Article

Structural Mining: Self-Consistent Design on Flexible Protein–Peptide Docking and Transferable Binding Affinity Potential

Zhijie Liu, Brian N. Dominy, and Eugene I. Shakhnovich

J. Am. Chem. Soc., **2004**, 126 (27), 8515-8528• DOI: 10.1021/ja032018q • Publication Date (Web): 19 June 2004

J|A|C|S
ARTICLES
Published on Web 06/19/2004

Structural Mining: Self-Consistent Design on Flexible Protein−Peptide Docking and Transferable Binding Affinity Potential

Zhijie Liu, Brian N. Dominy, and Eugene I. Shakhnovich*

Contribution from the Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, Massachusetts 02138

Received December 31, 2003; E-mail: eugene@belok.harvard.edu

Abstract: A flexible protein−peptide docking method has been designed to consider not only ligand flexibility but also the flexibility of the protein. The method is based on a Monte Carlo annealing process. Simulations with a distance root-mean-square (dRMS) virtual energy function revealed that the flexibility of protein side chains was as important as ligand flexibility for successful protein−peptide docking. On the basis of mean field theory, a transferable potential was designed to evaluate distance-dependent protein−ligand interactions and atomic solvation energies. The potential parameters were developed using a self-consistent process based on only 10 known complex structures. The effectiveness of each intermediate potential was judged on the basis of a *Z* score, approximating the gap between the energy of the native complex and the average energy of a decoy set. The *Z* score was determined using experimentally determined native structures and decoys generated by docking with the intermediate potentials. Using 6600 generated decoys and the *Z* score optimization criterion proposed in this work, the developed potential yielded an acceptable correlation of $R^2 = 0.77$, with binding free energies determined for known MHC I complexes (Class I Major Histocompatibility protein HLA-A*0201) which were not present in the training set. Test docking on 25 complexes further revealed a significant correlation between energy and dRMS, important for identifying native-like conformations. The near-native structures always belonged to one of the conformational classes with lower predicted binding energy. The lowest energy docked conformations are generally associated with near-native conformations, less than 3.0 Å dRMS (and in many cases less than 1.0 Å) from the experimentally determined structures.

Introduction

Molecular docking is widely used in modern drug discovery, and many approaches, such as DOCK[1−3] and AutoDock,[4−6] have been developed for evaluating protein−small molecule interactions. Full consideration of complex flexibility, especially ligand flexibility, is a common feature of current docking methods. In recent years, protein−protein docking has drawn significant attention, and some popular methods have been developed which were mainly based on geometric or chemical complementarity with respect to an inflexible protein.[7−13] While much attention has been paid to these areas of study, the intermediate challenge of protein−polypeptide docking is often neglected. One reason is simply the problem of classification, where protein−peptide docking is often grouped into either protein−small molecule or protein−protein docking according to peptide size. Another reason lies in the significant computational difficulties due to the flexibility of peptides and proteins, both of which should be addressed. Considering the low toxicity, synthetic accessibility, and other potentially useful features of polypeptides, it is important to study protein−peptide interactions and address the challenge of flexibility inherent to these systems.

In molecular docking, evaluating the binding affinity between the protein and ligand accurately and rapidly remains a principal challenge. Traditional force fields in molecular mechanics (MM) evaluate free energy using several techniques, such as free energy perturbation (FEP), thermodynamic integration (TI), etc.[14−17] Unfortunately, significant computational requirements prevent broad application of these techniques for lead screening. Alternatively, the development of empirical scoring functions has been found to be a practical compromise and has been used

(1) Ewing, T. J. A.; Kuntz, I. D. *J. Comput. Chem.* **1997**, *18*, 1175−1189.
(2) Ewing, T. J. A.; Makino, S.; Skillman, A. G.; Kuntz, I. D. *J. Comput. Aided Mol. Des.* **2001**, *15*, 411−428.
(3) Meng, E. C.; Shoichet, B. K.; Kuntz, I. D. *J. Comput. Chem.* **1992**, *13*, 505−524.
(4) 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.
(5) Morris, G. M.; Goodsell, D. S.; Huey, R.; Olson, A. J. *J. Comput. Aided Mol. Des.* **1996**, *10*, 293−304.
(6) Dominy, B. N.; Brooks, C. L., III. *Proteins* **1999**, *36*, 318−331.
(7) Zacharias, M. *Protein Sci.* **2003**, *12*, 1271−1282.
(8) Heifetz, A.; Eisenstein, M. *Protein Eng.* **2003**, *16*, 179−185.
(9) Ben-Zeev, E.; Eisenstein, M. *Proteins* **2003**, *52*, 24−27.
(10) Gardiner, E. J.; Willett, P.; Artymiuk, P. J. *Proteins* **2003**, *52*, 10−14.
(11) Dominguez, C.; Boelens, R.; Bonvin, A. M. *J. Am. Chem. Soc.* **2003**, *125*, 1731−1737.
(12) Chen, R.; Li, L.; Weng, Z. *Proteins* **2003**, *52*, 80−87.
(13) Chen, R.; Weng, Z. *Proteins* **2003**, *51*, 397−408.
(14) Ajay; Murcko, M. A. *J. Med. Chem.* **1995**, *38*, 4953−4967.
(15) Goodford, P. J. *J. Med. Chem.* **1985**, *28*, 849−857.
(16) Searle, M. S.; Williams, D. H.; Gerhard, U. *J. Am. Chem. Soc.* **1992**, *114*, 10697−10704.
(17) Kollman, P. *Chem. Rev.* **1993**, *93*, 2395−2417.

in exploring protein−ligand interactions, using quantitative structure−activity relationships (QSAR), three-dimensional QSAR such as comparative molecular field analysis (CoMFA), and especially expanded master equation methods (ME).[14,16−18] Using these methods, one can construct a scoring function that exhibits a good correlation with known experimental binding free energies as well as determined or computed physiochemical properties from known structures in the training set.[19−25] These approaches generally perform well in closely related protein−ligand systems rather than being universally applicable.[19,20] Other scoring approaches use knowledge-based methods, which adopt mean field theory (MF) to derive an averaged interaction potential based on the statistical distribution of structural features among protein−ligand complexes in a training set.[26−31] Because of statistical limitations, this type of potential has difficulties in representing distance-dependent interactions, such as electrostatic energy.[26−28] Following the original SMoG method,[27] Muegge et al. approached this problem using a large number of training structures.[25] Many of these approaches have been explored in drug discovery. However, their efficiency and universality depend on the size and structural diversity of the training set. In particular, the master equation method also requires the experimental binding free energies of each structure within the training set. As for the protein−peptide complexes studied here, these methods will be difficult to apply because of the limited availability of determined structures and corresponding experimental binding free energies. Our group proposed a *Z* score optimization approach to potential development with applications to protein folding.[30,32] Based on the large number of decoys generated from a few typical structures, *Z* score optimization can generate a transferable potential directly applicable to protein−peptide docking with possible applications for binding free energy prediction.

In this work, we have developed a flexible docking method using a Monte Carlo annealing simulation. This approach is rooted in an earlier Monte Carlo-based approach to folding in fully atomic and flexible protein models.[33] As opposed to the majority of docking approaches, the method proposed here considers not only the flexibility of the ligand but also that of the protein. To represent short-range dispersion and long-range electrostatic interactions, our approach is based on a distance-dependent potential, which is then parametrized by *Z* score optimization. The optimization algorithm is self-consistent, evaluating the *Z* score of each intermediate potential on the basis of decoys generated with the potential obtained in the previous optimization steps. The optimized potential is found to yield a high correlation with binding free energies determined for known MHC I complexes that are not in the training set (Class I Major Histocompatibility protein HLA-A*0201). When the optimized potential is applied to protein−peptide docking, a significant correlation between energy and distance root-mean-square (dRMS) of the generated conformations is observed, and native-like conformations can be identified in most cases.

Methods

Structure Sets. We systematically searched the protein data bank (PDB) (release April 2001, no. 96) and found 443 entries for protein−peptide complexes. All structures were classified using structural classification of proteins (SCOP), and redundant structures were removed, leaving 87 complexes. An additional selection criterion was then applied, eliminating peptides with unnatural amino acids, complexes involving small molecules or metal ions in the binding site, and complexes with other structural defects. In the end, 25 structures remained, and few of them had experimental binding free energies available.

Ten from the remaining 25 protein−peptide complex structures were selected as the training set to optimize the potential. These complexes comprise a diverse set, as they belong to different protein families and are involved in different disease processes (Table 1). The peptide ligands also have different lengths, and their composition consists of all 20 natural amino acids. All these structures are high-resolution X-ray structures (<2.3 Å), with the exception of one NMR structure that is related to an important apoptosis process (PDB ID 1bxl). The remaining 15 structures comprise the testing set.

Classification of Atom Types. The coordinates of hydrogen atoms in native structures are seldom determined from X-ray crystallographic experiments. In our docking process, hydrogens were treated as one atom type and were only considered in removing clashes but not in energy evaluation. Because our potential mainly considers a solvent-accessible surface-based solvation energy and distance-dependent contact energies, which are related to electrostatic interactions, van der Waals interactions, and salt-bridge and hydrogen-bond interactions, the atom types in the method will be classified by their atomic numbers, partial charges, and van der Waals radii. All of these parameters are taken from the parm99 parameter set of the AMBER 7.0 program.[34−36] The partial charge is the principal factor in classification due to its significant variability. On the basis of a cutoff of 0.2 charge unit, all atoms including hydrogen atoms and metal ions were classified into 12 atom types (supplement A, Supporting Information). Metal ions not involved in binding were not considered. Neglecting hydrogens and metal ions, this left 10 remaining atom types and a total of 400 ($10 \times 10 \times 4$) parameters to be optimized within the energy function.

Flexible Molecular Docking Method. (a) Monte Carlo Annealing Process. A Monte Carlo annealing simulation protocol was adopted in our flexible docking process.[37] The initial temperature was set to be 3% of the energy of the native structure, the cooling rate was 0.992, the initial acceptance rate was 0.750, and the simulation was not terminated until the energy difference between the two nearest accepted steps converged to less than 0.0001 of the total energy for 10 000 times

(18) Bohm, H. J. *J. Comput. Aided Mol. Des.* **1994**, *8*, 243−256.
(19) Rognan, D.; Lauemoller, S. L.; Holm, A.; Buus, S.; Tschinke, V. *J. Med. Chem.* **1999**, *42*, 4650−4658.
(20) Logean, A.; Sette, A.; Rognan, D. *Bioorg. Med. Chem. Lett.* **2001**, *11*, 675−679.
(21) Wang, R. X.; Liu, L.; Lai, L. H.; Tang, Y. Q. *J. Mol. Model.* **1998**, *4*, 379−394.
(22) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. *J. Comput. Aided Mol. Des.* **1997**, *11*, 425−445.
(23) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. *J. Mol. Biol.* **1996**, *261*, 470−489.
(24) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. *J. Mol. Biol.* **1997**, *267*, 727−748.
(25) Muegge, I.; Martin, Y. C. *J. Med. Chem.* **1999**, *42*, 791−804.
(26) DeWitte, R. S.; Ishchenko, A. V.; Shakhnovich, E. I. *J. Am. Chem. Soc.* **1997**, *119*, 4608−4617.
(27) DeWitte, R. S.; Shakhnovich, E. I. *J. Am. Chem. Soc.* **1996**, *118*, 11733−11744.
(28) Shimada, J.; Ishchenko, A. V.; Shakhnovich, E. I. *Protein Sci.* **2000**, *9*, 765−775.
(29) Luo, H.; Sharp, K. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 10399−10404.
(30) Mirny, L. A.; Shakhnovich, E. I. *J. Mol. Biol.* **1996**, *264*, 1164−1179.
(31) Jiang, L.; Gao, Y.; Mao, F.; Liu, Z.; Lai, L. *Proteins* **2002**, *46*, 190−196.
(32) Vendruscolo, M.; Mirny, L. A.; Shakhnovich, E. I.; Domany, E. *Proteins* **2000**, *41*, 192−201.
(33) Shimada, J.; Kussell, E. L.; Shakhnovich, E. I. *J. Mol. Biol.* **2001**, *308*, 79−95.
(34) Wang, J.; Cieplak, P.; Kollman, P. *J. Comput. Chem.* **2000**, *21*, 1049−1074.
(35) Cieplak, P.; Caldwell, J.; Kollman, P. *J. Comput. Chem.* **2001**, *22*, 1048−1057.
(36) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179−5197.
(37) Binder, K.; Heermann, D. W. *Monte Carlo simulation in statistical physics: an introduction*; Springer-Verlag: Berlin, 1992.

Table 1. Database of Protein−Peptide Complexes

PDB ID	resolution/ Å	protein	ligand	length	exposure degree/%
1a30[a]	2.00	aspartic protease−hiv-1 protease	EDL	3	16.2
1awq[a]	1.58	isomerase−cyclophilin a	HAGPIA	6	53.9
1be9[a]	1.82	third pdz domain from the synaptic protein psd-95	KQTSV	5	40.7
1bxl[a]	NMR	apoptosis- bcl-xl	GQVGRQLAIIGDDINR	16	40.2
1cka[a]	1.50	oncogene protein-c-crk (*Mus musculus*)	PPPALPPKK	9	58.8
1eg4[a]	2.00	dystrophin ww domain fragment	NMTPYRSPPPYVP	13	61.7
1elw[a]	1.60	chaperone-tpr1-domain of hop	GPTIEEVD	8	48.8
1gux[a]	1.85	transcription reg-retinoblastoma protein	DLYCYEQLN	9	46.4
1ycq[a]	2.30	oncogene protein, *Xenopus laevis* mdm2	ETFSDLWKLLP	11	42.4
2fib[a]	2.10	blood coagulation-human γ-fibrinogen carboxyl-terminal fragment	GPRP	4	32.2
1i31[b]	2.5	exocytosis(μ2 adaptin subunit (ap50) of ap2 clathrin adaptor	FYRALM	6	61.8
2cbl[b]	2.1	proto-oncogene n-terminal domain of cbl	SDGYTPEPA	9	64.8
1g3f[b]	NMR	apoptosis bir3 domain of xiap	AVPI	4	43.8
1io6[b]	NMR	signaling protein growth factor receptor-bound protein 2 (grb2) sh3 domain	RHYRPLPPLP	10	57.8
1ab9[b]	1.6	serine protease (bovine γ-chymotrypsin)	CGVPAIQPVL	10	64.1
1bc5[b]	2.2	methyltransferase	NWETF	5	49.6
1duz[b]	1.8	human class I histocompatibility antigen (hla-a 0201)	LLFGYPVYV	9	26.1
1evh[b]	1.8	mena evh1 domain from murine	FPPPP	5	44.0
1f95[b]	NMR	contractile protein (dynein light chain 8 (dlc8))/apoptosis	MSCDKSTQT	9	44.2
1jhg[b]	1.3	regulatory protein (trp operon repressor mutant v58I)	W	1	37.3
1vwg[b]	1.46	biotin-binding protein (streptavidin)	CHPQGPPC	8	41.0
8tln[b]	1.6	hydrolase (metalloproteinase thermolysin (E.C. 3.4.24.27))	VK	2	39.0
2seb[b]	2.5	hla-dr4 class II histocompatibility antigen	AYMRADAAAGGA	12	38.7
1ce1[b]	1.9	therapeutic antibody campath-1h fab	GTSSPSAD	8	33.2
1pau[b]	2.5	apopain protease	DEVD	4	27.9
1shf[c]	1.9	fyn proto-oncogene tyrosine kinase (*Homo sapiens*)	PPPALPPKK	9	−
2cpl[c]	1.63	cyclophilin a	HAGPIA	6	−

[a] Ten complexes in training set. [b] Fifteen complexes in testing set. [c] Two unbound proteins in testing set.

continuously or the simulation reached the maximum number of 1.5 million steps. To accelerate computational speed, all atoms were mapped into a periodic cubic space, which can accommodate any size protein complex in theory. This cubic space is partitioned by a grid into smaller cubes in order to accelerate energy calculations. The spacing of the internal grid is 6.0 Å, and only the interactions within the same grid or neighboring grids will be counted. The contact distance cutoff between atoms was 6.0 Å, meaning that all interactions occurring past this cutoff are neglected. A simple starting configuration was chosen by separating the protein and ligand, maintaining their bound conformations. Using the Monte Carlo annealing algorithm, the ligand was then docked into the protein binding pocket. The Monte Carlo move set includes ligand rotation, which rotates the ligand in the coordinate space as a rigid body; ligand translation, which translates the ligand in the coordinate space as a rigid body; ligand torsion rotation, which rotates the partial ligand atoms around the randomly selected ligand backbone or side-chain torsion angle; protein side-chain torsion rotation, which rotates the related protein side-chain atoms around the selected protein side-chain torsion angle; and protein backbone torsion rotation, which rotates the partial protein atoms around the selected protein backbone torsion angles. All flexible torsion angles are consistent with the given rotamer library.[27,38] The detailed Monte Carlo process is illustrated in supplement B (Supporting Information).

In the Monte Carlo step involving the protein side-chain torsion rotation, first a ligand neighboring space is defined to cover the grids that ligand occupied and the neighboring grids. One protein residue in this space is then randomly selected, and its side-chain torsion angles are rotated by small random angles within the given thresholds from the rotamer library, respectively. The clash between the related atoms is examined, and the rotation will repeat again if it results in atom clashes in the generated complex. Because protein buried-cores are often highly compacted, most of the accepted protein side-chain torsion rotation after clash-checking is located on the protein surface. The obtained structure is additionally judged by the metropolis criterion based on binding energy evaluation.

(38) Liu, Z.; Jiang, L.; Gao, Y.; Liang, S.; Chen, H.; Han, Y.; Lai, L. *Proteins* **2003**, *50*, 49−62.

For each protein−ligand complex, multiple Monte Carlo docking simulations are started from different random seeds. The dRMSs between generated decoys and the native structure are computed, and a stable distribution is ensured from enough simulations. Additionally, the dRMSs between generated decoys are also computed, and all decoys are clustered into different conformational classes by dRMS with a cutoff of 3.0 Å. The docking simulations continue until no new conformational classes are generated. For most protein−ligand complexes in this work, a total of 100 simulations are enough to ensure that no new conformational classes were generated further and to sample the conformational space thoroughly. Therefore, 100 simulations with different random seeds are conducted for each protein−ligand complex.

(b) Restraint. A harmonic restraint called a surface restraint between the protein and ligand was adopted to keep the ligand in contact with the protein surface. The restraint was applied between the nearest atom pair of the protein and ligand for which the distance is the shortest between any protein−ligand atom pairs, and finally was coarse-grained by determining the distance between the geometric centers of the grids in which the restrained atoms were located (eq 1). d is the distance between the centers of the nearest grid cells, and d_0 is the contact distance cutoff, 6.0 Å.

$$E_{\text{constraint}} = \begin{cases} 0, & d < d_0 \\ (d - d_0)^2, & d \geq d_0 \end{cases} \quad (1)$$

To further accelerate the computational speed, we shrank the conformational space accessible by the ligand by introducing pocket restraints. The restraint was computed between the geometric centers of the protein binding pocket and the ligand (eq 1). d is the distance between the above geometric centers, and d_0 is the distance between the geometric centers in the native complex structure plus a buffer of 8.0 Å. The accessible volume is large enough to include the protein binding pocket as well as neighboring regions. Because the restraint information can be easily obtained from rough protein structural analysis or experimental data, the methods applied in this study will be applicable to "real-world" problems.

From eq 1, we can see that both restraints will be zero and will not contribute to the binding free energy when the ligand approaches the

neighborhood of the protein binding pocket. In other words, these restraints will not influence the energy evaluation nor the conformational preference within the binding pocket.

Energy Functions for Protein−Peptide Binding Affinity Evaluation. (a) A Pairwise Atomic, Distance-Dependent Potential between the Protein and Peptide. Our group has developed a knowledge-based potential, SMoG, which can accurately predict the binding free energies of many protein−small molecule complexes. An obvious defect in the potential is its weakness in considering electrostatic-like interactions, which are highly distance-dependent. To fully characterize all atom-pair interactions, especially the distance-dependent interactions, a distance-dependent energy function is proposed, which is illustrated in eq 2. In our design strategy, R_{ij} is the actual distance between a pair

$$E_{\mathrm{pl-contract}\ ij} = A_{ij}(R_{ij} - B_{ij})^{C_{ij}} + D_{ij} \quad (|A_{ij}| \le 1.0,\ -1 \le C_{ij} \le 2,\ |D_{ij}| \le 1.0) \quad (2)$$

of atoms of types i and j; A_{ij} is the force constant related to the atom pair; B_{ij} is the typical interaction distance between atoms of types i and j; the exponent C_{ij} determines the interaction's distance dependence; and D_{ij} corresponds to a basic packing background. Several constraints were applied to the parameters: B_{ij} was set to be greater than the clash distance (0.75 of the sum of van der Waals radii) and less than the contact distance (1.6 of the sum of van der Waals radii); C_{ij} took the discrete values of −1, 0, 1, or 2, which characterize the electrostatic interaction (−1), van der Waals or packing effect (0, 1), and hydrogen-bond or salt-bridge interaction (2), respectively. These constraints have the obvious benefit of shrinking the accessible solution space, providing a tractable optimization problem. In order to avoid the significant difference between A_{ij}'s and D_{ij}'s, A_{ij} and D_{ij} were normalized together to ensure the parameters have comparable values by iteration optimization—A_{ij}' and D_{ij}' are the corresponding normalized parameters used in the final potential (eqs 3 and 4)—and a boundary condition was established which enforces the pairwise energy (E_{ij}) to be 0 when the distance between the atoms is equal to or greater than a 6.0 Å cutoff.

$$U = \sqrt{\sum_{i,j}(A_{ij}^{\ 2} + D_{ij}^{\ 2})} \quad (3)$$

$$A_{ij}' = A_{ij}/U, \quad D_{ij}' = D_{ij}/U \quad (4)$$

(b) A Solvation Energy Function Based on the Atomic Solvent-Accessible Surface (ASAS). A simplified solvation energy function based on the solvent-accessible surface was adopted, the form of which is similar to the pairwise contact potential (eq 5). The atomic solvent-accessible surface was computed by an approximate analytic method (eq 6),[39] in which the parameters P_i and P_{ij} were refit for the polypeptide systems.

$$E_{\mathrm{sol}\ i} = \begin{cases} A_i \cdot (S_i - B_i)^{C_i} + D_i, & S_i \ge B_i \\ 0.0, & 0.0 \le S_i \le B_i \end{cases} \quad (|A_i| \le 1.0,\ \ 0.0 < B_i < 30.0,\ \ 0 \le C_i \le 2,\ \ |D_i| \le 1.0) \quad (5)$$

$$S_i = T_i \prod_j (1.0 - P_i P_{ij} b_{ij}/T_i) \quad (6)$$

Compared with surface areas computed using the program Naccess,[40] we obtained a high correlation of $R^2 = 0.824$ and a slope of 1.03 for atoms in both native structures and decoy structures. In eq 6, T_i is the theoretical isolated surface area of atom i, S_i is the computed surface after deduction from neighboring atom contacts, P_i is a single-body scaling factor for T_i from atom i's own effect, P_{ij} scales the two-body effect of a neighboring atom j on the accessible surface of atom i, and b_{ij} is the contact surface area between atoms i and j.

(c) Gō Potential for the Internal Energy of the Protein and Ligand. Differences in the protein or ligand internal energy, resulting from movement around torsion angles, were computed using a Gō potential. This ensured that their conformations remained near native, while permitting reasonable flexibility during the docking process.[33,41−43] For most "real-world" docking studies, however, the internal Gō energy of the ligand would be inaccessible because the bound structure is unknown and the peptide ligand is too flexible to assume a single conformation. In many cases, however, the Gō internal energy term does not contribute significantly to the total complexed energy. This suggests that this energy term could be reliably neglected in future docking studies. This argument is more thoroughly described later in this paper. However, for those complexes in which the ligand is known to have a specific bound conformation, for example, the peptide substrate bound to Bcl-X_L that maintains an α-helical structure in complex,[44] a Gō internal energy term may be useful.

The total energy in docking is the sum of the protein−ligand contact energy, the protein and ligand internal energies, and the protein and ligand solvation energies (eq 7). The adjustable parameters A_{ij}, B_{ij}, C_{ij},

$$E_{\mathrm{total}} = E_{\mathrm{pl-contact}} + w_{\mathrm{pg}}E_{\mathrm{p-go}} + w_{1g}E_{\mathrm{l-go}} + w_{\mathrm{ps}}E_{\mathrm{p-sol}} + w_{\mathrm{ls}}E_{\mathrm{l-sol}} \quad (7)$$

and D_{ij} in distance-dependent protein−peptide contact energy and A_i, B_i, C_i, and D_i in solvation energy form the final potential and will be optimized in the following self-consistent Z score optimization process.

Potential Optimization. (a) *Z* Score Optimization. Z score optimization has been successfully used in developing protein folding potentials.[30] This method is based on the simple thermodynamic hypothesis that the native structure of a protein has the lowest energy (or free energy or potential of mean force if solvent degrees of freedom and short-scale motions of the protein are taken into account), proposed by Anfinsen in 1961.[45,46] Here we applied this approach to protein−ligand interactions and assumed that the native conformation of the protein complex is the conformation with the lowest binding free energy. There are two typical Z score functions. One is called the critical Z score (Z_C) and is based on a continuous random energy model (REM), which presumes that energies of both the native structure and decoy structures are random Gaussian variables. The critical Z score is related to the gap between the native energy and the average energy of the decoys (eq 8), where $\sigma(E_i)$ is the standard deviation of the decoy energies, $\langle E_{ci}\rangle$ is the average energy of the decoys, and i refers to one protein−ligand complex. Another Z score function is called the gap Z score (Z_G), which presumes that there is a significant gap between native energy and the lowest energy decoy (eq 9). We have considered the merits of both functions and here propose a combined Z score function (eq 10),

$$Z_{\mathrm{C}\ i} = \frac{E_{\mathrm{native}\ i} - \langle E_{\mathrm{C}\ i}\rangle}{\sigma(E_i)} \quad (8)$$

$$Z_{\mathrm{G}\ i} = E_{\mathrm{native}\ i} - E_{\mathrm{lowest}\ i} \quad (9)$$

$$Z_i = Z_{\mathrm{C}\ i} + w_i\frac{Z_{\mathrm{G}\ i}}{\sigma(E_i)} = \frac{E_{\mathrm{N}\ i} - \langle E_{\mathrm{C}\ i}\rangle}{\sigma(E_i)} + wi\frac{E_{\mathrm{N}\ i} - \langle E_{\mathrm{lowest}\ i}\rangle}{\sigma(E_i)} \quad (10)$$

to ensure a distribution of decoy energies and a significant gap between the native energy and the lowest energy decoy. For multiple protein−ligand complexes, two average Z scores, $\langle Z\rangle_1$ and $\langle Z\rangle_2$, are computed,

(39) Hasel, W.; Hendrickson, T. F.; Still, W. C. *Tetrahedron: Comput. Methodol.* **1988**, *1*, 103−116.
(40) Hubbard, S. J.; Campbell, S. F.; Thornton, J. M. *J. Mol. Biol.* **1991**, *220*, 507−530.
(41) Go, N.; Abe, H. *Int. J. Pept. Protein Res.* **1983**, *22*, 622−632.
(42) Go, N.; Abe, H. *Biopolymers* **1981**, *20*, 991−1011.
(43) Abe, H.; Go, N. *Biopolymers* **1981**, *20*, 1013−1031.
(44) Sattler, M.; Liang, H.; Nettesheim, D.; Meadows, R. P.; Harlan, J. E.; Eberstadt, M.; Yoon, H. S.; Shuker, S. B.; Chang, B. S.; Minn, A. J.; Thompson, C. B.; Fesik, S. W. *Science* **1997**, *275*, 983−986.

which emphasizes a smaller Z_i score gap (eq 11) and a more

$$\langle Z\rangle_1 = \begin{cases} Z_i, & \text{if any } Z_i \geq 0 \\ \dfrac{M}{\sum_{i=1}^{M} 1/Z_i}, & \text{if all } Z_i < 0 \end{cases} \quad (11)$$

representative average Z_i score gap (eqs 12−14), respectively. Z score optimization was performed using a Monte Carlo annealing simulation to parametrize the potential used to distinguish native structures from decoy structures.

$$\langle Z\rangle_2 = \begin{cases} Max(Z_i), & \text{if any } Z_i \geq 0 \\ Z_{\text{avg}}\dfrac{M}{\sum_{i=1}^{M} Q_i}, & \text{if all } Z_i < 0 \end{cases} \quad (12)$$

$$Z_{\text{avg}} = \sqrt[n]{\prod_i^n Z_i} \quad (13)$$

$$Q_i = \begin{cases} Z_i/Z_{\text{avg}}, & (|Z_i| \geq |Z_{\text{avg}}|) \\ Z_{\text{avg}}/Z_i, & (|Z_i| < |Z_{\text{avg}}|) \end{cases} \quad (14)$$

(b) Self-Consistent Potential Optimization Process in an MPI Parallel Package. A self-consistent algorithm written in the C language and using MPI parallel controlling was developed to link decoy generation through a docking method and potential optimization (supplement C, Supporting Information). The whole process has four steps. In the initialization step, the complexes are prepared and their native structures are relaxed to reduce clashes; an initial potential, which can be random or predetermined, is constructed. In the second step, the flexible docking method generates decoys based on the initial potential. In the third step, Z score optimization is used to parametrize the potential within a Monte Carlo annealing process. This process includes perturbation and normalization of the potential by adjustable parameters A_{ij}, B_{ij}, C_{ij}, D_{ij}, A_i, B_i, C_i, and D_i, decoy energy computation, Z score computation, Monte Carlo metropolis evaluation of Z score difference to determine whether the perturbed potential is acceptable, and convergence evaluation, until an improved potential is obtained. Finally, in the last step, the decoy database is augmented with a new set of decoy structures generated using the improved potential. The latter two steps are repeated until the final Z score reaches a predetermined converged value or the number of decoys in the database reaches a threshold. In our work, the initial potential is derived on the SMoG2001 knowledge-based method.[26−28] The initial distance-dependent potential was chosen to match closely the SMoG2001 potential.[47] In the SMoG2001 potential, the interaction energy between two atoms is the sum of interaction energies associated with two distance bins (0.0−3.5 Å and 3.5−4.5 Å). The distance-dependent functional described in this work (eq 2) is fit according to eq 15, where M_{ij} is the

$$E_{\text{pl-contact } ij} = (M_{ij} - N_{ij})(R_{ij} - 3.5^2)^1 + N_{ij} \quad (15)$$

SMoG2001 energy between atom types i and j at a distance of 3.5−4.5 Å, and N_{ij} is the energy between types i and j at a distance of 0.0−3.5 Å. An initial set of 500 decoys was generated, and 150 decoys were added in each round until the convergence criterion of potential was reached. The final potential was chosen for the docking study as described below.

(c) dRMS Virtual Energy Function for Consideration of Protein Flexibility in Docking. While actual docking (see below) will be carried out with the full potential developed by the above self-consistent process, as a first step in evaluating the role of protein flexibility we use dRMS as a virtual energy function. While such an energy function requires knowledge of the final structure of the complexes and as such cannot be used for docking purposes, it is useful for the analysis of the minimal requirements of degree of protein flexibility needed to achieve accurate docking with the more realistic energy function that will be used throughout this study. By computing the distances between all protein−ligand atom pairs, the dRMS is obtained as the root-mean-square deviation of the distances between the docked conformation and the native conformations. The dRMS potential, similar to a Gō potential, assigns the most favorable energy to the native configuration. The energy function increases proportionally to dRMS, providing a smooth funnel landscape. This also provides a long-range attractive potential that is not contained in a typical Gō potential.

Results

Flexibility Consideration in Protein−Peptide Docking. Which aspects of flexibility should be considered in protein−peptide docking? Instead of using the optimized potential, the dRMS virtual energy function was used to study sampling issues related to the importance of flexibility in docking. In the Monte Carlo docking simulation with dRMS, after each movement, the dRMS between the generated conformation and the native structure is computed. The difference in dRMS virtual energy to the last step is judged by the Metropolis criterion to determine whether the current movement is acceptable. Sample complexes from the training set with different ligand degrees of freedom and different degrees of solvent exposure in the binding pocket were selected as test cases (Table 2). Initially, two docking approaches were employed to investigate the importance of flexibility. In the first approach, full ligand flexibility was considered while the protein was kept rigid. It was found that the complexes with a solvent-exposed protein binding pocket (degree of exposure >50%), such as 1cka and 1awq, achieve native-like conformations (dRMS <1.0 Å) with relatively high frequency (>50%). However, the complexes containing a buried protein binding pocket (degree of exposure <40%) had native-like conformations less frequently (frequency <20%), and the average dRMS of the docked conformations was high (about >4.0 Å). In the case of 1bxl, the binding pocket is not significantly buried; however, native-like conformations were rarely observed. The reason for this may be the size and the flexibility of the 16-residue ligand (>15 residues). To achieve more consistent docking results, additional protein flexibility may be necessary.

In the second approach, flexible protein side-chain torsions were introduced in the docking process. The results show that more native-like conformations were generated for all complexes, and the average dRMSs of the docked conformations were significantly reduced. Among the test cases, the complexes with an exposed binding pocket, especially the cyclophilin A complex 1awq, demonstrated significant improvement. The frequency of native-like conformations for this system was higher than 90%, and the resulting docked peptides were structurally similar to one another, with an average dRMS of less than 0.8 Å relative to the native conformation. Although the frequency of native-like conformations in buried binding

(45) Anfinsen, C. B.; Haber, E.; Sela, M.; White, F. H. *Proc. Natl. Acad. Sci. U.S.A.* **1961**, *47*, 1309−1314.
(46) Anfinsen, C. B. *Science* **1973**, *181*, 223−230.
(47) Ishchenko, A. V.; Shakhnovich, E. I. *J. Med. Chem.* **2002**, *45*, 2770−2780.

Table 2. Effect of Flexibility of Protein and Ligand on Docking Results

PDB ID	ligand	length	exposure degree/%	ligand flexibility only: decoys (%) <1.0 Å	ligand flexibility only: decoys (%) <2.0 Å	ligand flexibility only: average dRMS/Å	additional protein side-chain flexibility: decoys <1.0 Å	additional protein side-chain flexibility: decoys <2.0 Å	additional protein side-chain flexibility: average dRMS/Å
1a30	EDL	3	16.2	6.0	8.0	4.45	12.0	38.0	2.86
1awq	HAGPIA	6	53.9	58.0	80.0	1.02	94.0	98.0	0.78
1bxl	GQVGRQLAIIGDDINR	16	40.2	6.0	26.0	3.94	12.0	46.0	3.27
1cka	PPPALPPKK	9	58.8	82.0	100.0	0.42	94.0	100.0	0.31
2fib	GPRP	4	32.2	20.0	24.0	4.26	34.0	42.0	3.17

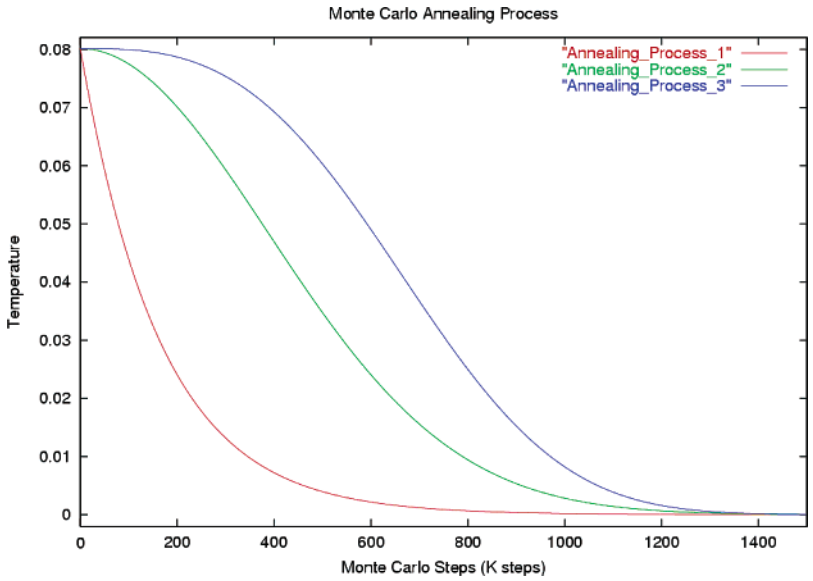

Figure 1. Three kinds of annealing protocols exploited in flexible protein−peptide docking. From curves 1 to 3, the temperature decreasing rates in Monte Carlo annealing docking process ranged from a fixed rate to a more flexible rate, and the system was kept under higher temperature for a longer time.

Table 3. Results from Different Annealing Protocols in Flexible Docking with dRMS Virtual Potential

PDB ID	pocket exposure degree/%	annealing protocol 1: decoys (%) <1.0 Å	annealing protocol 1: average dRMS/Å	annealing protocol 2: decoys (%) <1.0 Å	annealing protocol 2: average dRMS/Å	annealing protocol 3: decoys (%) <1.0Å	annealing protocol 3: average dRMS/Å
1cka	58.8	94%	0.31	100%	0.20	100%	0.18
2fib	32.2	34%	3.17	72%	1.53	90%	0.77

pockets was significantly smaller than that in the exposed binding pockets, protein side-chain flexibility still dramatically improved the docking results (Table 2).

In addition to considering the flexibility of the protein and ligand, making changes to the annealing protocol may be helpful in overcoming the high energy barriers to significantly improve the docking accuracy. Docking with the dRMS virtual scoring function will produce native-like structures much more frequently than with any transferable potential, and therefore provides an upper bound on the expected docking efficiency. In a modified annealing protocol, instead of decreasing the temperature at a constant rate, a variable rate was used in the Monte Carlo annealing docking. This had the effect of maintaining a higher temperature in the beginning stages (Figure 1). The docking results show that the revised annealing process improved the docking accuracy further. For complexes with solvent-exposed binding pockets, such as 1cka, the high frequency of native-like structures was maintained while the average dRMS was reduced. For complexes with buried binding pockets, such as 2fib, the frequency of native-like conformations increased to greater than 60%, and the average dRMS was less than 1.5 Å (Table 3).

This analysis revealed that there were conformational constraints due to the accessibility from the protein binding pocket and the size and flexibility of the peptide. In such docking processes, not only full ligand flexibility but also protein side-chain flexibility should be considered. It may also be necessary to include partial protein backbone flexibility when hinge movements are involved.[48−51] Moreover, successful protein−peptide docking also required an annealing protocol in which the initial high temperature is cooled more slowly in order to allow the system to overcome high energy barriers due to structural constraints.

In fact, docking with our optimized potential also demonstrated similar high energy barriers due to structural constraints. If the docking started from the native bound state, it seldom generated far-native structures which had distinguishable conformational differences and larger dRMSs (>5.0 Å) compared

(48) McCammon, J. A.; Gelin, B. R.; Karplus, M.; Wolynes, P. G. *Nature* **1976**, *262*, 325−326.
(49) Rose, R. B.; Craik, C. S.; Stroud, R. M. *Biochemistry* **1998**, *37*, 2607−2621.
(50) Rose, R. B.; Craik, C. S.; Douglas, N. L.; Stroud, R. M. *Biochemistry* **1996**, *35*, 12933−12944.
(51) Allikas, A.; Ord, D.; Kurg, R.; Kivi, S.; Ustav, M. *Virus Res.* **2001**, *75*, 95−106.

Table 4. Potential Parameters of Ligand Solvation Contribution

atom type	CN	CP	CC	OC	OH	NM	NA	NH	N3	S
A_i	0.0740	−0.0497	0.0000	−0.0430	−0.0211	0.0439	0.0000	0.0000	−0.0717	0.0444
B_i	4.199	13.395	0.000	13.954	15.502	3.404	0.000	0.000	7.412	17.482
C_i	0.5	0.5	0.0	0.5	0.5	0.5	0.0	0.0	1.0	1.0
D_i	−0.0668	0.0152	−0.0478	−0.0323	0.0308	−0.0087	0.0905	−0.0114	0.0001	0.0561

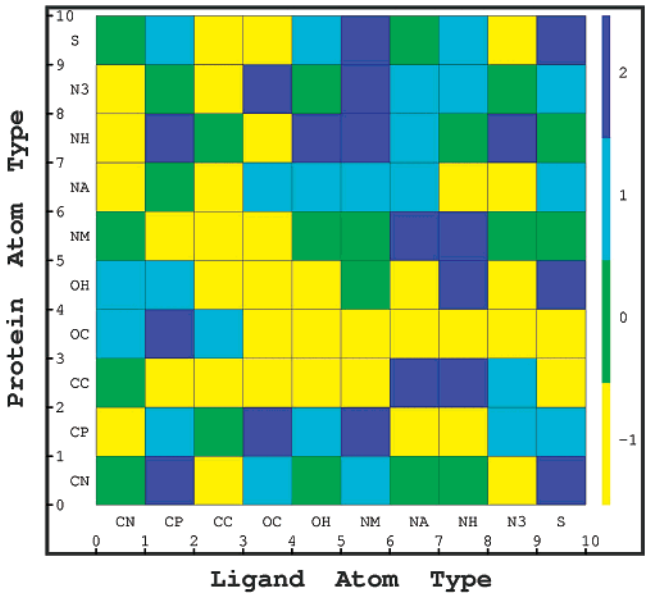

Figure 2. Potential parameters C_{ij} in contact energy between protein−peptide atom pairs. Squares with different colors of blue, cyan, green, and yellow represent different C_{ij} values of 2, 1, 0, and −1, respectively, which also represent different distance dependences of contact interactions between protein−peptide atom pairs, such as harmonic bond interaction, hydrophobic or van der Waals interaction, and Coulomb-like electrostatic interaction.

to native structure, and most of the generated conformations had dRMSs of less than 2.0 Å. Finally, with full flexibility included with regard to the ligand and protein side chains, and using a Monte Carlo annealing process with a variable cooling rate, the docking algorithm and the optimized potential can often generate near-native conformations with dRMSs of less than 3.0 Å.

Potential Analysis. (a) Analysis of Contact Parameters. On the basis of 6600 generated decoy complex structures and 42 intermediate potentials, the final optimized potential was established using the self-consistent procedure described in the Methods. The potential parameters were analyzed to determine whether they reflect well-established physical relationships. Regarding the distance dependence of contact energy parameters, we focused on the parameter C_{ij} describing the functional form of distance dependence (Figure 2). Optimized functionals for atom-pair interactions are physically intuitive. For example, a Coulomb-like distance dependence ($C_{ij} = -1$, $e \propto 1/r$) was found between carboxyl−oxygen and carboxyl−oxygen pairs (square $x_{[4]}y_{[4]}$). Hydrophobic or van der Waals-like interactions between well-separated pairs, showing a weak distance dependence ($C_{ij} = 0$, 1), were found between neutral carbon and neutral nitrogen atom pairs (square $x_{[1]}y_{[1]}$, $x_{[6]}y_{[1]}$, $x_{[1]}y_{[6]}$, $x_{[6]}y_{[6]}$). A bond-like harmonic distance dependence ($C_{ij} = 2$) was detected in pairs representing hydrogen-bond interactions between ligand hydroxyl−oxygens and protein aromatic nitrogens (square $x_{[5]}y_{[8]}$), disulfide bonds between sulfur−sulfur atom pairs (square $x_{[10]}y_{[10]}$), and salt bridges between oppositely charged ligand carboxyl−oxygens and protein sp^3 nitrogen atom pairs (square $x_{[4]}y_{[9]}$). Because the optimized potential is a mean force potential, there are some functional relationships that cannot be clearly associated with simple pair potentials. These C_{ij} parameters enforce distinct distance dependence regarding interactions between various atom types, which may be used to evaluate binding affinity more accurately. The C_{ij} parameters also appear coarsely symmetric with respect to the diagonal line ($x_{[0]}y_{[0]} \rightarrow x_{[10]}y_{[10]}$), implying similar but not identical properties between related atom types located in ligands and proteins, respectively. All optimized potential parameters are listed in supplement D (Supporting Information).

(b) Analysis of the Solvation Contribution. The solvation contribution to the potential is based on a model that is dependent on the atomic surface area. Here we examine whether the parameters determined for this model are physically intuitive (Table 4). The most significant surface dependencies, C_i, are associated with the fully charged nitrogen (atom type N3) and exposed hydrophobic sulfur (atom type S). However, the solvation energies of these two atom types have proportionality constants, A_i, of opposite signs, as expected when comparing polar and apolar solvation. Other atom types demonstrated a relatively weak dependence on the atomic surface area (C_i), especially the buried carbonyl carbon (atom type CC). It was also found that all hydrophobic atom types have positive proportionality constants (A_i) while polar atom types have negative proportionality constants (A_i), again consistent with a physical description of polar solvation.

The physical nature of the solvation terms can be further illustrated by examining the effects of ligand binding to an MHC I (Major Histocompatibility protein class I) protein complex (PDB ID 1duz). The solvation contributions of the free state and the bound state of the peptides were computed from the potential (Figure 3). When the ligand was in the free state, a large positive contribution to the solvation energy came from the hydrophobic residues, such as leucine, tyrosine, phenylalanine, valine, proline, while when the ligand was in the bound state, a smaller positive or even negative contribution was made by the same residues. Therefore, if we consider the solvation energy difference during binding, it demonstrates that the hydrophobic residues will often provide a favorable solvation contribution in binding, consistent with other models of hydrophobic solvation. For the protein−peptide complexes studied here, solvation contributes 25−40% of the energy in binding.

(c) Binding Free Energy Prediction. The binding affinities of five MHC I HLA-A*0201 protein complexes with different peptides were computed using the optimized potential, based on the energy difference between the bound and free states. All complexes have known crystal structures and experimental binding free energies and do not have similar structures in the training set (Figure 4a). An acceptable correlation of $R^2 = 0.770$

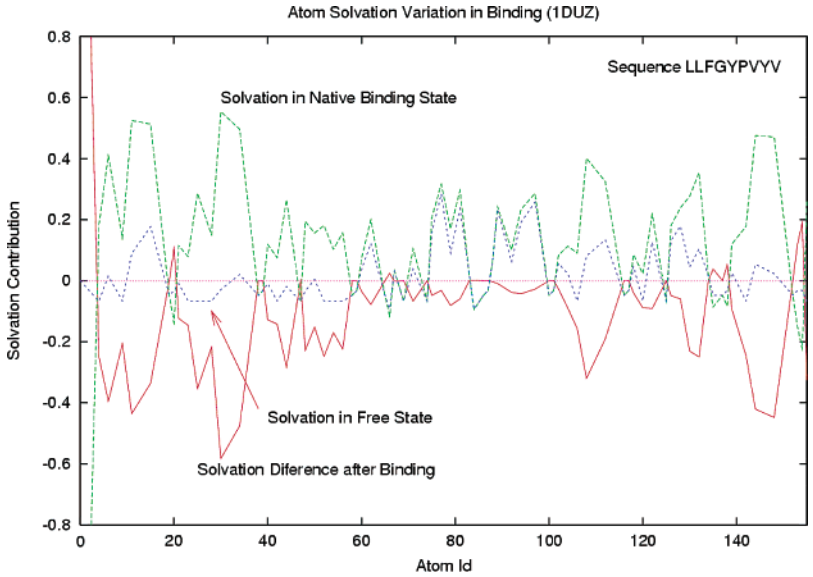

Figure 3. Analysis of the ligand solvation contribution of the MHC I complex 1duz.

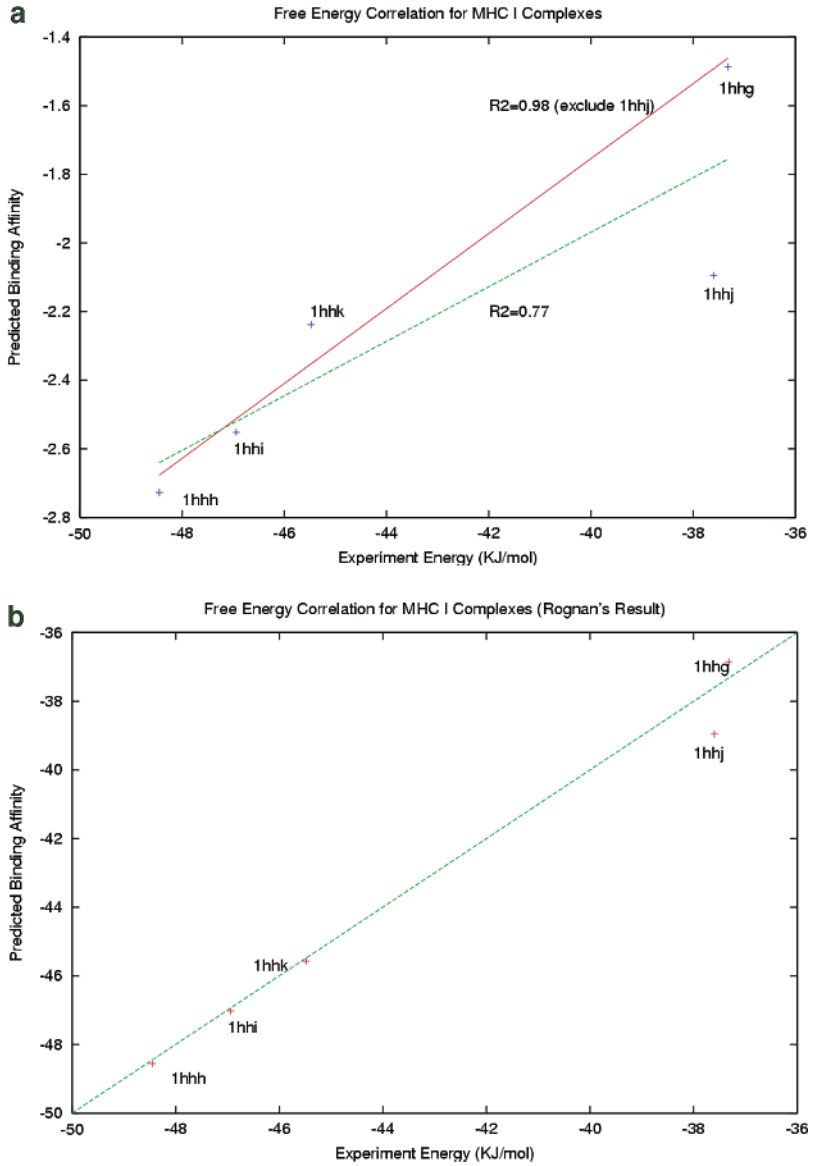

Figure 4. Comparison between this work and Rognan's work on predicted binding affinity for five determined X-ray structures of MHC I HLA-A*0201 complexes. (a) Fitting correlation of the developed docking potential. (b) Fitting correlation of empirical scoring function in Rognan's work ($R^2 = 0.895$).

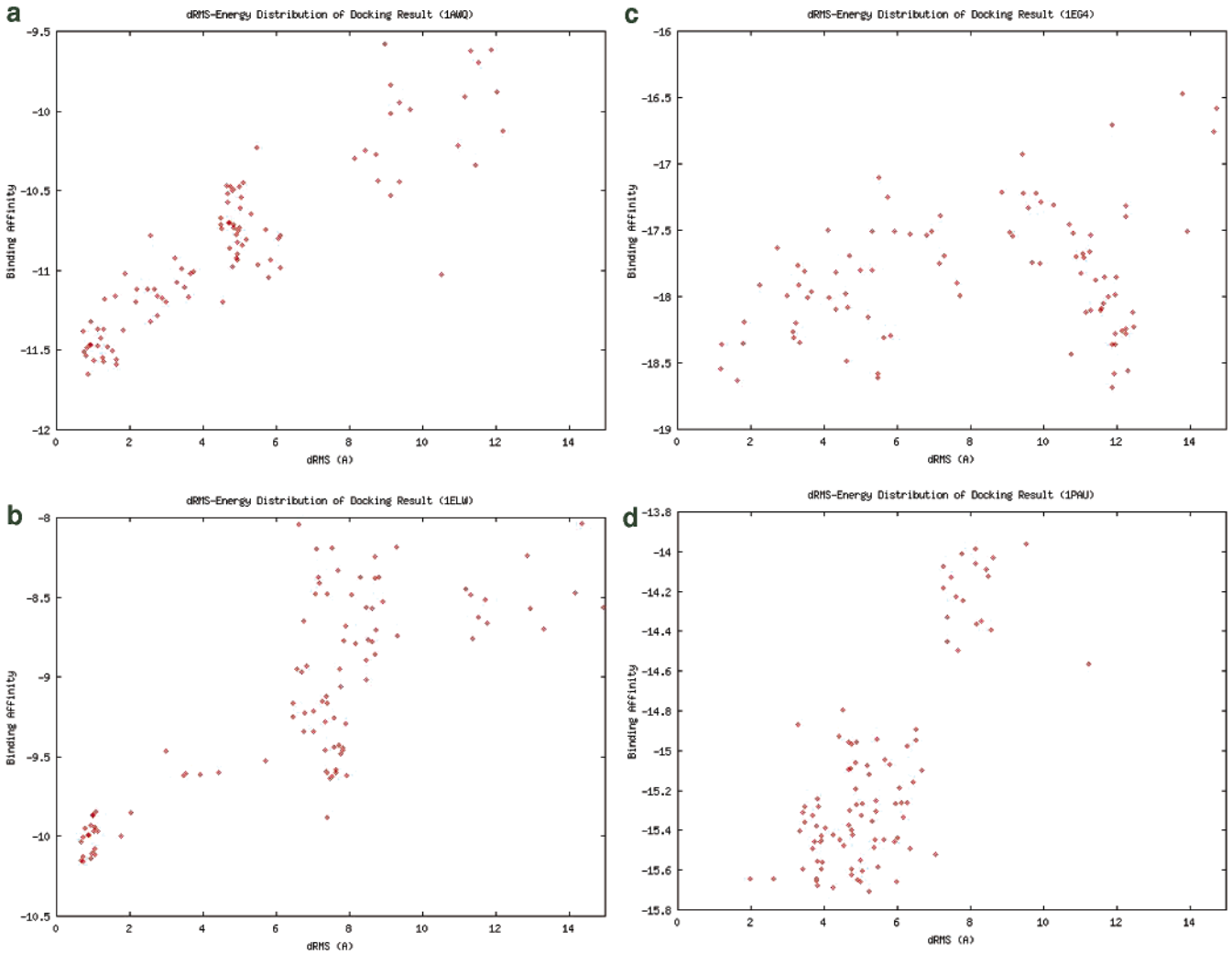

Figure 5. Accuracy of docking results: dRMS−energy correlation of protein−peptide docking. (a,b) Successful docking results (here results for complexes 1awq and 1elw are listed). (c) Successful docking result in which the native-like docked conformation can be distinguished by additional conformational clustering (result for complex 1eg4 listed). (d) Failed prediction result in which native-like conformations cannot be distinguished (here the result for complex 1pau is listed).

between the predicted binding affinities and the experimental binding free energies was obtained, comparable to the $R^2 = 0.895$ computed using an empirical scoring function specifically developed for MHC I protein complexes in Rognan's work (Figure 4b).[19,20] The predictions made by these two approaches are also quite similar in that both overestimate the binding affinity for 1hhj while predicting well the binding affinities for the remaining four complexes. By including the Bcl-X_L apoptosis complex[44] in the database, the correlation is maintained and even improved to yield $R^2 = 0.795$.

High correlation with regard to MHC I protein complexes as well as an additional Bcl-X_L complex suggests that our potential may be a transferable potential. It should be noted that the implied transferability is principally determined by the distribution of atom-pair interactions in the generated decoys, but not by the specific structures within the training set. Because the potential is based on atom-pair interactions, it converges given a reasonable number of decoys. For example, in supplement E (Supporting Information), the distance distribution of atom pairs between protein S atom type and peptide CN atom type in the decoys covers all S−CN distance possibilities in current PDB structures, though the decoys are developed from limited training complexes and the S atom has a lower frequency than other atom types in proteins. Therefore, if the training set can provide typical atom-pair distance favorites, the developed potential might be transferable to more protein−peptide complexes.

Protein−Peptide Docking Results. The docking algorithm and the potential were examined further by docking protein−peptide complexes from both the training set and the testing set. For each complex, 100 docked conformations were generated by Monte Carlo annealing with different random seeds, and their dRMS of protein−ligand heavy atom pairs were computed. Basically, the relationship between dRMS and energy in all cases followed the principal trend that conformations with lower energies have smaller dRMSs. A few cases demonstrate some deviations from this overall trend (Figure 5). A comprehensive analysis of these docking results is shown in Table 5. Successful simulations were conducted on 8 of 10 complexes in the training set (1awq, 1be9, 1bxl, 1cka, 1elw, 1gux, 1ycq, and 2fib) and 7 of 15 complexes in the testing set (1g3f, 1io6, 1ab9, 1bc5, 1duz, 1jhg, and 2seb), in which the generated conformations with the lowest energy were native-like structures with dRMS less than 3.0 Å.

A docked structure with the lowest energy, corresponding to the MHC I complex (1duz), is shown (Figure 6) and has the highest dRMS (3.01 Å) of all the minimum energy docked

Table 5. Flexible Protein−Peptide Docking Results

PDB ID	resolution/ Å	ligand	length	exposure degree/%	smallest dRMS/Å	lowest energy dRMS/Å
1a30[c]	2.00	EDL	3	16.2	2.38	5.37
1awq[a]	1.58	HAGPIA	6	53.9	0.73	0.85
1be9[a]	1.82	KQTSV	5	40.7	0.79	1.05
1bxl[a]	NMR	GQVGRQLAIIGDDINR	16	40.2	1.44	1.44
1cka[a]	1.50	PPPALPPKK	9	58.8	0.64	0.80
1eg4[b]	2.00	NMTPYRSPPPYVP	13	61.7	1.17	11.89
1elw[a]	1.60	GPTIEEVD	8	48.8	0.66	0.71
1gux[a]	1.85	DLYCYEQLN	9	46.4	0.68	1.79
1ycq[a]	2.30	ETFSDLWKLLP	11	42.4	0.79	1.63
2fib[a]	2.10	GPRP	4	32.2	0.57	2.62
1i31[b]	2.5	FYRALM	6	61.8	1.96	8.07
2cbl[b]	2.1	SDGY(PO4)TPEPA	9	64.8	1.68	9.25
1g3f[a]	NMR	AVPI	4	43.8	0.75	2.15
1io6[a]	NMR	RHYRPLPPLP	10	57.8	1.64	2.66
1ab9[a]	1.6	CGVPAIQPVL	10	64.1	0.60	0.68
1bc5[a]	2.2	NWETF	5	49.6	0.76	1.42
1duz[a]	1.8	LLFGYPVYV	9	26.1	2.78	3.01
1evh[b]	1.8	FPPPP	5	44.0	0.64	4.62
1f95[b]	NMR	MSCDKSTQT	9	44.2	0.91	3.34
1jhg[b]	1.3	W	1	37.3	1.01	2.44
1vwg[d]	1.46	CHPQGPPC	8	41.0	3.83	5.16
8tln[b]	1.6	VK	2	39.0	0.95	4.40
2seb[a]	2.5	AYMRADAAAGGA	12	38.7	1.04	1.04
1ce1[e]	1.9	GTSSPSAD	8	33.2	2.12	4.62
1pau[e]	2.5	DEVD	4	27.9	1.97	5.22
1shf[f]	1.9	PPPALPPKK	9	−	1.04	6.90
2cpl[f]	1.63	HAGPIA	6	−	0.68	3.21

[a] Complexes which were accurately docked. [b] Complexes which have native-like docked conformations can be distinguished by additional conformational clustering. [c−e] Complexes which failed in prediction because native-like conformations cannot be distinguished ([c]1a30 has the deeply buried binding pocket; [d]1vwg's ligand has to bend extremely to fit the pocket). [f] Result of docking peptide to unbound protein structure.

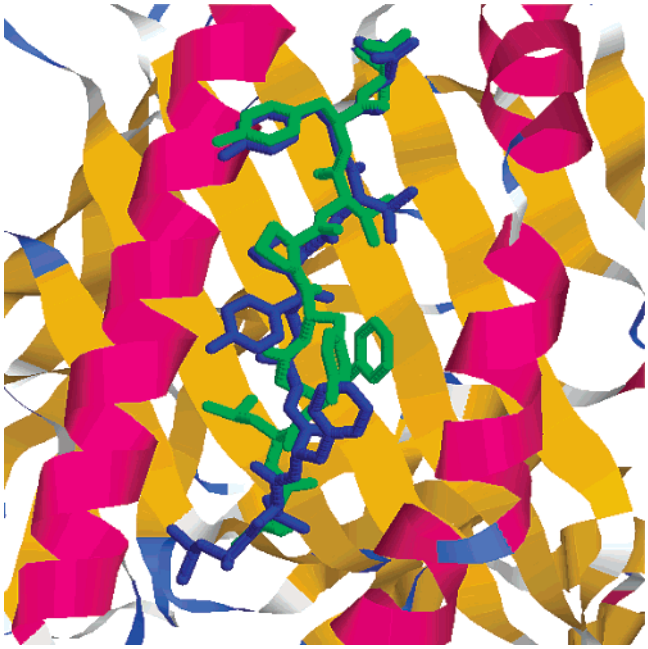

Figure 6. Structural comparison between the native conformation and the docked conformation of the MHC I complex (1duz). The native conformation of the peptide is colored green, and the docked conformation is colored blue.

complexes. The structural comparison between the minimum energy docked conformation and the native conformation revealed that the docking process identifies the ligand's native orientation in the binding site. In the minimum energy docked structure, the C-terminal conformation of peptide is almost identical to that in the native structure, while the N-terminal portion of the peptide just as closely resembles that in the native state. Generally, however, this is still a successful docking.

Further, for one complex in the training set (1eg4) and five complexes in the testing set (1i31, 2cbl, 1evh, 1f95, and 8tln), although the docked conformations with the lowest energy were not near-native structures, the docking simulations were still successful because the native-like conformations were among the 10 lowest energy conformations and could be easily identified (Figure 5c). Those non-native conformations with the lowest energy were often conformations trapped in a local energy minimum, in which the ligand had diffused away from the protein binding pocket. These trapped conformations could be avoided by considering more restrictive pocket restraints. In a few cases, the ligand adopted an alternative binding mode within the binding pocket.

In the end, a total of 9 out of 10 complexes in the training set and 12 out of 15 complexes in the testing set were accurately docked, where the averaged dRMS of native-like conformations was 0.83 Å in the training set and 1.23 Å in the testing set. The training set performed better than the testing set because it had better structural resolution, as well as being trained specifically, etc. Docking simulations on 1 of 10 complexes in the training set and 3 of 15 complexes in the testing set failed. Although these cases still generated near-native conformations, they were hard to distinguish because there were no obvious structural classes in the generated conformations and the energy−dRMS distribution was not clustered (see Figure 5d). A buried protein binding pocket might be the main reason for failure since all of these complexes had deep binding sites with a solvent exposure of less than 40%. Ligand flexibility might be another consideration.

Figure 7. Structural comparison between the docked conformation and the native conformation for the SH3 domain—stereoview of superposed structures of docking peptide to unbound protein structure in 1shf and native bound complex structure in 1cka. Green, native bound peptide structure in 1cka complex; blue, docked peptide structure to unbound protein structure in 1shf; remainder, superposed structures of unbound 1shf protein and bound 1cka protein.

To validate the prediction ability in unbound protein−peptide docking, two unbound protein structures, 1shf and 2cpl, are tested. 1shf and 1cka have the homologous SH3 domains from different species, for which the protein core structures are similar while the binding sites have small differences caused by neighboring loops. 2cpl has the unbound protein structure as in 1awq. The docking result shows that there are still similar energy−dRMS relationships for these two cases and the conformations with the smallest dRMSs, around 1.0 Å, belong to the top conformations with lower energies and can be easily identified by additional structural clustering. Superposing the above-obtained 1shf peptide conformation onto the 1cka native structure gives great structural fitness, as shown in Figure 7.

In general, the flexible docking algorithm can generate native-like protein−peptide complex conformations within 3.0 Å of the crystal structure. These docked conformations are often among the lowest energy structures and can be distinguished by additional conformational clustering when necessary.

Computational Speed. The computational speed of the program is benchmarked on an Intel-P4 2.8 GHz PC with Redhat 9.0 linux system. Two complexes, 1cka and 1awq, which have different protein sizes and peptide sizes, are tested. In supplement F (Supporting Information) we list the CPU time for a 1.5-million-step Monte Carlo docking. Because of pocket restraint in energy function, the protein size does not have significant effect on the speed, while the size and flexibility of peptide are the principal factors. Basically there is a linear correlation between the peptide flexibility and computational time. For a complex system consisting of a 200-residue protein and a 10-residue peptide, it will take not more than 2 h on the above computer, which is fast enough considering the flexibility of the system.

Discussion

***Z* Score Optimization Criterion.** Efficient *Z* score optimization should consider all complexes within the training set and include a reasonable optimization criterion. In our potential optimization, eq 11 was used initially, which was biased toward the complexes with a smaller *Z* gap ($|Z_i|$), such as 1cka. By optimizing on the basis of eq 11, the final *Z* score converges rapidly in the initial steps of the self-consistent optimization process, yielding a $\langle Z \rangle$ close to −5.0 (see Figure 8). However, this approach would produce a biased potential, which cannot generate efficient decoys for all training set complexes necessary for further optimization. Therefore, to eliminate this bias in the potential optimization process, eq 12 was adopted to more uniformly consider all complexes within the training set. This equation takes into account the average *Z* value as well as both boundaries of the *Z* gaps. *Z* score optimization attempts to maximize the energy gap between the native structure and the corresponding decoys, while the decoy generation procedure attempts to minimize the gap. As in traditional min−max

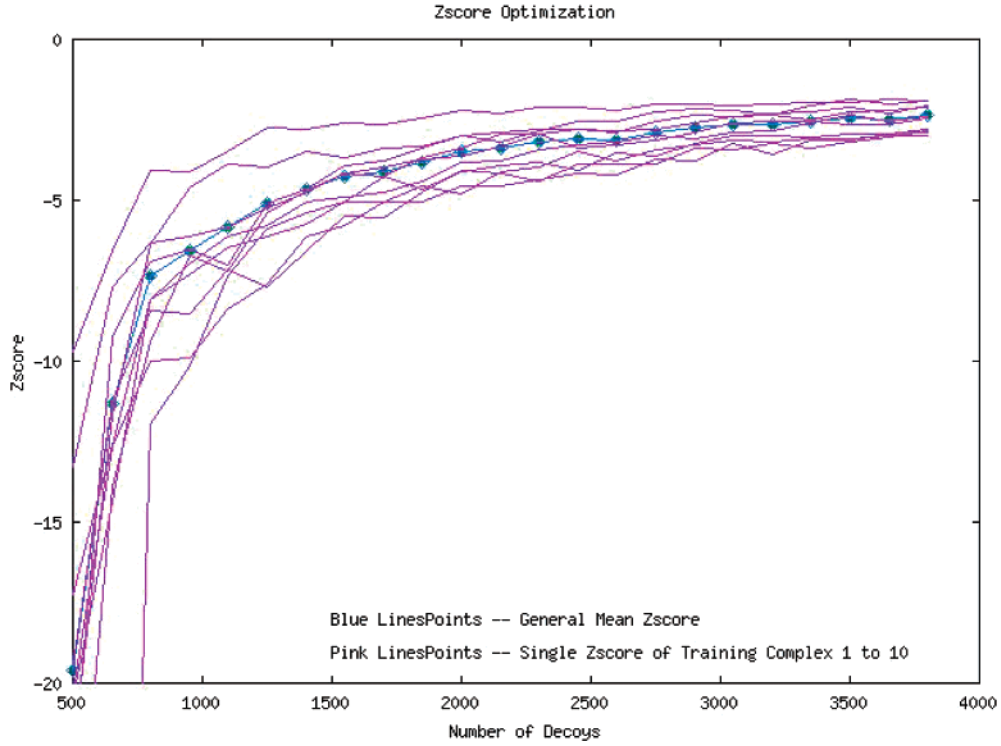

Figure 8. *Z* score optimization process.

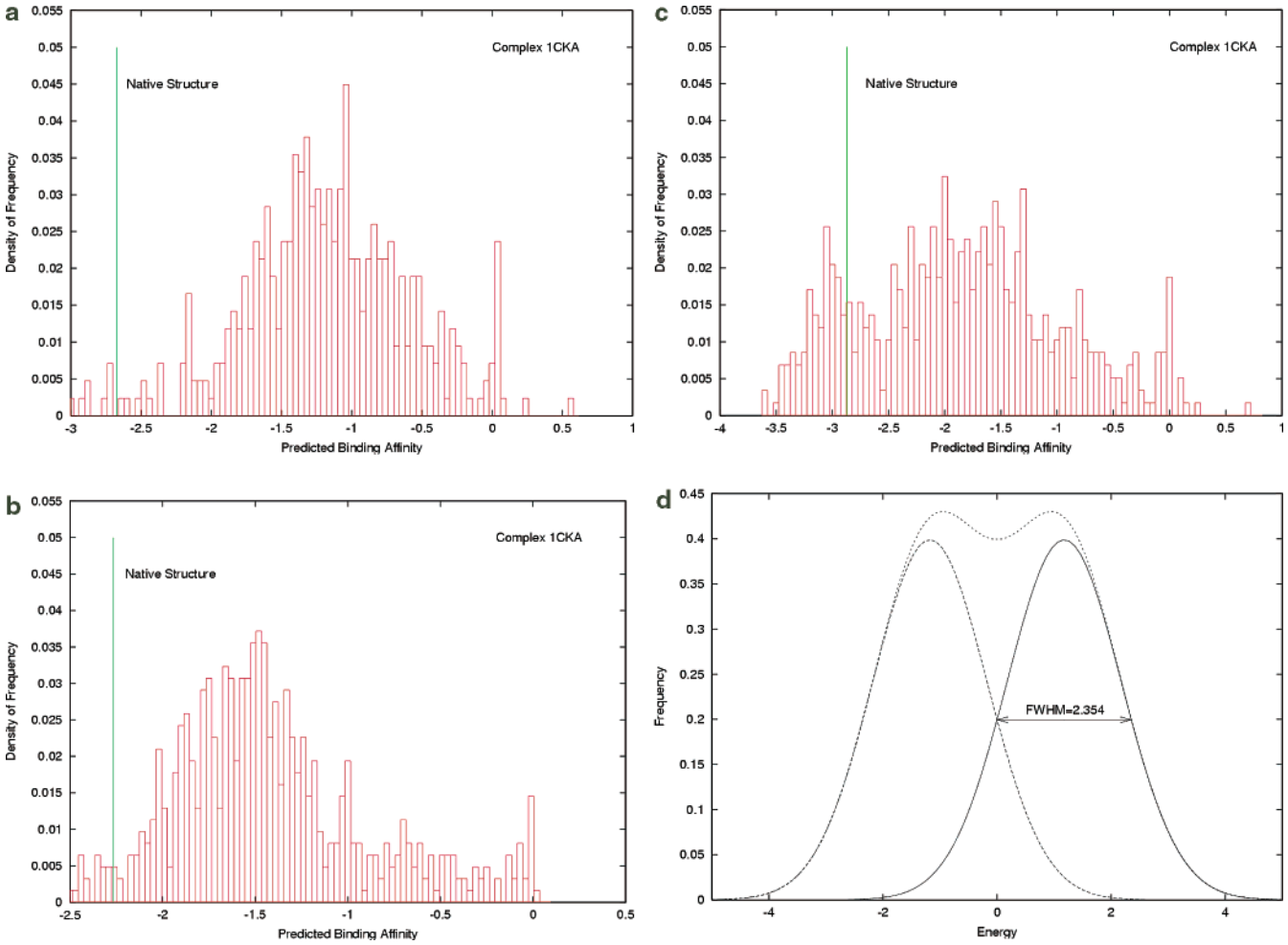

Figure 9. Energy distribution of generated decoys for 1cka. (a) Energy distribution of decoys in intermediate potential optimization process (total 3800 decoys), which basically followed Gaussian distribution and only a few native-like decoys appeared. (b) Energy distribution of decoys in the later potential optimization process (total 5600 decoys). More native-like decoys are generated. (c) Energy distribution with more near-native decoys (total 5600 decoys), which formed a shifted Gaussian distribution. (d) Criterion to distinguish two Gaussian distributions by fwhm.

algorithms, the solution to the optimal potential can be found from the above self-consistent process.

As stated previously, the energy of the generated decoys for each complex follows a Gaussian distribution determined by the critical Z score (Z_C). In addition, native structures are always located in the extreme tail of this energy distribution (Figure 9a). When the potential has been optimized well, more native-like decoys will be generated. The energies of these decoys may be similar to the energies of the native structures and would form a shifted Gaussian distribution around the native energy (Figure 9b). Because these native-like decoys are not significantly distinct from the native structures, further optimization to distinguish them from the native structures will overfit the potential and push Z_C to zero. Therefore, a new convergence criterion should be determined for Z_C.[52,53] A negative value of full-width half-maximum (fwhm), which in our case was -2.354, is the standard to distinguish two Gaussian distributions, and could be a reasonable optimization criterion for Z_C (Figure 9d). In our potential optimization, the final Z_C reached was -2.421.

The proposed Z_C criterion was supported by designed potential optimization based on Z_C. After the initial potential was developed from the initial 3500 docked decoys, it was optimized further using two different processes. In one process, additional decoys were generated by docking starting from the native bound conformations, producing more near-native decoys. The energies of these new decoys form an energy distribution inclusive of the native energies (Figure 9c). Another optimization process generated decoys by docking from far-native initial conformations. A comparison was conducted using the Z_C score optimization process. The correlation between the predicted binding affinity using the optimized potentials and the experimental binding free energies of the same five MHC I protein complexes revealed significant differences. The first process, based on near-native decoys, produces a potential that has failed to converge. The correlation with experimental binding affinities using this potential was quite poor (Figure 10). As in the combined Z score (eq 10), when Z_G was used in potential optimization, which more strictly distinguishes the native-like decoys from the native structure, the resulting potential was even worse. In the second process, based on more far-native decoys, the potential converges well and produces potentials with high correlation to experimental binding affinities.

As for the optimization criterion of gap Z score (Z_G) within the combined Z score (eq 10), the theoretical cutoff should be zero, because the critical Z score (Z_C) attempts to make the

(52) Mirny, L. A.; Shakhnovich, E. I. *J. Mol. Biol.* **1998**, *283*, 507−526.
(53) Mirny, L. A.; Finkelstein, A. V.; Shakhnovich, E. I. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 9978−9983.

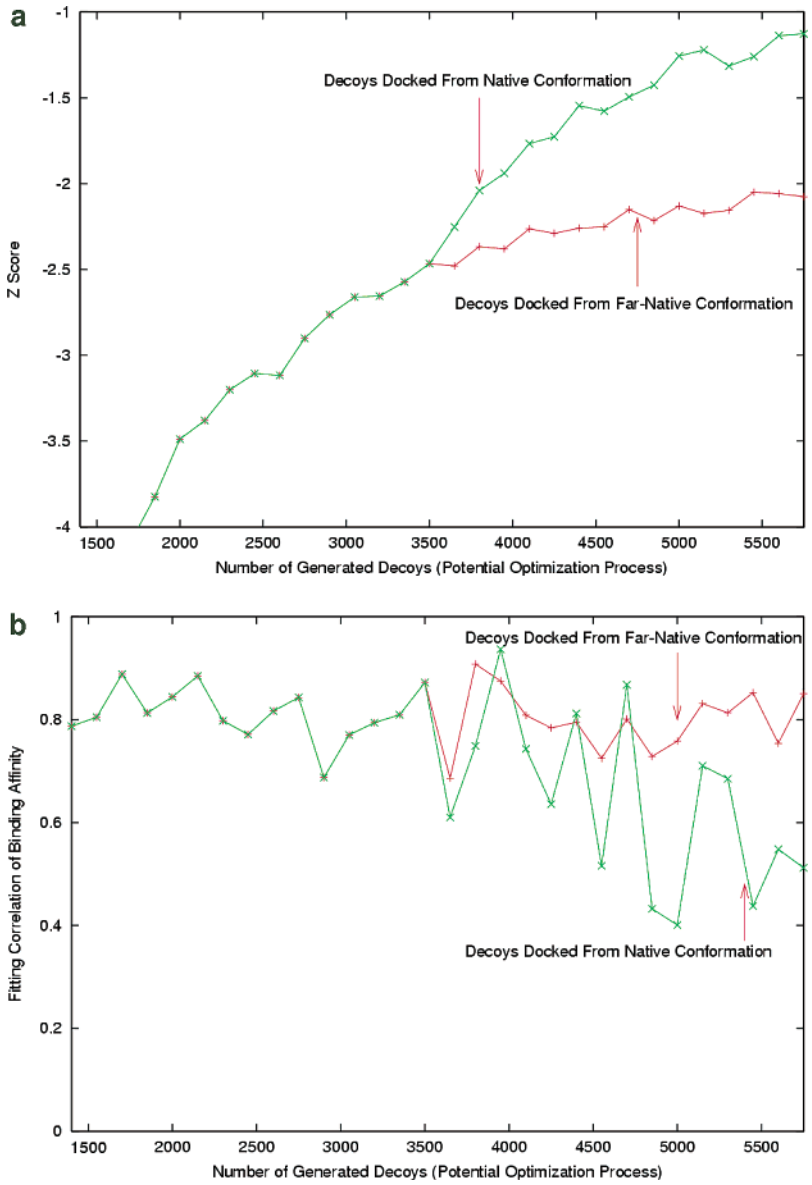

Figure 10. Determination of optimization criterion for critical *Z* score (Z_C) by comparison of two different potential optimization processes: 1, potential optimization based on more artificial near-native decoys generated by docking starting from native bound conformation; 2, potential optimization based on general decoys docked from the far-native state. (a) *Z* score comparison; process 1 was harder to converge. (b) Fitness correlation of binding affinity for five known MHC I complexes.

energies of decoys and the native structures continuously distributed. Our final Z_G reached a value of −0.395, and still has a small energy gap.

Gō Energy Contribution. In this work, Gō energies in the protein and peptide were introduced in order to restrain protein and peptide conformations, improving docking sampling efficiency. Here we investigate the importance of the Gō "restraints" as it pertains to the training set complexes (Table 6). First, the contribution of the Gō energies to the total energy was computed for the native structures. The smaller contribution of the ligand Gō energy (<6%) and larger contribution of protein Gō energy (>35%) roughly illustrates that the ligand Gō energy is less crucial while the protein Gō energy may be important. A more accurate description of the importance of the Gō energy should consider its contribution to the energy difference between docked conformations, which is the key to identifying final native-like docked structures. It was found that the ligand Gō energy range contributes minimally to the total energy range (6%) while the protein Gō energy contributes more significantly (>16%). Although the ligand Gō energy does not appear to strongly distinguish near-native from far-native decoys, the ligands nevertheless exhibit significant conformation variability, with total dRMS ranging from 0.0 to 15.0 Å. Therefore, the ligand Gō energy does not contribute significantly to the energy−dRMS distribution and is less important for the thermodynamic aspects of docking. Since the ligand Gō restraint does not perturb the obtained results, this energy term can be ignored in future docking. The protein Gō energy restraint, however, is still important, together with the protein−peptide distance-dependent contact energy and the atomic surface-based solvation energy.

Conclusion

In summary, to overcome high-energy repulsive barriers in protein−peptide docking, a docking method was developed which not only considers the full flexibility of the ligand but

Table 6. Gō Energy Contribution to Total Conformational Energy

		contribution of Gō energies to total energy of native structures		contribution of Gō energy range to total energy range of docked conformations	
PDB ID	ligand size	ligand Gō energy/%	protein Gō energy/%	ligand Gō energy/%	protein Gō energy/%
1a30	3	0.54	61.28	2.75	37.77
1awq	6	0.92	61.36	1.35	25.59
1be9	5	1.27	49.64	2.01	23.56
1bxl	16	4.02	51.25	4.40	17.33
1cka	9	3.15	35.89	2.03	16.39
1eg4	13	2.06	64.49	4.46	24.15
1elw	8	1.96	58.19	2.02	31.21
1gux	9	1.38	66.34	3.86	21.11
1ycq	11	5.22	43.74	5.14	16.51
2fib	4	0.39	64.37	1.66	36.89

also takes into account protein side-chain flexibility, which was proven to be crucial *in our preliminary calculation*—docking with an idealized dRMS virtual energy function. Using a physical optimization criterion and a training set of only 10 protein−peptide complex structures, a transferable potential was designed for actual docking process. The optimization procedure involves a novel iterative method based on *Z* score minimization and decoy generation. The potential considers protein−ligand atom-pair interactions and an atomic solvation contribution. The optimized potential accurately predicts binding affinity in protein−peptide complexes. With the optimized potential, the flexible docking algorithm could recover the binding state of most protein−peptide complexes with high precision. Both the docking method and rapid potential might have potential applications to database screening in drug discovery.

Acknowledgment. The authors are grateful to Concurrent Pharmaceuticals for funding this research and to Jun Shimada for his help at the initial stages of the project.

Supporting Information Available: Supplement A, table of protein−peptide potential atom types; supplement B, flexible docking algorithm used in the Monte Carlo annealing process; supplement C, parallel self-consistent process for potential optimization and decoy generation; supplement D, potential parameters dataset; supplement E, distribution of distance between protein-S and peptide-CN atom pair in decoys and PDB structures; and supplement F, benchmark on docking speed. This material is available free of charge via the Internet at http://pubs.acs.org.

JA032018Q