doi:10.1006/jmbi.2001.5147 available online at http://www.idealibrary.com on IDEAL® *J. Mol. Biol.* (2001) **314**, 607–617

JMB

Binding Mode Prediction for a Flexible Ligand in a Flexible Pocket using Multi-Conformation Simulated Annealing Pseudo Crystallographic Refinement

Nobuyuki Ota and David A. Agard*

Howard Hughes Medical Institute and the Department of Biochemistry and Biophysics University of California at San Francisco, San Francisco CA 94143-0448, USA

**Corresponding author*

We describe multi-conformation simulated annealing-pseudo-crystallographic refinement (MCSA-PCR), a technique developed for predicting the binding mode of a flexible ligand in a flexible binding pocket. To circumvent the local-minimum problem efficiently, this method performs multiple independent cycles of simulated annealing with explicit solvent, "growing" the ligand in the binding pocket each time. From the ensemble of structures, a pseudo-crystallographic electron density map is calculated, and then conventional crystallographic refinement methods are used to best fit a single, optimal structure into the density map. The advantage of the MCSA-PCR method is that it provides a direct means to evaluate the accuracy and uniqueness of the calculated solution, provides a measure of ligand and protein dynamics from the refined *B*-factors, and facilitates comparison with X-ray crystallographic data. Here, we show that our MCSA-PCR method succeeds in predicting the correct binding mode of the VSV8 peptide to the major histocompatibility complex (MHC) receptor. Importantly, there is a significant correlation between the experimentally determined crystallographic water molecules and water density observed in the pseudo map by MCSA-PCR. Furthermore, comparison of different approaches for extracting a single, most probable structure from the calculated ensemble reveals the power of the PCR method and provides insights into the nature of the energetic landscape.

Keywords: binding; simulated annealing; refinement; molecular dynamics; multi-conformation

Introduction

Ligand-protein interactions are central and ubiquitous phenomena in a variety of biological processes from enzymatic catalysis to signal transduction. A theoretical understanding of ligand-protein interactions is essential for understanding the mechanism of a biological system and for searching for new drugs against a wide variety of diseases.[1–3] A major step toward understanding the interactions between proteins and their ligands will be to reliably and accurately predict the conformation of both the ligand and the binding pocket. To date, various approaches have been developed to tackle this "docking problem".[4] The earliest approach was to treat both ligand and protein as rigid bodies and use rotation and translation searches around the binding pocket to predict the appropriate binding mode. This approach is exemplified in early versions of the DOCK programs.[5] With significant advances in computer technology, an extension of the rigid method to a flexible method[6–13] has become available recently and shows impressive improvements in binding-mode predictions and faster screening for new lead compounds.

Despite recent advances in flexible approaches, there are still many difficulties to overcome.[14] First,

Abbreviations used: MCSA, multi-conformation simulated annealing; PCR, pseudo crystallographic refinement; MHC, major histocompatibility complex; VSV8, vesicular stomatitis virus nucleoprotein (52-59) RGYVYQGL; SEV9, Sendai virus nucleoprotein (324-332) FAPGNYPAL; MD, molecular dynamics; NOE, nuclear Overhauser effect; rmsd, root-mean-squared deviation; CPU, central processing unit; TCR, T-cell receptor.

E-mail address of the corresponding author: agard@msg.ucsf.edu

0022-2836/01/030607–11 $35.00/0

the accuracy of the binding modes predicted by those methods relies heavily on the accurate estimation of ligand-protein interaction energies, or their scoring functions.[15,16] Usually, those scoring functions are somewhat simplified when compared to the empirical functions used in molecular dynamics (MD) simulations and represent a trade-off between accuracy and computational efficiency. Second, complete side-chain rotamer sampling of flexible ligands in flexible binding pockets is still problematic.[10,14] When a ligand is docked into a modeled binding pocket, the conformation of the complex is easily trapped into one of the local minimum conformations close to the initial model. Even with simulated annealing at high temperatures, it is still difficult to sample all possible rotamers, because steric clashes between the ligand and the protein prevents transitioning from one rotamer to another. Third, it is known that in some cases, water molecules play a critical role in determining the conformation of a ligand in a binding pocket by mediating interactions with the protein.[17,18] Fourth, it is possible that a part of the ligand may still be flexible or mobile even in the bound state.[18] Most of the current prediction methods cannot give information on the dynamics of the ligand within the binding pocket.

A feasible way to overcome the problems listed above is to grant flexibility to both ligand and protein using a standard MD simulation[19,"20] with a more accurate force-field.[6–15] However, MD simulations are computationally very intensive and often cannot sufficiently sample conformational space to cover the range of binding modes, due to multiple local minima. Here, we present a new computational approach for tackling the docking problem with a flexible ligand in a flexible binding pocket: binding prediction using a combination of multiple simulated annealing and pseudo crystallographic refinement (MCSA-PCR). Finding the correct binding mode for a ligand-protein complex essentially requires that the conformation corresponding to the global energetic minimum be found. To overcome the local minima problem, in the MCSA-PCR method we adopted a ligand-growing procedure combined with simulated annealing. Growing out ligand side-chains (or other functional groups) during the simulation provides better sampling of ligand conformational space and leads to a better prediction of the binding mode. Although, in theory, simulated annealing[21] should converge to the global optimum for a simulation of infinite duration, in practice, simulated annealing is not guaranteed reach to the global minimum.[22] To help solve this problem, we repeated the simulated annealing with the side-chain growth protocol 100 times using random starting velocities. From the ensemble of 100 annealed structures, an averaged set of pseudo-crystallographic structure factors were calculated and Fourier transformed to generate a pseudo electron density map, providing a probabilistic picture of the simulation solution space. The pseudo map provides valuable dynamics information that cannot be obtained by most standard docking procedures. Furthermore, the maps enable us to make direct comparisons to experimental X-ray crystallographic density maps. Finally, a single representative structure was determined by fitting the model into the averaged pseudo electron density map. This was done by refining a single model against the set of averaged pseudo structure factors using standard X-ray crystallographic simulated annealing methodologies.[22] A concept similar to that underlying the MCSA-PCR method had been used for predicting GCN4 coiled coil structure.[23] However, in our MCSA-PCR method we expand the prediction method by adding molecular dynamics simulations with the molecular growth performed in a bath of explicit solvent. Furthermore, X-ray crystallographic concepts are applied more rigorously, so that direct comparisons between predictions and experimental data are facilitated. The MCSA-PCR method is not aimed at screening as many ligands as possible, but at predicting the binding of a flexible ligand in the context of a flexible binding site as accurately as possible. Therefore, the MCSA-PCR method would prove useful for the refinement of pre-modeled ligand-protein complexes in the later stage of binding mode predictions and as a prelude to free-energy perturbation binding energy calculations.

Here, we apply the new method to determine the binding mode of major histocompatibility complex (MHC) class I H-2 K^b complex with a viral peptide derived from vesicular stomatitis virus nucleoprotein(52-59), RGYVYQGL, VSV8. The crystal structure has been solved at 2.3 Å resolution.[24] Class I MHC molecules bind short peptide fragments (eight to ten residues) derived from intracellular proteins and display them on infected cells so that the T-cell receptors (TCR) on $CD8^+$ cytotoxic T-lymphocytes can recognize them and discriminate foreign from self.[24,25] The structural conformation of the peptide/MHC complex has been a very important target immunologically, as well as a challenging target for a flexible docking problem.[26] However, we chose this system primarily as a test case for developing a general methodology to predict binding of flexible ligand-receptor complexes, not specifically to predict the peptide/MHC complex.[27–31] Thus, procedures that have been highly optimized for just this system complex were not pursued in this study.

To make our tests rigorous, we tried to predict the binding mode of the complex starting with randomized conformations of the peptide and conformations of the receptor in which orientations of side-chains that interact with the ligand have been randomized. The peptide and MHC conformations were randomized separately by heating to 2000 K. Then, the peptide and the receptor were put back together as a complex and were used as the starting point for our prediction test cases. The experimental conformations of the ligand and the side-chains of the receptor are completely lost in our starting model (Figure 1(a)), making it quite challenging to predict the binding mode of the complex accurately. We carried out the prediction of the complex under three distinct scenarios at three levels of difficulty. The first scenario is the case where the locations of all backbone atoms of the peptide are previously known. Second, the locations of only the CA atoms are known. Lastly, nothing is known precisely, but there is approximate information on the locations of the N and C termini. We report the results obtained by using the MCSA-PCR method under these three scenarios and discuss the validity and applicability of the MCSA-PCR methods for the general problem of predicting the binding of flexible ligands to flexible receptors.

Results and Discussion

Overall procedure

The general steps of the MCSA-PCR prediction method are illustrated in Figure 1(a) to (e). First, the bond lengths of the side-chains of the conformationally randomized ligand (Figure 1(a)) are reduced to 0.3 Å and the side-chain interactions are eliminated by turning off the van der Waals and electrostatic interactions (Figure 1(b)). The shrunken ligand is then gradually heated to 2000 K and equilibrated for 10 ps (Figure 1(b)). To obtain a possible binding mode of the ligand, simulated annealing is carried out by growing the shrunken side-chains back to the original size with full energetic interactions. During simulated annealing, the temperature of the system is gradually cooled from 2000 K to 300 K. The van der Waals and electrostatic interactions are also returned proportionally to their original values. After cooling to 300 K, the intramolecular and intermolecular interactions of the ligand are checked. If the intra and inter-atomic energies are extremely high (>0 kcal mol^{-1}, >1000 kcal mol^{-1}, respectively (1 cal = 4.184 J)), indicating that this particular structure has a steric clash between the peptide and MHC or that the peptide takes a physically impossible conformation, the model is eliminated. Otherwise, the annealed model is stored as one of the possible binding structures. The processes above are repeated until 100 structures are obtained (Figure 1(c)). Then, an averaged pseudo electron density map is calculated by averaging amplitudes and phases calculated from each of the 100 annealed structures (Figure 1(d)). Finally, to obtain the single structure to fit the model to the pseudo map, crystallographic simulated annealing refinement targeting the pseudo structure factors is used."[22] The final refined model is presented as the

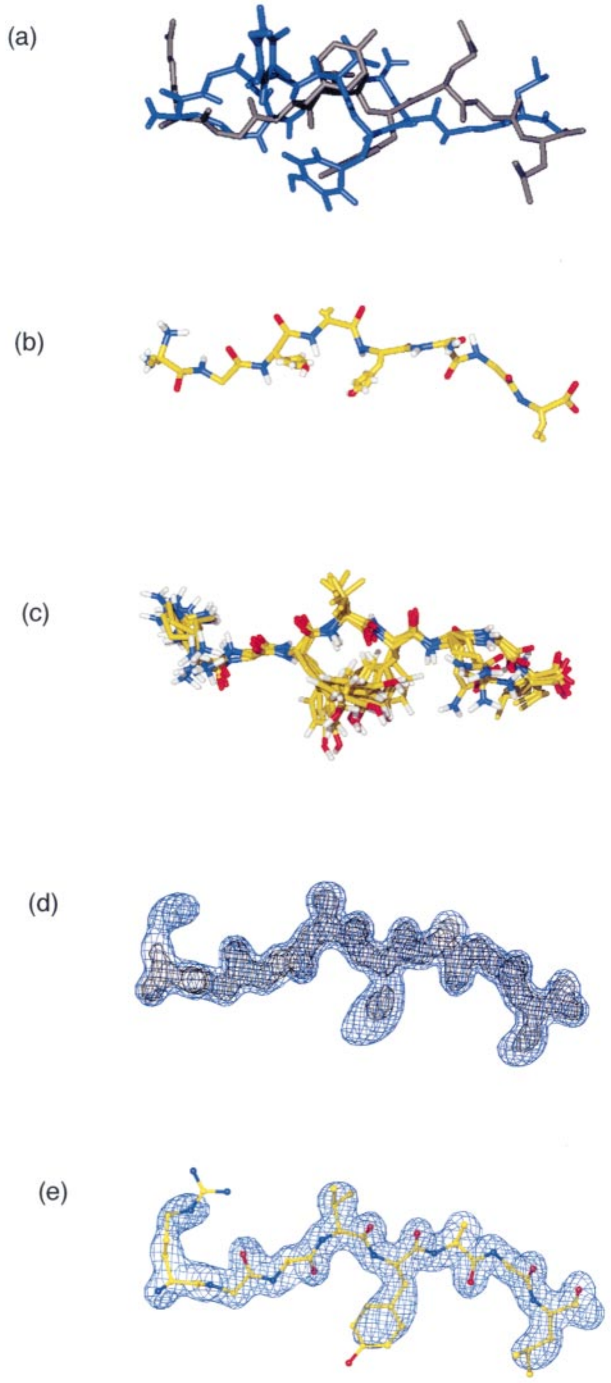

Figure 1. Procedure for multi-conformation, simulated annealing pseudo crystallographic refinement (MCSA-PCR). (a) The randomized VSV8 structure is shown in blue with the superimposed X-ray structure in gray. The MHC protein is not shown. (b) The VSV8 peptide with shrunken side-chains is shown as a stick model. (c) Collected annealed structures of the VSV8 peptide. Only five of the 100 annealed structures are shown here, for clarity. (d) Pseudo electron density maps around the VSV8 peptide. The map was calculated by Fourier transform using the averaged 100 annealed structures. The 1 σ and 0.5 σ maps are shown in black and blue, respectively. (e) A single representative structure of the VSV8 with the pseudo map. The final structure was fit into the pseudo map by using crystallographic refinement targeting the pseudo structure factors. The refined structure is shown as a ball and stick model with the 0.5 σ pseudo electron density map. Figures with the electron density maps were generated using Molray.[43]

predicted binding structure of a peptide/MHC complex, as shown in Figure 1(e).

Validation analyses

In the prediction and evaluation processes, the following three scenarios are considered. The first case is one in which the backbone conformation of the ligand is known, say from a series of previous structural studies on related compounds. For example, this situation would apply to serine proteases such as α-lytic protease with different inhibitors.[32] In those cases, little deviation is observed in the backbone atoms of the various inhibitors. Therefore, only the side-chain conformations need to be scrutinized to find the accurate binding mode. Thus, for this test, the backbone atoms of VSV8 peptide were restrained to the X-ray positions during the simulation. The second scenario is where only limited information about the binding conformation is known. In this case, the location of only a small number of atoms would be specified, such as the C^{α} atoms in a peptide ligand. For instance, this could apply when only a limited number of nuclear Overhauser effect (NOE) distance restraints around the binding pocket are available to specify the binding conformation due to experimental difficulties associated with the ligand's weak affinity. [33] The last scenario is where only an approximate location of the ligand is known, while the detailed conformations of both ligand backbone and side-chains are unknown. In this scenario, only the ends of the peptide (N and C termini atoms) are weakly restrained, so that the ligand can be kept within the binding pocket and yet all the ligand atoms can move freely within the pocket. This would exemplify the situations seen in MHC/peptide complexes where the approximate location of the anchor residues are often known.[24,25] For each scenario, we performed MCSA-PCR procedures under various conditions for the predictions of the complex conformation. First, the distance-dependent dielectric constant is set to $\varepsilon = 4r$ *in vacuo*. Although there is no physical justification for this dielectric approximation, it is still in widespread use for simulations *in vacuum*. Second, the dielectric constant is set to 1 *in vacuo*. Third, the dielectric constant is set to 1 with explicit TIP3 water solvent. Lastly, the standard simulated annealing without the MCSA-PCR procedure was performed as a control.

The overall results obtained by all methods above are listed in Table 1. For evaluation of the various methods, the root-mean-squared deviations (rmsd) of backbone and side-chain are plotted in Figure 2(a) and (b), respectively. The starting model and the final refined models determined by MCSA-PCR methods are shown along with the X-ray structure in Figure 3. As would be expected in scenario 1, all MCSA-PCR methods either *in vacuo* or in solution end up with very small backbone rmsd values (<0.3 Å; Table 1, Figure 2(a)), since the backbone atoms are

Table 1. The rmsd (Å) of predicted VSV8 peptides in the three distinct scenarios and the standard simulated annealing

	$E = 4r$[a]	$E = 1$[b]	Solvent[c]	Standard[d]
Backbone atoms restraints				
Backbone	0.24	0.23	0.27	0.85
Side-chain	2.41	2.30	1.78	4.04
CA atoms restraints				
Backbone	0.65	0.62	0.55	1.01
Side-chain	2.61	2.64	1.95	3.73
NC atoms restraints				
Backbone	0.96	1.08	0.76	2.24
Side-chain	2.84	2.88	2.15	3.84

[a] MCSA-PCR method *in vacuo* with dielectric constant set to 4*r* in a distance-dependent manner.

[b] MCSA-PCR method *in vacuo* with dielectric constant set to 1.

[c] MCSA-PCR method in explicit TIP3 water molecules with dielectric constant set to 1.

[d] Standard simulated annealing method *in vacuo* as a control.

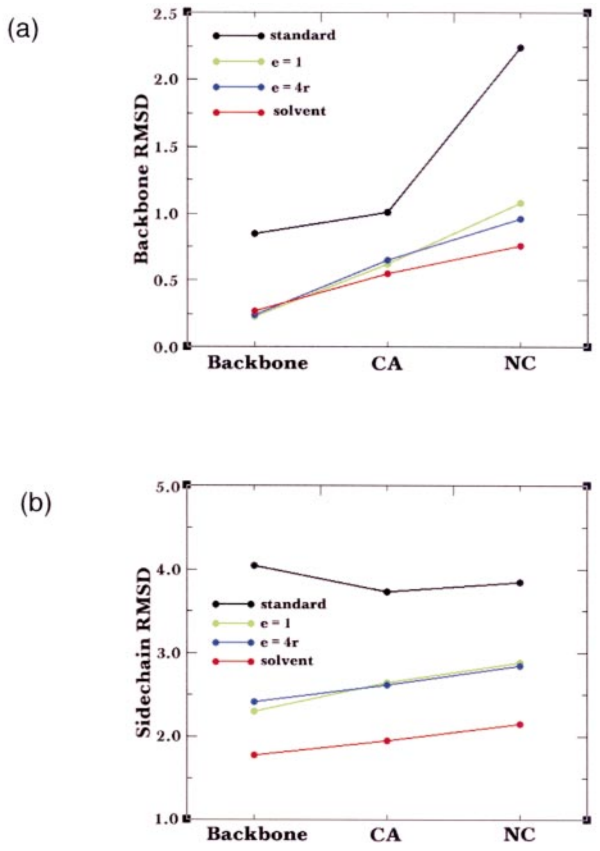

Figure 2. Root-mean-squared deviation (rmsd) of VSV8 structures predicted by various methods in three distinct scenarios. (a) The backbone atom rmsd values of VSV8 are plotted for the method using a bath of TIP3 water molecules (red), for the method having a dielectric constant of 1.0 (green), for the method having a distance-dependent dielectric constant (blue), and for the standard simulated annealing protocol (black). (b) Side-chain atom rmsd values of VSV8 are plotted for methods using solvent (red), dielectric constant of 1.0 (green), distance-dependent dielectric constant (blue), and standard simulated annealing (black).

restrained to the X-ray positions. The slightly larger rmsd of the solvent method than those of vacuum methods appears to result from the larger deviation due to water's interaction with the peptide. However, in the case of N and C termini restraints (scenario 3), the explicit solvent method shows the best results, indicating that the solvent play an important role in determining and stabilizing the peptide/MHC complex. Overall, all MCSA-PCR methods are superior to the standard simulated annealing method.

The side-chain rmsd values show an interesting trend (Figure 2(b)). Again, the MCSA-PCR method with explicit solvent shows the best results for all scenarios. For the *in vacuo* simulations, whatever the dielectric constants used, the rmsd values are always larger than that of the solvent method by at least 0.7 Å (Figure 2(b)). The errors of the *in vacuo* simulation are observed mainly in the long polar side-chains at Arg1 and Gln6. In the absence of solvent, these side-chains bend over to the surface of the receptor to optimize van der Waals interactions (Figure 3(a)). These errors are reduced in the solvent method so that the long polar side-chains are correctly exposed to the solvent as the X-ray structure has suggested (Figure 3(c)-(e)). The results obtained by standard simulated annealing clearly illustrate the side-chain sampling problems of this method. In the standard method, the side-chain conformations are easily trapped within one of the possible rotamers. Even with high-temperature simulations, it is still difficult to sample other rotamers, due to the steric clashes between the ligand and the receptor. Consequently, the standard simulated annealing method shows much larger side-chain rmsd values (>3.5 Å) (Table 1). Interestingly, the rmsd of the side-chains by the standard method in scenario 1 is the largest among all scenarios. It is probably due to the fact that the backbone atoms are restrained more closely to the X-ray positions at the sacrifice of accuracy of the side-chain rotamer conformations. For instance, the side-chains of Tyr3 and Tyr5 were incorrectly exposed to the solvent instead of being buried inside the pocket (Figure 3(b)).

For all scenarios, the MCSA-PCR method with explicit water molecules demonstrated the best results (Table 1 and Figure 3(c)-(e)). The prediction of the flexible ligand by MCSA-PCR method can converge within the range of 0.8 Å rmsd for backbone atoms and ca 2.0 Å rmsd for side-chains with respect to X-ray structures (Table 1), starting from a ligand and a binding pocket with the randomized side-chain conformations (Figure 1(a)). If

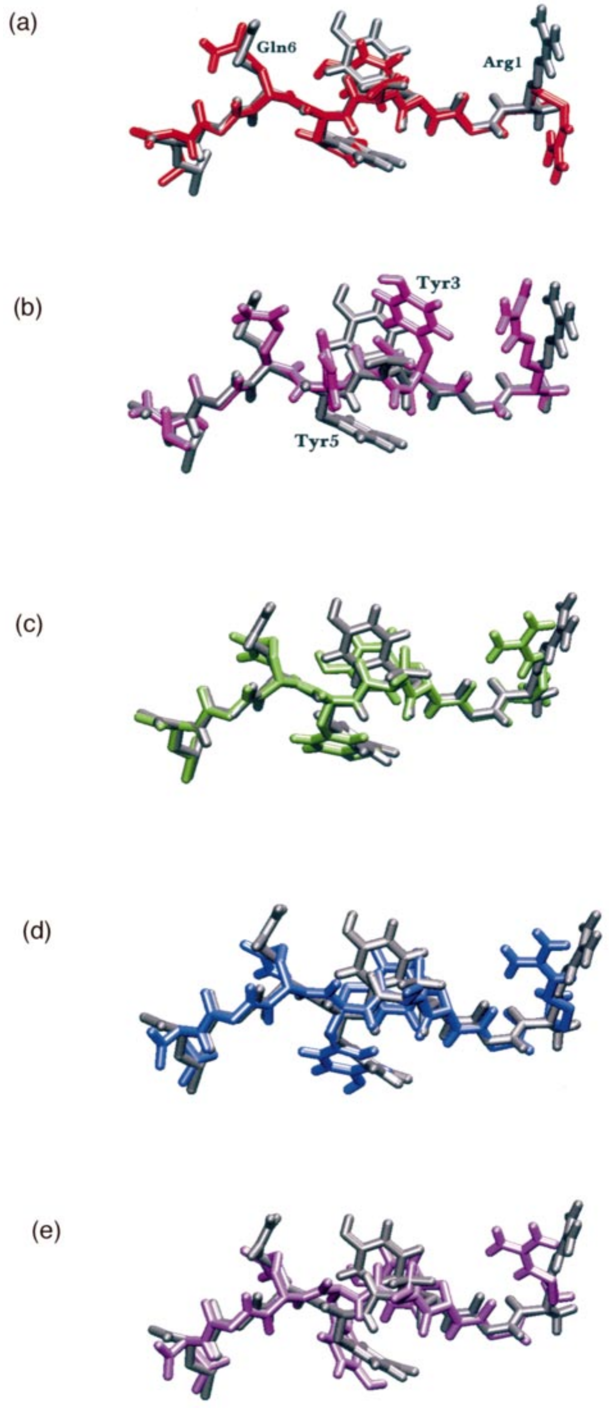

Figure 3. Graphic representations of the predicted VSV8 structures. (a) The refined structure by MCSA-PCR with unit dielectric constant *in vacuo* in scenario 1, where backbone atoms are restrained. The refined VSV8 structure is shown in red with the X-ray structure for comparison. The major errors were observed at the long polar side-chain of Arg1. (b) The refined structure by a standard simulated annealing protocol in scenario 1 is shown in purple. The side-chains are mispositioned by this method, especially at buried residues such as Tyr4 and Tyr6. (c)-(e) The refined structures by MCSA-PCR in explicit solvent. The structures in scenarios 1, 2, and 3 are shown in green, blue, and purple, respectively. All structures are shown with the X-ray structure (gray) for comparison. The use of explicit solvent reduced the error of the side-chain conformation of Arg1 and correctly exposed it to the solvent. The difference at the guanidino group of Arg1 between the predicted model and the crystal structure is reasonable, because that region of the electron density was invisible both in the pseudo map and the experimental map.

additional conformational information, such as NOE data, are available, better results could be obtained as shown in either scenarios 1 or 2 (Table 1).

Pseudo electron density map and dynamics Information

Perhaps the most useful feature of the MCSA-PCR method is that pseudo electron density maps become available, which are directly comparable to experimentally determined electron density maps. A standard molecular dynamics (MD) simulation at 300 K usually fails to sample a complete conformational space within a feasible simulation duration (~nanoseconds) due to multiple local minima. However, our use of a large collection of independent simulations is able to overcome most multiple local minima within a system. Coordinate information from all the simulations is collected and visualized by means of the pseudo electron density maps, which represent the overall landscape of the simulation solution space. As shown below, these maps can also provide direct insights into dynamics and disorder of the system being modeled.

Examination of the pseudo maps of MHC/peptide complexes suggests that we can obtain information about the flexibility or dynamics of the ligand and protein. Using the different contour levels of the density maps (Figure 1(d)), it is possible to identify which part of a ligand is more flexible than others, as evidenced by its exploration of multiple conformations during the simulation. The stable or most populated part of the ligand should be seen clearly in the high-contour level of maps. As shown in the 1σ map (where σ corresponds to the standard deviation of the map, Figures 1(d) and 4(a)), mainly the backbone atoms of the ligand can be seen, indicating that the backbone atoms are very stable in the receptor. In the lower-density maps (Figures 1(d) and 4(a)), most of the side-chain conformations become visible, suggesting that the side-chains are more flexible and mobile than the backbone atoms in general. If a portion of the ligand is extremely mobile, the density cannot be seen even in the low σ level. For example, the guanidino group of Arg1, which is exposed to the solvent, appears to be so mobile that part of the map is completely invisible, as shown in Figure 4(a). It is important to note that the guanidino group of Arg1 was also invisible in the experimental electron density map.[24] Thus, this pseudo electron density map provides faithful insights into the flexibility and dynamics of both ligand and protein in the same manner as found in experimental crystallographic maps.

Prediction of bound water

Remarkably, there was a significant correlation between the water density in the pseudo map

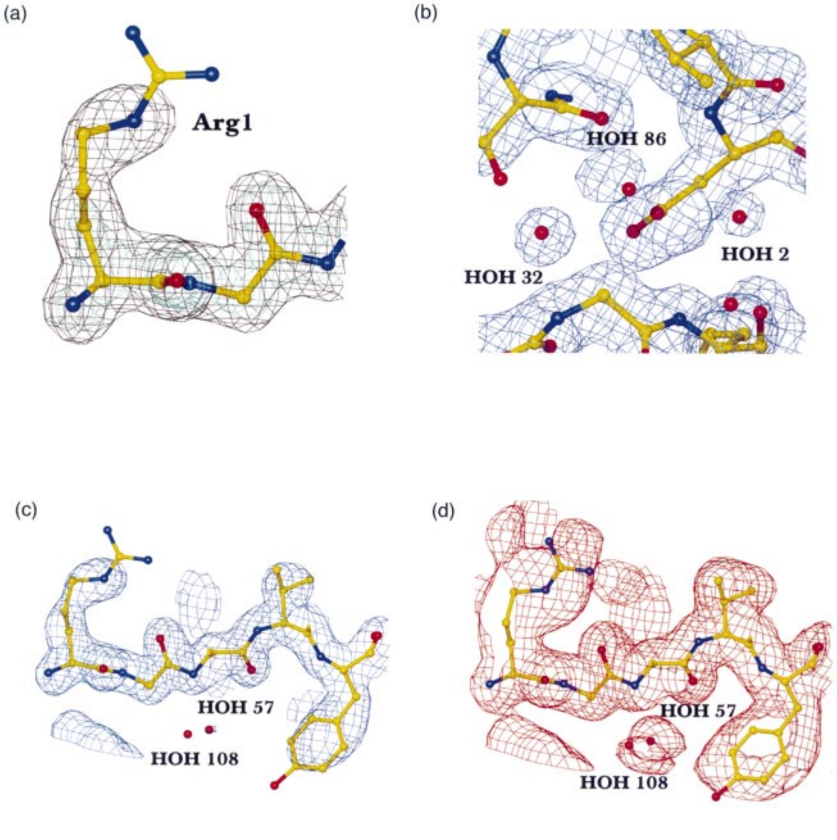

Figure 4. The refined structure with the pseudo electron density map. (a) The refined atoms of VSV8 near Arg1 are shown with the pseudo map. The 1 σ and 0.5 σ maps are shown in cyan and black, respectively. The top of the guanidino group of Arg1 is not visible, suggesting that it is very flexible and mobile. The experimental electron density map shows the similar invisibility around Arg1 as predicted by the PCR method. (b) The water pseudo density shown in the 0.5 σ pseudo map. The pseudo map is shown with the X-ray structure and the crystallographic water molecules. The three water molecules, HOH2, HOH32, and HOH86 are predicted correctly as a water density in the pseudo map. (c) The buried water density in the 1 σ pseudo map. At the 1 σ level, HOH 57 is barely predicted, while HOH108 is not observed. (d) The buried water prediction in the 0.5 σ pseudo map. The HOH57 and HOH108 molecules are clearly observed as bulky density, suggesting that the water molecules are very mobile at those positions. The Figures were generated by Molray.[43]

using the MCSA-PCR method and the location of the bound water molecules determined by X-ray crystallography. From the pseudo maps in scenario 1 (Figure 4(b)-(d)), we could predict accurately the location of 12 of the 16 experimentally observed bound water molecules around the peptide. Two of the four unobserved water molecules (HOH110 and HOH111) were totally exposed to the bulk solvent (data not shown). The ordered water molecules were observed at various σ levels, suggesting various stabilities and residence times (Figure 4(b)-(d)). Water molecules such as HOH32 (Figure 4(b)) buried in the deep groove of the MHC receptor were observed in the high-σ maps, indicating high stability, or long residence time. For water molecules that are either close to the surface or appear to be exposed to the bulk water, such as HOH86, density was not observed in the high-σ map, but could be seen as a bulky density in the low-σ maps (Figure 4(b)). This indicates that HOH86 is more mobile than HOH32 and has a shorter residence time at the location.

The MCSA-PCR method could even predict the existence of water molecules buried under the ligand as shown in Figure 4(c) and (d). The successful prediction of the buried waters is attributed to the slow growth of peptide side-chains. In the early stage of the annealing, the water molecules can sneak into the open space between the ligand and the protein without interference from peptide side-chain atoms. In the later stages of the annealing, the water molecules located at the side-chain positions will be occluded by the growing side-chains. However, it is still possible that the water molecules sometimes hinder the side-chains from being positioned correctly. To eliminate unrealistic conformations due to mislocated water molecules during the iterative simulated annealing process, an energy filter was added, since those water molecules usually experience steric clashes and high-energy conformations (see Materials and Methods).

Interestingly, there were two buried water molecules (HOH42 and HOH26) that were not observed in the pseudo maps (data not shown). Although the reason for this discrepancy is unknown, it may be due to the differences between the conditions of the system, the simulation in the solution phase and the crystalline state in the X-ray crystallography."[24] It is possible that the discrepancy resulted from the errors in the simulations. Further studies will be necessary to clarify this discrepancy and to improve the prediction of the ordered molecules.

The importance of pseudo crystallographic refinement using multiple conformations

While our experiments clearly demonstrate that running multiple simulations provides a powerful mechanism for exploring the energetic landscape, several different approaches can be envisioned by which a single optimal structure is obtained from the calculated ensemble. Here, we propose the novel approach of building a population density map and then using pseudo crystallographic simulated-annealing refinement to best fit the distribution. Other choices would be choosing the lowest calculated energy solution, the mean solution or some Boltzmann-weighted solution. To address this issue, we plot the rmsd of each of the 100 annealed structures with respect to the X-ray coordinates *versus* the calculated energy of that structure. This was done for the tightly constrained case, the fixed backbone case (scenario 1, Figure 5(a)) and the fully flexible loosely tethered case (scenario 3, Figure 5(b)). The calculated energy included the interaction energy between the ligand and the protein as well as the self-energy within the ligand. Note the quite broad distribution in energies and conformations, with the most challenging case (scenario 3) showing the greatest spread in rmsd values (Figure 5(b)). For scenario 1, the lowest energy conformation had an rmsd of 2.30 Å, while the mean rmsd from the 100 structures was 2.23 Å, and a Boltzmann-weighted mean was 2.24 Å. Because the calculated enthalpies are so large, it was necessary to re-scale *kT* to a larger value (50 kcal/mol) to avoid having the lowest energy totally dominate the average. By contrast, the MCSA-PCR-refined structure had the much lower rmsd of 1.92 Å. Similarly for scenario 3, the MCSA-PCR method produced a significantly lower rmsd (2.37 Å) than for the lowest energy structure (2.55 Å), the mean structure (2.67 Å) or the Boltzmann-weighted structure (2.70 Å).

The large energetic and rmsd dispersion evidenced in Figure 5(a) and (b) again underscores the need to have broad conformational sampling. What was perhaps most surprising, was how much better the PCR method performed compared to either choosing the lowest energy conformation or to averaging a number of the lowest energy conformations (*via* the Boltzmann average). We believe that this is revealing important features of the energetic landscape and the nature of the errors in the calculations. First, there is obviously sufficient error in the calculations such that there is only very limited correlation between energy and rmsd. Second, averaging structures can be problematic. For example, in cases where the structures have notably different conformations, such as different side-chain rotamers, the average structure can be the worst choice. By contrast, the PCR method would refine to the most probable of the conformations. Third, from these results, we envision an energetic landscape that contains both broad and narrow regions of low energy (Figure 5(c)). Although the narrow region may represent the global energy minimum, we know from the high rmsd that it is in error. By contrast, the broad regions will be more highly sampled in our 100 simulations and hence will dominate the PCR refinement. Thus, our data suggest that sampling density in a region of configurational space is a better metric of error than is energy. PCR provides

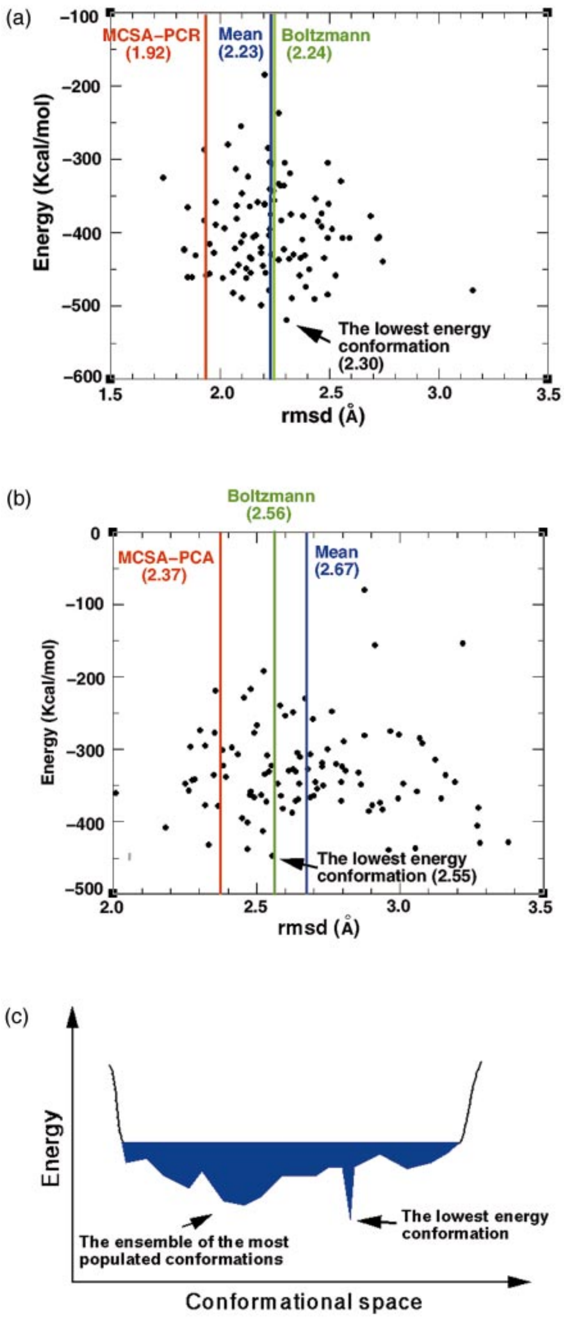

Figure 5. Scatter plot of the energy of each annealed structure *versus* the rmsd between the annealed structure and the X-ray structure. The 100 annealed structures are shown as a black dot. The rmsd values of the single representative structures derived by various methods, MCSA-PCR, mean, and scaled Boltzmann-weighted average are shown as a red line, a blue line, and a green line, respectively. (a) In the case of scenario 1, the backbone atoms of VSV8 peptide were positionally restrained to the X-ray structure during the annealing process. The plot indicates little correlation between the rmsd value and the energy. It is important to note that the single optimal structure derived by MCSA-PCR (red line) is the best representative structure of the ensemble structures compared with others, the lowest energy structure (arrow), the mean structure (blue line) and the scaled Boltzmann-weighted structures (green line). (b) In the case of scenario 3, only the N and C termini were weakly restrained to the X-ray structure. Although the plot shows the broader distribution, the MCSA-PCR structure (red line) is the closest to the X-ray structure. Other methods do not perform as well as MCSA-PCR to extract the single representative structure using the same 100 ensemble structures. (c) Schematic energy profile around the conformational space of the flexible binding mode. The blue region indicates the accessible conformational space of the binding state at room temperature. In this hypothetical energy profile, the most dominant structure may not be the lowest energy conformation because of its narrow conformational space. To predict the binding mode of the flexible ligands accurately, it appears to be necessary to find the most populated structure, which is obtained by considering each contribution by all possible configuration. Our test cases demonstrate that MCSA-PCR is the best method to derive a single representative structure from the ensemble structure.

an efficient means to extract the most probable structure from the ensemble average.

The MCSA-PCR method as a refinement method for accurate binding mode prediction

In conclusion, the MCSA-PCR method shows several advantages for predicting a binding mode of a flexible ligand in a flexible binding pocket. The simulated annealing with the side-chain growth method can circumvent the multiple local minima problem and enhance the sampling performance for the side-chain rotamers, leading to a more accurate prediction of the side-chain conformations than is possible using a standard simulated annealing protocol. The pseudo electron density maps provide a graphical representation of ligand and receptor conformations as seen in conventional X-ray crystallographic maps. Map contour levels reveal the statistical information from the simulation can faithfully reproduce dynamics information for both the ligand and protein. Through the pseudo maps and the pseudo structure factors, we can determine a single representative conformation, which is directly comparable to the X-ray structure. Since the experimentally observable data in X-ray crystallography is not refined structures but diffraction data or structure factors, correlation coefficient analysis or *R*-free analysis between experimentally obtained structure factors and pseudo structure factors from MCSA-PCR can be envisioned. The direct comparable nature of MCSA-PCR to the experimental data may provide more accurate information to improve current force fields and better methodologies used in MD.

Besides the advantages above, the primary advantage of the MCSA-PCR method is, in fact, its flexibility and expandability. It is possible to manipulate the growth method for only a part of the side-chains of interest or to expand it to include the backbone atoms of either ligand or protein.

When other experimental information, such as NOE data, is available, it can be incorporated readily into the refinement of binding predictions.[33] The MCSA-PCR method is not designed either for screening a large number of compounds in a short time or for searching the binding pocket of a ligand but, instead, is aimed at most accurately predicting the binding mode of a flexible ligand in a known binding pocket, providing dynamics information and locating ordered water positions between ligands and proteins. Although in this study we focused on predicting the binding mode of a peptide to MHC as a test case, in general, the MCSA-PCR procedure can be applicable for various studies such as non-peptide drug design and protein-protein or protein-DNA interactions. Therefore, this method should be useful as a final refinement procedure in the later stage of binding mode prediction, homology modeling, and drug/vaccine designs.

Materials and Methods

Force field and stochastic boundary molecular dynamics

The OPLS force field[34] with polar hydrogen atoms was used for the MHC protein and the VSV8 peptide. The TIP3P model[35] was used for the simulated annealing with solvent water molecules. The polar hydrogen atoms were added to the X-ray structure by HBUILD.[36] We performed four distinct simulations under three different scenarios to find the best approach to predict the ligand binding mode. For *in vacuo* simulations, all MHC side-chain atoms within 15 Å from any VSV8 peptide atom were free to move and backbone atoms were weakly restrained (10 kcal mol^{-1} Å^{-2}). For the solvated simulation, we used the stochastic boundary condition.[37] The simulation consisted of a spherical region with a radius of 14 Å centered on C$^{\alpha}$ atom of Val4. The size of the spherical solvent region is rather small but is sufficient to cover the entire peptide and all side-chain atoms of MHC around the binding pocket. Water molecules within 2 Å of any heavy atom of the MHC and the VSV8 peptide were deleted. During the overlay of water molecules onto the system, no previous knowledge about crystal water was used. Inside the sphere (''reaction region''), standard molecular dynamics simulations were carried out. Protein atoms outside of the sphere (''reservoir region'') are rigid.

The Verlet algorithm[38] was used to integrate the equation of motion. Temperature coupling[39] was used to keep the temperature to the target temperature from 2000 K to 300 K during the simulated annealing. The coupling constant was 100 ps^{-1}. With the SHAKE algorithm,[40] the geometry of each water molecule was kept rigid and the bond lengths of the VSV8 peptide and MHC were kept constant. A time-step of 0.5 fs was used for the entire simulated annealing.

Preparation of initial coordinates

To evaluate the accuracy and validity of the MCSA-PCR method, a randomized peptide/MHC complex was prepared as a starting structure for our test cases. The VSV8/MHC complex was taken from the Protein Data Bank (accession code 2VAA). Initially, all crystal water molecules were eliminated and then VSV8 and MHC were separated. The VSV8 peptide was heated to 2000 K for 10 ps *in vacuo* with N and C termini atoms being restrained. By this high-temperature simulation, the side-chains and backbone atoms are basically randomized. For the MHC binding pocket, the side-chains were randomized in a similar fashion, yet the backbone atoms were restrained to the X-ray structure, since the MHC backbone structure with a different bound peptide (SEV9/MHC) is nearly identical with that of the VSV8/MHC complex.[24] The randomized VSV8 peptide was put directly back into the randomized MHC binding pocket. Then, the obvious atomic clashes in the complex were corrected manually without any knowledge of the X-ray complex structure. The randomized structure was minimized briefly (100 steps) and thermalized gradually from 10 K to 2000 K and equilibrated for 1.0 ps. This equilibrated complex was always used as a starting structure for all side-chain growth methods and a standard simulated annealing method as a control. The starting VSV8 coordinate is shown in Figure 1(a). The rmsd of backbone atoms and side-chain atoms with respect to the X-ray structure is 2.71 Å and 5.67 Å, respectively.

Shrunken side-chain and simulated annealing with side-chain growth

For the side-chain growth method, an extra heating and equilibration were carried out for the randomized complex above by assigning new parameters. The bond lengths of all of the side-chains were reduced to 0.3 Å. The van der Waals and electrostatic interactions of the side-chains were turned off, but those of backbone atoms remained intact. The shrunken system was heated gradually from 10 K to 2000 K and was equilibrated for 1.0 ps. We used the simulated annealing method to find a global minimal conformation of the complex. The 2000 K equilibrated system was gradually cooled by 25 K in every 50 steps from 2000 K to 300 K. After reaching 300 K, an extra 100-step equilibration was carried out. The final coordinate sets without an extra energy minimization were stored as one of the possible conformations of the complex. It is known that even though the simulated annealing method is a powerful tool to find a global minimum, there is no guarantee of reaching the global minimum each time. Also, it is possible that the system may have multiple stable conformations at 300 K. Therefore, multiple trials by simulated annealing and the ensemble average should give a better representation of the complex structure. We repeated this simulated annealing method with side-chain growth 100 times. Every time simulated annealing was repeated, new randomly drawn velocities from a Maxwell-Boltzmann distribution are assigned to each atom of the system. All simulations were carried out using the program X-PLOR[41] running on an SGI R10,000. The CPU times for single simulated annealing in case of vacuum and solution were about 25 minutes and 35 minutes, respectively.

Pseudo structure factors and pseudo electron density map generation from 100 annealed structures

The pseudo electron density map was calculated from the 100 ensemble conformations generated during the simulated annealing with side-chain growth. The averaged pseudo electron density maps were calculated as follows. First, the pseudo structure factors, F_{pseudo}, were calculated by using the following equation:

$$F_{\text{pseudo}}(\mathbf{h}) = \sum_i Q_i f_i(\mathbf{h}) \exp(-B_i((\Gamma\mathbf{h})^2/4)) \times \exp(2\pi i \mathbf{h}\Gamma(\mathbf{r}))$$

where Q_i is the occupancy and B_i is the individual atomic temperature factor for atom i. We set the occupancy for each atom to unity. Each atomic temperature factor is assigned to 15 Å^2. In our prediction method, the system has no symmetry, in other words, it is treated as $P1$ with a single molecule in an asymmetric unit. The $P1$ unit cell was designed to encapsulate the entire MHC/peptide complex with an extra cushion space (6 Å for each side). The unit cell size was $a = 63.891$ Å, $b = 80.919$ Å, $c = 62.003$ Å. The MHC/peptide complex was centered in the unit cell. Γ is the 3×3 matrix operator that converts orthogonal coordinates to fractional coordinates. Γ^* is the transpose of Γ. The term f_i is the atomic scattering factor (f) for atom type i obtained from the International Tables for Crystallography.[42]

Then, the structure factors were averaged from the ensemble of 100 structures generated during the course of the simulated annealing with side-chain growth:

$$\mathbf{F}_{\text{avepseudo}} = \langle \mathbf{F}_{\text{pseudo}} \rangle$$

Where $\langle\ \rangle$ is the ensemble average generated by the simulated annealing method. The averaged electron density maps were then calculated by fast Fourier transformation using the ensemble-averaged amplitudes and phases:

$$\rho_{\text{pseudo}} = \frac{1}{V} \sum_{hkl} |\mathbf{F}_{\text{avepseudo}}| \times \exp[-2\pi i(hx + ky + lz) + i\alpha_{\text{avepseudo}}]$$

where ρ_{pseudo} is an average pseudo electron density, V is the volume of the unit cell, and $\alpha_{\text{avepseudo}}$ is the averaged phase corresponding to the structure factors.

Generation of a representative structure through pseudo crystallographic refinement

The pseudo electron density map shows the statistical information of the possible binding modes of the peptide within the MHC. In other words, the probability of the location of each residue of the peptide can be represented by the contours of the map. Therefore, finding the most probable binding mode of the peptide is equivalent to finding the best fit of the peptide into the pseudo map. This is fundamentally the same task as solved in crystallographic refinement. We adopted a standard crystallographic refinement method[22] for the model fitting process.

A standard crystallographic refinement is, in essence, a search for the global minimum of a target function. The target function is composed of the chemical potential energy of the system, E_{chem}, and the experimental penalty energy resulting from the difference between the model and the experimental data (E_{data}):

$$E = E_{\text{chem}} + w_{\text{data}} \sum_{hkl} (|\mathbf{F}_{\text{avepseudo}}| - k|\mathbf{F}_{\text{calc}}|)^2$$

where hkl are the indices of the reciprocal lattice points of the crystal, and k is a relative scale factor. The weighting factor, w_{data}, was chosen so that the forces arising from E_{data} and E_{chem} can be balanced against each other.[22] Automated protocols to obtain initial estimates for optimal weighting as implemented in X-PLOR[41] were used for this study.

In order to find the global minimum of the target function, we used simulated annealing with side-chain growth as shown above. The same annealing procedure as seen in the side-chain growth was used, except imposing the extra E_{data} function in the refinement. After the simulated annealing, an extra 100-step minimization was performed to optimize the model fitting to the pseudo map. As shown in Figure 1(e), the refined model was best fit into the electron density map after the simulated annealing process. During the annealing process, the water molecules were not included in the system.

References

1. Kuntz, I. D., Meng, E. C. & Shoichet, B. K. (1994). Structure-based molecular design. *Accts Chem. Res.* **27**, 117-123.
2. Whittle, P. J. & Blundell, T. L. (1994). Protein structure-based drug design. *Annu. Rev. Biophys. Biomol. Struct.* **23**, 349-375.
3. Lybrand, T. P. (1995). Ligand-protein docking and rational drug design. *Curr. Opin. Struct. Biol.* **5**, 224-228.
4. Bamborough, P. & Cohen, F. E. (1996). Modeling protein-ligand complexes. *Curr. Opin. Struct. Biol.* **6**, 236-241.
5. Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R. & Ferrin, T. E. (1982). A geometric approach to macromolecular-ligand interactions. *J. Mol. Biol.* **161**, 269-288.
6. Wang, J., Kollman, P. A. & Kuntz, I. D. (1999). Flexible ligand docking: a multistep strategy approach. *Proteins: Struct. Funct. Genet.* **36**, 1-19.
7. Jones, G., Willett, P., Glen, R. C., Leach, A. R. & Taylor, R. (1997). Development and validation of a generic algorithm for flexible docking. *J. Mol. Biol.* **267**, 727-748.
8. Miller, M. D., Kearsley, S. K., Underwood, D. J. & Sheridan, R. P. (1994). FLOG: a system to select ''quasi flexible'' ligands complementary to a receptor of known three-dimensional structure. *J. Comput. Aided Mol. Des.* **8**, 153-174.
9. Rarely, M., Kramer, B., Lengauer, T. & Klebe, G. (1996). A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **261**, 470-489.
10. Rosenfeld, R., Vajda, S. & DeLisi, C. (1995). Flexible docking and design. *Annu. Rev. Biophys. Biomol. Struct.* **24**, 677-700.
11. Leach, A. R. (1994). Ligand docking to proteins with discrete side-chain flexibility. *J. Mol. Biol.* **235**, 345-356.
12. Gehlhaar, D. K., Verkhivker, G. M., Rejto, P. A., Sheman, C. J., Fogel, D. B., Fogel, L. J. & Freer, S. T. (1995). Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: conformationally flexible docking by evolutionary programming. *Chem. Biol.* **2**, 317-324.
13. Oshiro, C. M., Kuntz, I. D. & Dixon, J. S. (1995). Flexible ligand docking using a genetic algorithm. *J. Comput. Aided Mol. Des.* **9**, 113-130.

14. Carlson, H. A. & McCammon, J. A. (2000). Accommodating protein flexibility in computational drug design. *Mol. Pharmacol.* **57**, 213-218.
15. Zacharias, M., Luty, B. A., Davis, M. E. & McCammon, J. A. (1994). Combined conformational search and finite-difference Posson-Boltzmann approach for flexible docking. *J. Mol. Biol.* **238**, 455-465.
16. Totrov, M. & Abagyan, R. (1994). Detailed ab initio prediction of lysozyme-antibody complex with 1.6 Å accuracy. *Nature Struct. Biol.* **1**, 259-263.
17. Levitt, M. & Park, B. H. (1993). Water: now you see it, now you don't. *Structure,* **1**, 223-226.
18. Ota, N., Stroupe, C., Ferreira-da-Silva, J. M. S., Shah, S. A., Mares-Guia, M. & Brunger, A. T. (1999). Non-Boltzmann thermodynamic integration (NBTI) for macromolecular systems: relative free energy of binding of trypsin to benzamidine and benzylamine. *Proteins: Struct. Funct. Genet.* **37**, 641-653.
19. Karplus, M. & Petsko, G. A. (1990). Molecular dynamics simulations in biology. *Nature,* **347**, 631-639.
20. Rastelli, G., Thomas, B., Kollman, P. A. & Santi, D. V. (1995). Insight into the specificity of thymidylate synthase from molecular dynamics and free energy perturbation calculations. *J. Am. Chem. Soc.* **117**, 7213-7227.
21. Kirkpatrick, S., Gelatt, C. D. & Vecchi, M. P., Jr (1983). Optimization by simulated annealing. *Science,* **220**, 671-680.
22. Brunger, A. T., Karplus, M. & Petsko, G. A. (1989). Crystallographic refinement by simulated annealing: application to a 1.5 Å resolution structure of crambin. *Acta Crystallog. sect. A,* **45**, 50-61.
23. Delano, W. L. & Brunger, A. T. (1994). Helix packing in proteins: prediction and energetic analysis of dimeric, trimeric, and tetrameric GCN4 coiled coil structures. *Proteins: Struct. Funct. Genet.* **20**, 105-123.
24. Fremont, D. H., Matsumura, M., Stura, E. A., Peterson, P. A. & Wilson, I. A. (1992). Crystal structures of two viral peptides in complex with murine MHC class I H-2 K^b. *Science,* **257**, 919-927.
25. Bjorkman, P. J., Saper, M. A., Samraoui, B., Bennett, W. S., Strominger, J. L. & Wiley, D. C. (1987). Structure of the human class I histocompatibility antigen, HLA-A2. *Nature,* **329**, 506-512.
26. Rammensee, H. G. (1997). Chemistry of peptides associated with MHC class I and class II molecules. *Curr. Opin. Immunol.* **7**, 85-96.
27. Rognan, D., Reddehase, M. J., Koszinowski, U. H. & Folkers, G. (1992). Molecular modeling of an antigenic complex between a viral peptide and a class I major histocompatibility glycoprotein. *Proteins: Struct. Funct. Genet.* **13**, 70-85.
28. Rognan, D., Scapozza, L., Forker, G. & Daser, A. (1995). Rational design of nonnatural peptides as high-affinity ligands for the HLA-B*2705 human leukocyte antigen. *Proc. Natl Acad. Sci. USA,* **92**, 753-757.
29. Gulukkota, K., Sidney, J., Sette, A. & DeLisi, C. (1997). Two complementary methods for predicting peptides binding major histocompatibility complex molecules. *J. Mol. Biol.* **267**, 1258-1267.
30. Froloff, N., Windemuth, A. & Honig, B. (1997). On the calculation of binding free energies using continuum methods: application to MHC class I protein-peptide interactions. *Protein Sci.* **6**, 1293-1301.
31. Zhang, C., Anderson, A. & DeLisi, C. (1998). Structural principles that govern the peptide-binding motifs of class I MHC molecules. *J. Mol. Biol.* **281**, 929-947.
32. Bone, R., Silen, J. L. & Agard, D. A. (1989). Structural plasticity broadens the specificity of an engineered protease. *Nature,* **339**, 191-195.
33. Maurer, M. C., Trosset, J.-V., Lester, C. C., DiBella, E. E. & Scheraga, H. A. (1999). New general approach for determining the solution structure of a ligand bound weakly to a receptor: structure of a fibrinogen Aα-like peptide bound to thrombin(S195A) obtained using NOE distance constraints and an ECEPP/3 flexible docking program. *Proteins: Struct. Funct. Genet.* **34**, 29-48.
34. Jorgensen, W. L. & Tirado-Rives, J. (1987). The OPLS potential functions for proteins. Energy minimizations for crystals of crystals of cyclic peptides and crambin. *J. Am. Chem. Soc.* **110**, 1657-1666.
35. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. (1983). Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926-935.
36. Brunger, A. T. & Karplus, M. (1988). Polar hydrogen positions in proteins: empirical energy function placement and neutron diffraction comparison. *Proteins: Struct. Funct. Genet.* **4**, 148-156.
37. Brunger, A. T., Brooks, C. L., III & Karplus, M. (1985). Active site dynamics of ribonuclease. *Proc. Natl Acad. Sci. USA,* **82**, 8458-8462.
38. Verlet, L. (1967). Computer experiments on classical fluids. I. Thermodynamic properties of Lennard-Jones molecules. *Phys. Rev.* **159**, 98-105.
39. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., Dinola, A. & Haak, J. R. (1984). Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81**, 3684-3690.
40. Ryckaert, J.-P., Ciccotti, G. & Berendsen, H. J. C. (1977). Numerical-integration of cartesian equations of motion of a system with constraints - molecular dynamics of *N*-alkanes. *J. Comput. Phys.* **23**, 327-341.
41. Brunger, A. T. (1993). *X-PLOR Version 3.1: A system for X-ray Crystallography and NMR*, Yale University Press, New Haven, CT.
42. Hahn, T. (1983). *International Tables for Crystallography*, vol. A, Dordrecht Kluwer Academic Publisher, Hingham, MA.
43. Harris, M. & Jones, T. A. (2001). Molray - a web interface between O and the POV-Ray ray tracer. *Acta Crystallog. sect. D,* **57**, 1201-1203.

Edited by J. Thornton

(Received 12 June 2001; received in revised form 1 October 2001; accepted 1 October 2001)