, TreeDock enters an inner loop and steps through *M*-contact points; for each *M*-contact point, TreeDock brings molecule *M* in touch with *F* so that the two anchors are tangent and the two contact points are coincident, as shown in Figure 1.

There is a one-parameter family of configurations that keeps a pair of contact points coincident, while allowing rotation of molecule *M* about an axis through the centers of the anchors (and the contact points). TreeDock employs the merry-go-round algorithm to determine analytically the configuration of lowest energy within this family. The configuration found for a given *M*−*F* pair of contact points is compared with a previously found least-energetic configuration, and, if lower, replaces it as the "best so far".

To achieve adequate resolution in the scanning of contact points, the user tells TreeDock the value of a parameter *R* called the *search resolution* (in Å) and TreeDock organizes the search so that no atom of *M* is moved by more than *R* in a change from one configuration to a neighboring configuration. The parameter *R* can be varied by the user to trade precision for running time. Given *R*, TreeDock generates a set of *F*-contact points and a set of *M*-contact points.

Moving from one *M*-contact point to its neighbor necessarily rotates molecule *M*. To bar this rotation from moving the atom of *M* most distant from the *M*-anchor by more than *R*, the angle (in radians) between any *M*-contact point and its neighboring *M*-contact points is made less than

$$\Delta\theta_{\mathrm{rot}} = R/d_{\mathrm{max}} \tag{1}$$

where $d_{\mathrm{max}}$ is the distance from the center of the *M*-anchor to the farthest atom of *M*.

Interestingly, the spacing of *F*-contact points can be much coarser, because at each of them all relevant rotations are explored using the *M*-contact points; hence the spacing of *F*-contact points is limited only by the need to makes translations from one to its neighbors be less than *R*. This is achieved by making the angle (in radians) between neighboring *F*-contact points less than

$$\Delta\theta_{trans} = R/d_{\mathrm{fixed}} \tag{2}$$

where $d_{\mathrm{fixed}}$ is the radius of the fixed anchor. In this manner it is guaranteed that all atoms of *M* will not be placed by more than 2*R* Å from a previous location by any step in the search, so that the absolute best configuration will have no atom more than *R* distant from a configuration examined in the search.

**The Energy Function and Multidimensional Search Trees.** Critical to TreeDock's capacity to deliver results in a reasonable amount of time is its use of a multidimensional search tree,[32,33] to take advantage of a cutoff feature for the computation of the L-J potential. The L-J potential is given by

$$\sum_i \sum_j \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \tag{3}$$

where the *i* and *j* summations are taken over atoms of the *F* and *M* molecules, respectively. The constants $A_{ij}$ and $B_{ij}$ depend on the types of atoms, and $r_{ij}$ is the distance between atoms *i* and *j* of the molecules *F* and *M*, respectively. We used appropriate values of the constants $A_{ij}$ and $B_{ij}$ supplied in the X-PLOR program.[30,34]

The L-J potential, (eq 3), has a contribution from each pair of atoms, one from *M* and one from *F*; however, if the atoms of an *M*−*F* pair are separated by more than a separation parameter $r_m$, the contribution of the pair is negligible. Thus the time spent calculating energy can be greatly reduced by skipping pairs that contribute a negligible amount, i.e., pairs separated by more than $r_{\mathrm{m}}$. TreeDock saves time by building a multidimensional search tree, once for each *F*-contact point, and using the tree repeatedly in the inner loop to find, for each *M*-atom, the *F*-atoms separated from it by less than $r_m$. In this way it efficiently finds all the *M*−*F* pairs that contribute significantly to the energy. These atoms are found in an expected-time logarithmic in the number of atoms of the *F* molecule and linear in $r_m$.

Multidimensional search trees,[32,33] or kd-search trees, have been used in computer science extensively for efficient reporting of proximity queries and range searching in many different applications. They generalize the binary search algorithm to *k*-dimensions. For each *F*-contact point explored, TreeDock changes to cylindrical coordinate frame ($\rho$, $\varphi$, *z*) in which the center of the *M*-anchor is at the origin and the center of the *F*-anchor is on the negative *z*-axis. The tree-building subroutine takes the unsorted list of coordinate triples ($\rho_i$, $\varphi_i$, $z_i$) for centers of the *F*-atoms in this frame and makes a two-dimensional tree for the coordinates $\rho$ and *z*, ignoring the $\varphi$-coordinate. To do this, the subroutine finds the median of the *z*-coordinates of the *F*-atoms and assigns all the atoms to two subtrees, one for atoms having *z*-coordinate above the median and one for atoms having a *z*-coordinate at or below the *z* median. The subroutine then flips to the $\rho$-dimension and splits each subtree into two subtrees, one for atoms having $\rho$-coordinates greater than the $\rho$ median of the subtree, etc.; then the subroutine flips back to *z*, and so

(32) Bentley, J. *Commun. ACM* **1975**, *18*, 509−517.
(33) Preparata, F.; Shamos, M. *Computational Geometry*; Springer-Verlag: Berlin, 1985.
(34) Engh, R. A.; Huber, R. *Acta Crystallogr.* **1991**, *A47*, 392−400.

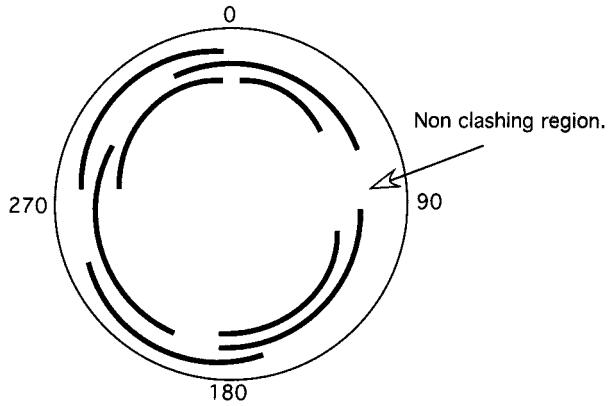


***Figure 2.*** The angular interval in which the molecules clash is the union of angular intervals in which any atom from the surface of the movable molecule *M* clashes with some atom from the fixed molecule *F* (dark arcs). The energy is computed only in the clash-free region if it exists.

on, recursively, until it builds a tree with no more than one atom in each leaf node.

Searching for a single atom can be done in logarithmic expected time in the number of atoms in the tree and searching for atoms within a certain range can be done in logarithmic time plus the size of the range. In our case, the size of the range will be small and will not affect the search time in any substantial way. Logarithmic search time is achieved only if the tree is roughly balanced, otherwise it becomes linear.

**The Merry-Go-Round Algorithm.** Each $M-F$ pair of contact points determines a one-parameter family of configurations indexed by a rotation. For the cylindrical coordinate frame adapted to a given $M-F$ pair, each rotation is about the *z*-axis by some angle $\theta$; this takes the coordinate $\varphi$ to $\varphi + \theta$ while leaving *z* and $\rho$ invariant. Within this one-parameter family of configurations, we will speak of the configuration $\theta$. The merry-go-round algorithm determines whether the minimum energy of the family of configurations defined by the $M-F$ pair is lower than a previously found "least energy so far", and if so, it determines that energy and the value of $\theta$ for which it occurs. For this, it finds within the set of configurations $0 \leq \theta \leq 2\pi$ the subset (possibly empty) in which the surface of *M* does not penetrate the interior of the *F* or vice versa and in which the surface atoms of *M* and *F* do not overlap enough to cause the L-J potential to be above a high threshold (see Figure 2). We say two atoms *clash* if their centers are separated by sufficiently less than the sum of van der Waals radii of the two atoms to push the L-J potential over a certain threshold. Speeded by its use of the multidimensional tree built before it is called, merry-go-round eliminates angles where the molecules clash and then, within the remaining range, searches for minimum energy configurations.

The set of angles at which the molecule *M* clashes with molecule *F* is the union of the sets of angles for which any atom of *M* clashes with any atom from *F* (see Figure 2). In many cases merry-go-round finds that the *M* and *F* clash over the whole of the $2\pi$ range, even before covering all of the atoms of *M*. Indeed, a single atom of *M* can clash with *F* for all rotations about the *z*-axis. When this occurs, the energy is very high and merry-go-round eliminates the *M*-contact point and goes to another one, if there is one. In contrasting cases when there exist angles free of clashes, merry-go-round finds the angle that minimizes the energy for the current choice of contact points by scanning the angular intervals that are clash-free and computing the energy within the intervals. Merry-go-round computes the energy using the set $C(m)$ for each atom *m* of *M*, which contains all the atoms of *F* for which the energy due to atom *m* is significant.

When called to deal with an $M-F$ pair of contact points, merry-go-round starts a loop through all centers of atoms of *M*. For each atom $m \in M$, merry-go-round uses the two-dimensional search tree to find the set $C(m)$ of all the *F*-atoms which some rotation about the *z*-axis brings *M* close to, meaning closer than the separation parameter $r_m$. The search tree keeps the time to do this logarithmic in the number of *F*-atoms. Within that loop, merry-go-round then enters an inner loop over all the *F*-atoms in $C(m)$ and eliminates configurations (indexed by $\theta$) for which there is a clash between *m* and any of the atoms in $C(m)$. It also computes a running union of clashing angles; if these grow to cover the whole set of rotations, Merry-go-round exits both loops, allowing TreeDock to go to the next $M-F$ pair of contact points. If there are nonclashing angles left after merry-go-round completes these loops, the set of them is a union of intervals of $\theta$. By explicit computation of energy within these intervals (taken at rotation increments that satisfy the search resolution parameter), merry-go-round computes the angle $\theta$ by which to rotate *M* so as to minimize the value of the energy function for the $M-F$ pair of contact points.

To determine analytically the angular ranges for clashing of *M* and $f \in C(m)$, merry-go-round uses a trigonometry formula that for an arbitrary distance *d* defines the (possibly empty) set of angles of rotation $\theta$ about the *z*-axis at which the centers of *M* and *F* are closer than *d*, i.e., merry-go-round determines analytically the interval(s) of $\theta$ for which

$$\cos(\theta + \varphi_m - \varphi_f) \geq \frac{\rho_m^2 + \rho_f^2 + (z_m - z_f)^2 - d^2}{2\rho_m\rho_f} \quad (4)$$

**Implementation and Testing of TreeDock.** TreeDock has been implemented in a C-program on a single SGI R10K workstation. Parameters for the scoring energy function were adopted from the X-PLOR program.[30] To identify surface atoms as potential anchors and distinguish them from interior atoms, we have determined the solvent-accessible surface area of each atom using the program Naccess,[35] which is an implementation of the Lee-and-Richards algorithm.[36]

To test the performance of the program, we have used coordinates of protein complexes deposited in the protein data bank (PDB). In each example, the two molecules of a complex were manually separated, and their relative orientation was randomized. Subsequently, TreeDock was asked to reassemble the complex while keeping the two molecules rigid. The results were then analyzed by plotting the energy of the target function against the RMSD. Identification of docked conformations that have low values of both indicate a successful prediction. The RMSD between a proposed complex and the known structure was determined by superimposing the *F*-molecule and computing the RMSD of all *M*-molecule atoms. However, the RMSD of the whole molecule may not be the best measure of the accuracy of prediction because distant atoms may have a large

(35) Hubbard, S.; Thornton, J. Department of Biochemistry and Molecular Biology; University College London: London, 1993.
(36) Lee, B.; Richards, F. M. *J. Mol. Biol.* **1971**, *55*, 379−400.

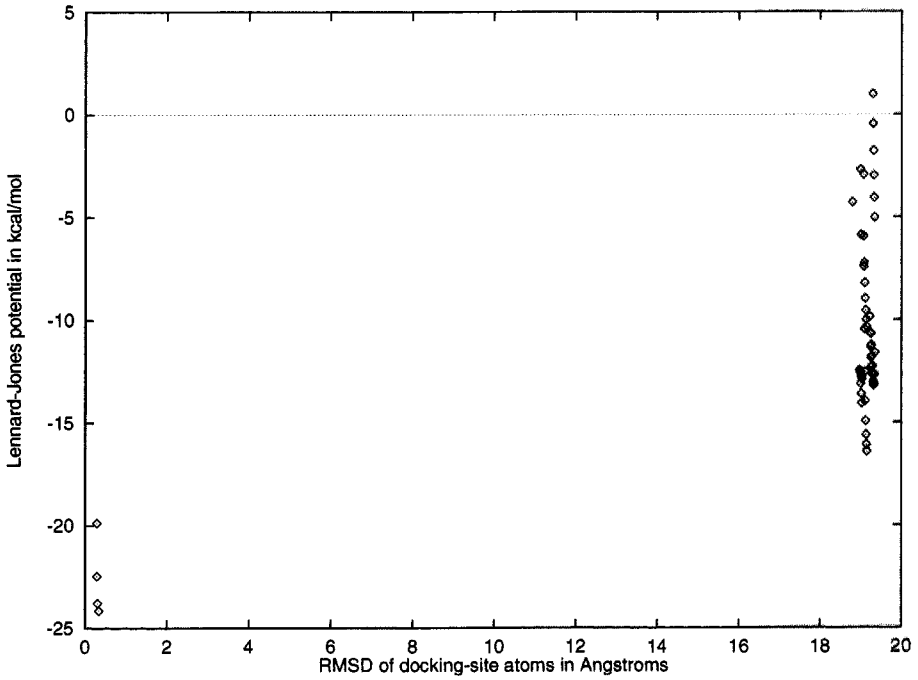


***Figure 3.*** L-J potential of all clash-free orientations versus RMSDDS of two domains of the T-cell receptor that were encountered during the search using anchors with 7.9% and 12.1% solvent-accessible surface areas.

deviation even though atoms near the docking site may be at nearly correct positions. For this reason we compute in addition to the RMSD of the whole *M*-molecule, RMSDM, a second RMSD limited to the docking site atoms, RMSDDS.

Testing TreeDock on known protein complexes had the purpose of answering the following questions: (i) Can TreeDock generally and always find the correct solution of the rigid docking problem? (ii) How much CPU time is needed for searching all reasonable docking conformations of typical small proteins with two known anchors? (iii) How does the resolution parameter *R* of TreeDock have to be set to find the correct solution? (iv) How does the shape of the molecular surface around the anchor atoms affect the search time? (v) Can we find a scoring function that yields valid results but is simple enough to allow a high-resolution search? (vi) Can the algorithm be used to dock small-molecular-weight ligands to proteins?

**Docking the Two Domains of the D10 T-Cell Receptor (TCR).** The first example is the docking of the $V_\alpha$ and $V_\beta$ domains that form the 28 kDa $F_v$ fragment of the D10 T-cell receptor whose structure was solved by NMR.[37] Here the interface is flat and primarily hydrophobic. The number of surface atoms of the $V_\beta$ domain is 1114 and that of the $V_\alpha$ domain is 903. There are 59 atoms of $V_\alpha$ making contact with 56 atoms of the $V_\beta$ domain in the complex. We use this example to test (i) whether TreeDock finds the correctly docked conformation if it is given at least one pair of atoms that contact each other and (ii) to estimate the search time for two different types of anchor pairs that are partially buried and largely exposed, respectively.

The two anchors, HE1 of TYR 335 of the $V_\beta$ domain and CG of GLN 106 of the $V_\alpha$ domain, were chosen from the middle of the docking site where the search space was expected to be small and the run time short. The search resolution *R* was chosen to be 1.0 Å. The solvent-accessible surface areas of the two anchors were 7.9% and 12.1%, respectively. After scanning the surfaces of the two anchors, 1078 contact-points on the surface of the movable anchor and 26 contact-points on the surface of the fixed anchor were found. The number of calls to merry-go-round were 28 028. A plot of the energy of orientations encountered during the search that had no surface penetration versus their RMSDDS is given in Figure 3. Four orientations with low energies were found that have low RMSDDS. The orientation with lowest L-J energy had an RMSDDS of 0.29 Å (RMSDM 0.43 Å). The L-J potential was −24.1 kcal/mol. The number of atomic contacts was 107, 47 of which resulted in positive (repulsive) energy (with distances as close as 70% of the ideal van der Waals distance), amounting to 17.3 kcal/mol, and 60 atomic contacts resulted in negative energy amounting to −41.4 kcal/mol. The run time was 2 min and 28 s of CPU time.

In a second case, two anchors, OE1 of GLN 337 of the $V_\beta$ domain and NE2 of GLN 36 of the $V_\alpha$ domain, are chosen with larger solvent-accessible surface areas, 14.0% and 34.8%, respectively. At a search resolution of 1.0 Å and since the solvent-accessible surfaces of the anchors are larger than in the previous example, more contact points were found, 23 and 5249, respectively. The number of calls to merry-go-round was 120 727. The orientation with the lowest energy had an RMSDDS of 0.96 Å (RMSDM 1.19 Å). This configuration was found in 39 min and 32 s of CPU time. The plot of L-J energy versus RMSDDS for clash-free orientations is given in Figure 4. The lowest L-J energy corresponded to low RMSDDS configurations. Thus the run time depends on the surface accessibility of the anchor atoms.

To model the uncertainty of knowing the exact atoms involved in docking, we fixed one anchor, OE1 of GLN 337 of the $V_\beta$ domain (solvent accessibility 14.0%), and chose 20 atoms from the $V_\alpha$ domain as anchors one at a time (average solvent

(37) Hare, B. J.; Wyss, D. F.; Osburne, M. S.; Kern, P. S.; Reinherz, E. L.; Wagner, G. *Nat. Struct. Biol.* **1999**, *6*, 574−581.

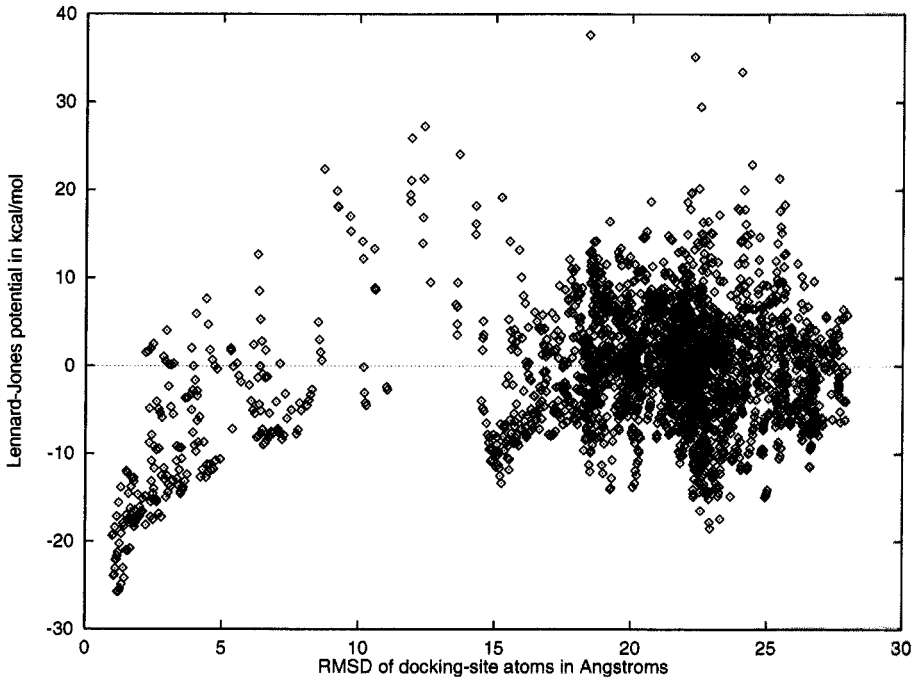


***Figure 4.*** L-J energy versus RMSDDS of all clash-free orientations of two domains of the T-cell receptor that were encountered during the search using anchors with 14.0% and 34.8% solvent-accessible surface areas.

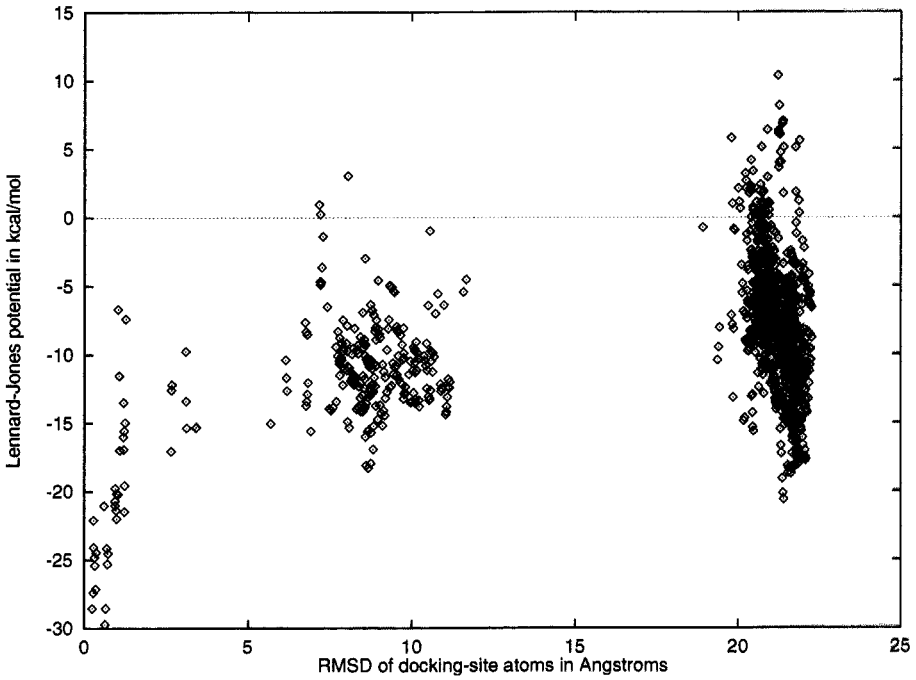


***Figure 5.*** L-J energy of all clash-free orientations of two domains of the T-cell receptor versus their RMSDDS. One anchor from one of the two molecules (with solvent accessibility of 14.0%) was picked. A total of 20 atoms from the other molecule were chosen one at a time as anchors (average solvent accessibility of 5.8%).

accessibility 5.8%). The run time was 41 min and 28 s of CPU time. The plot of the L-J energy of all clash-free orientations versus their RMSDDS is given in Figure 5. The lowest energy was −29.8 kcal/mol with an RMSDDS of 0.76 Å.

**Docking the Two Proteins CD2 and CD58.** The second complex on which we tested TreeDock is the complex composed of the two proteins CD2 and CD58.[38] There are 543 heavy atoms on the surface of CD2 and 520 on the surface of CD58. A total of 29 heavy atoms of CD2 are in contact with 23 heavy atoms of CD58. The interface is flat and primarily hydrophilic. The average solvent accessibilities of the two molecules are 24.7% and 18.8%, respectively.

We chose two anchors, OD2 of ASP 31 of CD2 and NH1 of ARG 44 of CD58, with solvent accessibilities of 17.4% and 41.0%, respectively. At a resolution of 1.0 Å, 37 contact points were found on the surface of the CD2-anchor and 3051 contact points on the surface of the CD58-anchor (the movable molecule). The number of calls to merry-go-round was 112 887.

(38) Wang, J. H.; Smolyar, A.; Tan, K.; Liu, J. H.; Kim, M.; Sun, Z. Y.; Wagner, G.; Reinherz, E. L. *Cell* **1999**, *97*, 791−803.

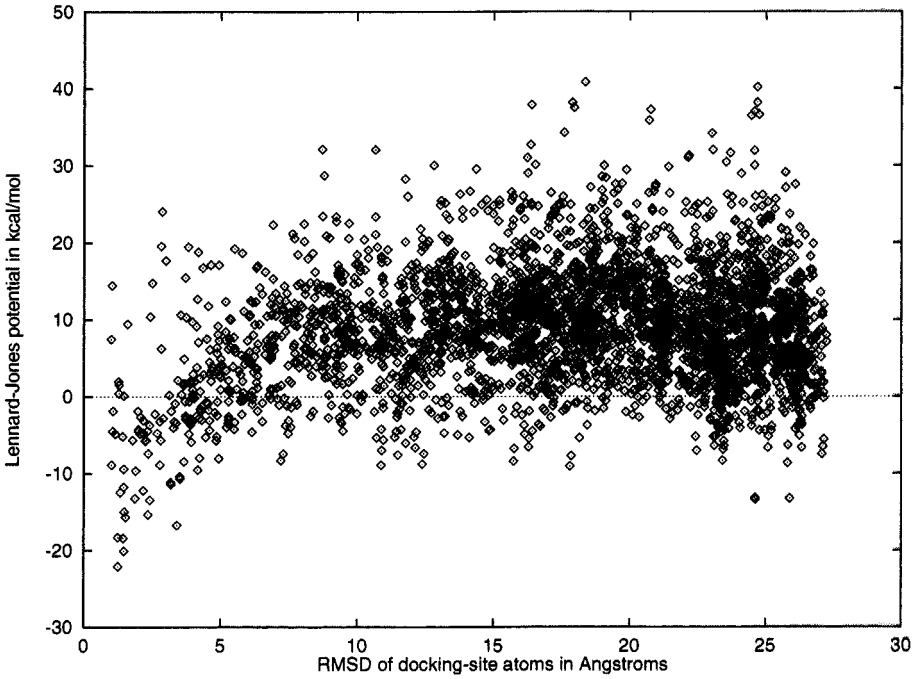


***Figure 6.*** L-J energy of all clash-free orientations versus RMSDDS of CD2 and CD58 that were encountered during the search using anchors with 17.4% and 41.0% solvent-accessible surface areas from the edge of the docking site.

The run time was 13 min and 33 s. A large number of collision-free orientations was found during the search. However, low values of the L-J energy corresponded to low RMSDDS orientations as is evident from Figure 6. The orientation with least energy had an RMSDDS of 0.715 Å (RMSDM 0.684).

**Docking a Compound to the Active Site of a Phosphatase.** TreeDock was further tested for the docking of an inhibitor, a tartrate, to the active site of the PTEN phosphatase, a tumor suppressor[39] which consists of 179-residue N-terminal domain and a 166-residue C-terminal domain. There are 19 atoms from the phosphatase making contact with 9 atoms of the tartrate. One atom from the docking site of the phosphatase was chosen (solvent-accessible area 4.2%, the average for the docking site is 3.1%) and all the atoms of the tartrate were chosen in turn as anchors. The search resolution was set at 0.7 Å. TreeDock spent 4 min and 11 s of CPU time to locate a 0.4 Å RMSDDS orientation which also corresponded to the lowest L-J energy, −18.881 kcal/mol, see Figure 7. We note that the same experiment was conducted with the search resolution set at 1.0 Å which failed to locate a simultaneously low energy and low RMSDDS configuration. This is due to the small diameter of the movable molecule, the tartrate, and shows the need for high-resolution search.

**Docking a T-Cell Receptor to a pMHC Complex.** To further model the uncertainty of knowing exactly which atoms of the two docking sites are known to be in contact, TreeDock was tested on a T-cell receptor in complex with peptide and MHC class II.[40] We chose 18 atoms from the docking site of the T-cell receptor to be potential anchors and 13 from the MHC docking site. The search resolution was set at 1.6 Å. Out of all possible pairings of the these atoms, 46 of the pairs resulted in low RMSD configurations. The best RMSDDS was 0.351 with an L-J energy of −67.193 kcal/mol, see Figure 8. The run time was 10 h and 30 mim of CPU time.

**Use of TreeDock To Determine the Docked Conformation of Experimentally Identified Inhibitors of the Antiapoptotic Protein Bcl-xL.** Recently Degterev et al. have identified inhibitors of the interaction between the antiapoptotic protein Bcl-xL and the proapoptotic Bak BH3 peptide.[41] This was achieved by screening a library of chemical compounds using a fluorescence-polarization assay where the fluorescence label was attached to the Bak peptide. NMR titrations were used to map the binding sites of the compounds on the Bcl-xL surface. The ligands identified bind with $K_D$ values in the low micromolar range. As is often observed in such a situation, many of the Bcl-xL resonances in the ligand-binding site are exchange broadened, in particular in the center of the binding site. Due to the line broadening, very few intermolecular NOEs can be observed, and those that are observed are from the periphery of the binding site that does not suffer from line broadening.[41]

To overcome this problem Lugovskoy et al.[31] have used TreeDock module supplemented with the sparse experimental intermolecular constraints and validation approach. In this hybrid approach, a family of ligand conformations was created and docked to Bcl-xL using TreeDock. The best ligand conformations were selected on the basis of shape complementarity and consistency with the experimental constraints. This was very successful and is described in the accompanying manuscript.[31] It showed that reasonable docked conformations can be obtained using such a procedure, and the values of the scoring function containing only van der Waals terms correlated very well with the experimentally measured affinities. This approach takes account of ligand mobility.

(39) Lee, J. O.; Yang, H.; Georgescu, M. M.; Di Cristofano, A.; Maehama, T.; Shi, Y.; Dixon, J. E.; Pandolfi, P.; Pavletich, N. P. *Cell* **1999**, *99*, 323−334.

(40) Reinherz, E. L.; Tan, K.; Tang, L.; Kern, P.; Liu, J.; Xiong, Y.; Hussey, R. E.; Smolyar, A.; Hare, B.; Zhang, R.; Joachimiak, A.; Chang, H. C.; Wagner, G.; Wang, J. *Science* **1999**, *286*, 1913−1921.

(41) Degterev, A.; Lugovskoy, A.; Cardone, M.; Mulley, B.; Wagner, G.; Mitchison, T.; Yuan, J. *Nat. Cell. Biol.* **2001**, *3*, 173−182.

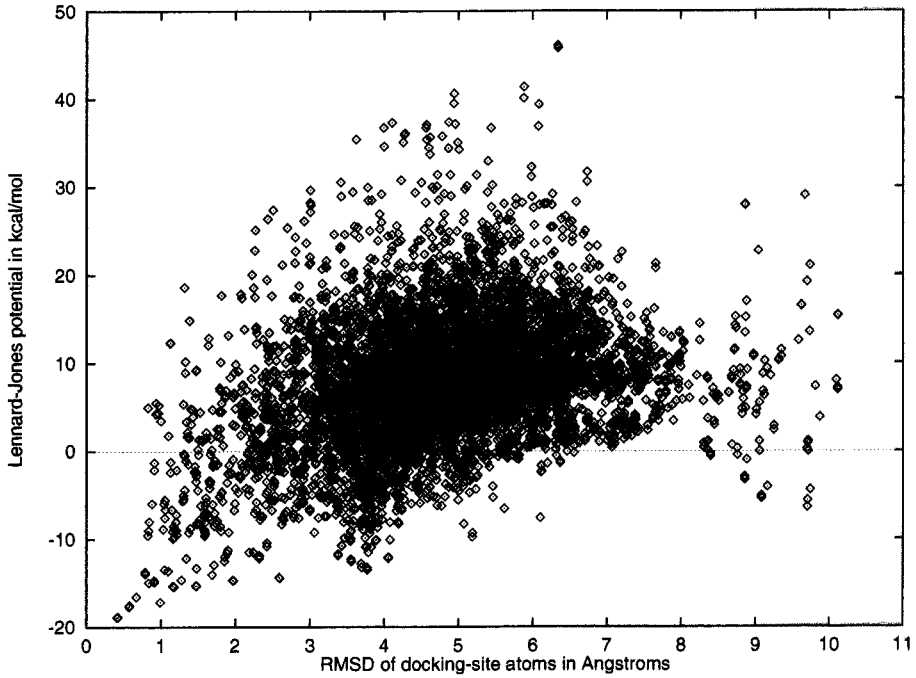


***Figure 7.*** L-J energy of all clash-free orientations versus RMSDDS of PTEN tumor suppressor and tartrate that were encountered during the search using one anchor from the phosphatase with 4.2% solvent-accessible surface area and all the atoms of the tartrate as anchors.

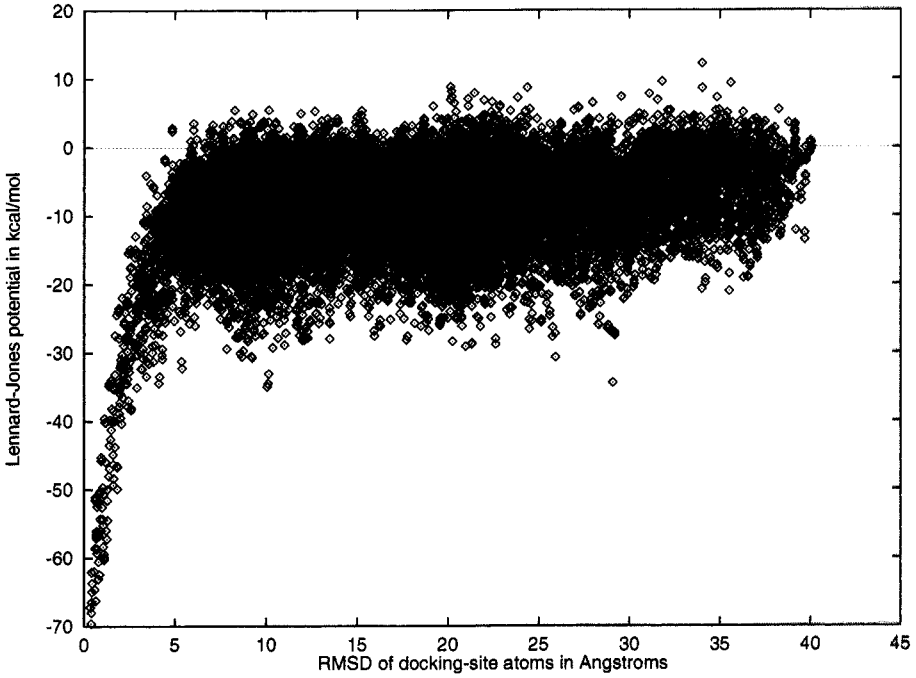


***Figure 8.*** L-J energy of all clash-fee orientations versus RMSDDS resulting from the docking of a T-cell receptor to pMHC complex that were encountered during the search using 18 anchors from the T-cell receptor with 13 anchors from pMHC complex.

## Discussion

We have developed a new program, TreeDock, that addresses the enumeration problem of protein docking. A key feature of our program is to discard efficiently all configurations where the two molecules are not in contact or have severe steric clashes. At this point, we limit the search to rigid body docking, and we only search for a minimum of the L-J potential to select for the best docking configuration. This searches essentially for the best steric surface complementarity. Due to the simplicity of the L-J potential and the dramatic reduction of search space, the residual contact space can be explored at very high resolution. The examples that we presented earlier were performed at the coarsest resolution that provided satisfactory energy values: any coarser and the search missed low energy configurations. A finer search did not result in much improvement in energy.

At the moment, our program only deals with rigid-body docking. This is certainly a severe restriction and limits us to cases where proteins do not change conformation significantly upon docking. However, this provides a baseline from which we will develop programs that include surface mobility. The merry-go-round algorithm we used here to explore the best

orientation of the *M*-molecule as a whole can also be used to explore the variation of the dihedral angles of the side chains of residues in the docking interface. Studies of this nature are underway.

So far we have limited our search of protein surfaces to smaller regions that are known to be involved in binding. This was to target our research to situations where some information about docking sites is available from mutational studies and/or NMR chemical shift mapping. Measurements of residual dipolar couplings[42] promise to provide additional restrictions of the search space that can be easily implemented in the program. As described in the section on results and test cases, identification of the best docked conformation for a single anchor pair with the typical protein size considered here takes from a fraction of a minute to several minutes on a single CPU SGI computer. The search is faster if the anchor atoms are less solvent accessible and slower if they are more exposed (larger number of permissive orientations). It is possible to perform a brute-force search with TreeDock over the entire surfaces of both proteins when using multiple CPUs. For proteins of the size of IgSF domains, such as the adhesion domains of CD2 or CD58, there are approximately 500 heavy surface atoms of either domain. Considering that a typical binding face involves about 50 atoms, only 10 to 20 representative surface atoms of one protein must be paired with all surface atoms of the other protein to include at least one correct pair. With the current performance of TreeDock (1 to 30 min CPU time per anchor pair) and 16 CPUs, this rigid docking search could be completed within 3 to 5 days. However, we are in the process of developing new algorithms to speed up the search procedures dramatically within the TreeDock framework before searching whole protein surfaces.

TreeDock will be very valuable when combined with some experimental data when the nature of the experimental constraints prevents a pure experimental determination of the complex. This is the case if a ligand binds to a protein with intermediate exchange kinetics causing line broadening for some of the protein resonances. In such a situation it is difficult to measure NOEs to the broadened resonances, which may be in the core of the binding site.

The use of TreeDock module supplemented with the limited experimental data set in conjunction with a validation procedure (described in the accompanying manuscript[31]) allowed identification of the docked conformation of inhibitors of Bcl-xL that were previously identified with high-throughput screening.[41] This indicates that our approach is reasonable and worth pursuing. Needless to say, implementation of electrostatics in the scoring function is desirable and will be pursued.

## Methods

TreeDock has been implemented in a C-program on a single SGI R10K workstation. The input to TreeDock consists of two slightly modified PDB-files, one for each molecule. Following the coordinates of each atom in a PDB file, its solvent-accessible surface (computed by the program Naccess[35]) is stored. The output of TreeDock is a single file containing the coordinates of the complex whose L-J energy is minimum. Parameters for the L-J potential were obtained from the X-PLOR program[30] and the work of Engh and Huber.[30,34]

**Acknowledgment.** This work was supported by NSF grant MCB 9527181 and NIH grants GM47467 and AI50900. Acquisition and maintenance of computers used for this work were supported in part by the Harvard Center for Structural Biology and the Giovanni Armenise-Harvard Foundation for Advanced Scientific Research. We thank Dr. Deani Cooper, Dr. Ann Ferentz, Dr. Brian Hare, Dr. Sven Hyberts, Dr. John Myers, and Dr. Jim Sun for useful discussions on all aspects of this research. We acknowledge Alexey Lugovskoy for his assistance in the computational studies of small-molecule inhibitors of the antiapoptotic protein Bcl-xL.

JA011240X

(42) Tjandra, N.; Omichinski, J.; Gronenborn, A.; Clore, G.; Bax, A. *Nat. Struct. Biol.* **1997**, *4*, 732−738.