preparations were >95% pure, as estimated by inspection of Coomassie blue-stained SDS-polyacrylamide gels. The concentration of HIV-1 RT was approximately 1×10^4 units/g, as determined by comparing polymerase activity with that of an RT standard (Worthington Biochemical Corporation, Lakewood, NJ). Murine leukemia virus RT used in this study was obtained from New England Biolabs (Beverly, MA) and Promega. The *E. coli* RNase H was purchased from Sigma (St. Louis, MO). Activities were tested according to the manufacturers' protocols.

Screening of RNase H Inhibitors. Chemical compounds were purchased from Enamine Co. Ltd. (Ukraine). The purity of compounds **1** and **2** analyzed in detail was >99% according to the manufacturer (Enamine Co. Ltd.). For the first and second rounds of screening, the oligo ribonucleotide 5′-GGUCUCUCUGGUAGACCAGA-3′, corresponding to nucleotides 455−475 of HXB2 with 6-carboxy-fluoroscein (FAM) conjugated at the 5′ end was annealed to the oligo deoxyribonucleotide 5′-TCTGGTCTAACCAGAGAGACC-3′ using a final concentration of 1 μM each in annealing buffer containing 100 mM NaCl. Enzyme reactions containing 100 ng RT and 0.1 μM substrate were preincubated at 37 °C, and reactions were initiated by adding reaction buffer to yield the following final concentrations of components: 45 μM annealed oligos, 10 mM Tris-HCl, pH 7.8, 15 mM KCl, 1 mM $MgCl_2$, 0.4 mM DTT, 0.01% v/v nonionic detergent IGEPAL (Sigma), and 0.2 mM EDTA. Alternatively, the reaction was set up on ice in the reaction buffer and the reaction initiated by shifting the temperature to 37 °C. Both assay protocols yielded essentially the same results. The reactions were stopped by adding 3 μL of loading buffer containing 50% formamide, 25% glycerol, 2.5 mM EDTA, and a nucleic acid stain SYBR Green (1:10000 dilution, Invitrogen, Tokyo, Japan). A 4 μL aliquot was subjected to 4% urea-containing 18% PAGE for 60 min at 100 V, and the products were visualized by Typhoon9400 imager. For the third screening, a previously described real-time monitoring assay was used with the following modifications.[21] For substrates, the following oligonucleotides were annealed at final concentrations of 40 and 1 μM, respectively, in annealing buffer: oligo ribonucleotide 5′-GAUCUGAGCCUGGGAGCU-3′ with FAM conjugated at the 5′ end, and oligodeoxyribonucleotide 5′-AGCTCCCAGGCTCAGATC-3′ with black hole quencher (BHQ) conjugated at the 3′ end. Enzyme reactions containing 100 ng RT, 4 μM oligoribonucleotide, and 0.1 μM oligodeoxyribonucleotide were carried out in a volume of 10 μL at 37 °C in reaction buffer for the indicated times. Fluorescence at 488 nm signal was monitored every 150 s using a Multimode Detector (Beckman Coulter, Miami, FL).

Replication Monitoring. To produce HIV-1, 293T cells were transfected with plasmids encoding the HIV-1 proviral DNA (pNL4−3) and culture supernatants containing viruses were collected at 48 h post-transfection. For HIV-1 infection, 1.5×10^4 primary peripheral blood mononuclear cells depleted for $CD8^+$ T cells (denoted as primary $CD4^+$ T cells in the text) by MACS CD8 microbeads (Miltenyi Biotec, Tokyo) or 2×10^3 NP2CD4CXCR4 cells were incubated at the room temperature for approximately 30 min with HIV-1-containing culture supernatant having approximately 250 pg of the viral antigen p24. IL-2 and anti-CD3 antibody-stimulated primary $CD4^+$ T cells were passaged every 2−4 days, and the culture supernatants were collected when cells were passaged. For NP2CD4CXCR4 cells, culture supernatants were collected at 4 days postinfection. The culture supernatants were subjected to an ELISA assay to measure the p24 antigen, as an indicator of virus production, using a Retro TEK p24 antigen ELISA kit according to the manufacturer's protocol (Zepto Metrix, Buffalo, NY). The ELISA reactions were measured using a ELx808 microplate photometer (BIO-TEK, Winooski, VT).

Measuring Cytotoxicity. IL-2 and anti-CD3 antibody-stimulated primary $CD4^+$ T cells were plated at a density of 2.5×10^4 cells per well in 96-well plates and maintained in culture for 6−7 days in the various concentration of chemical compounds. The cell proliferation was evaluated by the 3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide (MTT) assay (CellTiter 96 AQueous; Promega) according to manufacturer's instructions. The OD_{490} was measured by an ELISA reader (BIO-TEK). Wells with culture medium only were measured as background signal. For cell lines, assays were performed with 1×10^3 cells per well. The 50% cytotoxicity concentration was defined as a drug concentration by which the OD_{490} value reached the 50% level of the no drug control. The maximum concentrations of compounds we tested was 50 μM.

Protein Structure Preparation. An X-ray structure of HIV-1 RT was downloaded from the Protein Data Bank (PDB code: 1RTD[23]). This structure contains a DNA template, primer, dTTP containing two Mg^{2+} ions, and two Mg^{2+} ions in the RNase H active site in addition to the p66 and p51 domains. For the docking simulations, the DNA template, primer, and dTTP were deleted, and hydrogen atoms were added using the PDB2PQR web service.[37,38] The resulting.pqr file was modified manually, and was converted into a mol2 file using babel3 2.2 (OpenEye Scientific Software, Inc.).

Molecular Docking. The three-dimensional structures of the ligands used in this study were generated from a SMILES string by OMEGA 2.2.1 (OpenEye Scientific Software, Inc.). The binding models for HIV-1 RNase H inhibitors were predicted using the docking program GOLD 3.2 (CCDC Software Ltd., Cambridge, UK). The binding site was initially defined as all residues of the target within 10 Å of the Mg^{2+} ion coordinated by Asp443, Asp498, and Asp549. ChemScore was chosen as a fitness function and the standard default settings were used in all calculations.[39] The Mg^{2+} ions were set to allow hexavalent coordination. Early termination was allowed for searching the ligand docking poses. In early termination mode, calculation is stopped when the root-mean-square deviation (rmsd) on heavy atoms of the ligand among top three high scoring docking poses are within 15 Å of the target. An additional docking simulation was executed using Glide 4.0 (Schrodinger Inc.) with equivalent calculation conditions.

Acknowledgment. We thank Dr. Beutler's and Dr. Pommier's laboratory groups in the National Cancer Institute for measuring the IC_{50}s of compounds **1** and **2** against human RNase H1 and HIV-1 RT and against HIV-1 IN, respectively. We appreciate Dr. Sato's group for sharing purified CRF01 A_E RT (National Institute of Infectious Diseases, Tokyo). This work was supported by the Japan Health Science Foundation, the Japanese Ministry of Health, Labor, and Welfare (H18-AIDS-W-003) and the Japanese Ministry of Education, Culture, Sports, Science, and Technology (18689014 and 18659136).

Supporting Information Available: Alignment of reverse transcriptase amino acid sequences of the clade B B.FR.x.pNL43, clade C Indie-C1, and 93JP-NH1, a representative of CRF01_AE. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Nikolenko, G. N.; Svarovskaia, E. S.; Delviks, K. A.; Pathak, V. K. Antiretroviral drug resistance mutations in human immunodeficiency virus type 1 reverse transcriptase increase template-switching frequency. *J. Virol.* **2004**, *78*, 8761–8770.

(2) Nikolenko, G. N.; Palmer, S.; Maldarelli, F.; Mellors, J. W.; Coffin, J. M.; Pathak, V. K. Mechanism for nucleoside analog-mediated abrogation of HIV-1 replication: balance between RNase H activity and nucleotide excision. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 2093−2098. Epub 2005 Jan 2031.

(3) Luo, G. X.; Taylor, J. Template switching by reverse transcriptase during DNA synthesis. *J. Virol.* **1990**, *64*, 4321–4328.

(4) Peliska, J. A.; Benkovic, S. J. Mechanism of DNA strand transfer reactions catalyzed by HIV-1 reverse transcriptase. *Science* **1992**, *258*, 1112–1118.

(5) DeStefano, J. J.; Mallaber, L. M.; Rodriguez-Rodriguez, L.; Fay, P. J.; Bambara, R. A. Requirements for strand transfer between internal regions of heteropolymer templates by human immunodeficiency virus reverse transcriptase. *J. Virol.* **1992**, *66*, 6370–6378.

(6) DeStefano, J. J.; Roberts, B.; Shriner, D. The mechanism of retroviral recombination: the role of sequences proximal to the point of strand transfer. *Arch. Virol.* **1997**, *142*, 1797–1812.

(7) DeStefano, J. J.; Bambara, R. A.; Fay, P. J. The mechanism of human immunodeficiency virus reverse transcriptase-catalyzed strand transfer from internal regions of heteropolymeric RNA templates. *J. Biol. Chem.* **1994**, *269*, 161–168.
(8) Klumpp, K.; Mirzadegan, T. Recent progress in the design of small molecule inhibitors of HIV RNase H. *Curr. Pharm. Des.* **2006**, *12*, 1909–1922.
(9) Borkow, G.; Fletcher, R. S.; Barnard, J.; Arion, D.; Motakis, D.; Dmitrienko, G. I.; Parniak, M. A. Inhibition of the ribonuclease H and DNA polymerase activities of HIV-1 reverse transcriptase by *N*-(4-*tert*-butylbenzoyl)-2-hydroxy-1-naphthaldehyde hydrazone. *Biochemistry* **1997**, *36*, 3179–3185.
(10) Sluis-Cremer, N.; Arion, D.; Parniak, M. A. Destabilization of the HIV-1 reverse transcriptase dimer upon interaction with *N*-acyl hydrazone inhibitors. *Mol. Pharmacol.* **2002**, *62*, 398–405.
(11) Marchand, C.; Beutler, J. A.; Wamiru, A.; Budihas, S.; Mollmann, U.; Heinisch, L.; Mellors, J. W.; Le Grice, S. F.; Pommier, Y. Madurahydroxylactone derivatives as dual inhibitors of human immunodeficiency virus type 1 integrase and RNase H. *Antimicrob. Agents Chemother.* **2008**, *52*, 361−364. Epub 2007 Oct 2029.
(12) Tramontano, E.; Esposito, F.; Badas, R.; Di Santo, R.; Costi, R.; La Colla, P. 6-[1-(4-Fluorophenyl)methyl-1*H*-pyrrol-2-yl)]-2,4-dioxo-5-hexenoic acid ethyl ester a novel diketo acid derivative which selectively inhibits the HIV-1 viral replication in cell culture and the ribonuclease H activity in vitro. *Antivir. Res.* **2005**, *65*, 117–124.
(13) Tarrago-Litvak, L.; Andreola, M. L.; Fournier, M.; Nevinsky, G. A.; Parissi, V.; de Soultrait, V. R.; Litvak, S. Inhibitors of HIV-1 reverse transcriptase and integrase: classical and emerging therapeutical approaches. *Curr. Pharm. Des.* **2002**, *8*, 595–614.
(14) Moelling, K.; Schulze, T.; Diringer, H. Inhibition of human immunodeficiency virus type 1 RNase H by sulfated polyanions. *J. Virol.* **1989**, *63*, 5489–5491.
(15) Loya, S.; Hizi, A. The interaction of illimaquinone, a selective inhibitor of the RNase H activity, with the reverse transcriptases of human immunodeficiency and murine leukemia retroviruses. *J. Biol. Chem.* **1993**, *268*, 9323–9328.
(16) Tan, C. K.; Civil, R.; Mian, A. M.; So, A. G.; Downey, K. M. Inhibition of the RNase H activity of HIV reverse transcriptase by azidothymidylate. *Biochemistry* **1991**, *30*, 4831–4835.
(17) Davis, W. R.; Tomsho, J.; Nikam, S.; Cook, E. M.; Somand, D.; Peliska, J. A. Inhibition of HIV-1 reverse transcriptase-catalyzed DNA strand transfer reactions by 4-chlorophenylhydrazone of mesoxalic acid. *Biochemistry* **2000**, *39*, 14279–14291.
(18) Klumpp, K.; Hang, J. Q.; Rajendran, S.; Yang, Y.; Derosier, A.; Wong Kai In, P.; Overton, H.; Parkes, K. E.; Cammack, N.; Martin, J. A. Two-metal ion mechanism of RNA cleavage by HIV RNase H and mechanism-based design of selective HIV RNase H inhibitors. *Nucleic Acids Res.* **2003**, *31*, 6852–6859.
(19) Shaw-Reid, C. A.; Munshi, V.; Graham, P.; Wolfe, A.; Witmer, M.; Danzeisen, R.; Olsen, D. B.; Carroll, S. S.; Embrey, M.; Wai, J. S.; Miller, M. D.; Cole, J. L.; Hazuda, D. J. Inhibition of HIV-1 ribonuclease H by a novel diketo acid, 4-[5-(benzoylamino)thien-2-yl]-2,4-dioxobutanoic acid. *J. Biol. Chem.* **2003**, *278*, 2777−2780. Epub 2002 Dec 2711.
(20) Himmel, D. M.; Sarafianos, S. G.; Dharmasena, S.; Hossain, M. M.; McCoy-Simandle, K.; Ilina, T.; Clark, A. D., Jr.; Knight, J. L.; Julias, J. G.; Clark, P. K.; Krogh-Jespersen, K.; Levy, R. M.; Hughes, S. H.; Parniak, M. A.; Arnold, E. HIV-1 reverse transcriptase structure with RNase H inhibitor dihydroxy benzoyl naphthyl hydrazone bound at a novel site. *ACS Chem. Biol.* **2006**, *1*, 702–712.
(21) Parniak, M. A.; Min, K. L.; Budihas, S. R.; Le Grice, S. F.; Beutler, J. A. A fluorescence-based high-throughput screening assay for inhibitors of human immunodeficiency virus-1 reverse transcriptase-associated ribonuclease H activity. *Anal. Biochem.* **2003**, *322*, 33–39.
(22) Chan, K. C.; Budihas, S. R.; Le Grice, S. F.; Parniak, M. A.; Crouch, R. J.; Gaidamakov, S. A.; Isaaq, H. J.; Wamiru, A.; McMahon, J. B.; Beutler, J. A. A capillary electrophoretic assay for ribonuclease H activity. *Anal. Biochem.* **2004**, *331*, 296–302.
(23) Huang, H.; Chopra, R.; Verdine, G. L.; Harrison, S. C. Structure of a covalently trapped catalytic complex of HIV-1 reverse transcriptase: implications for drug resistance. *Science* **1998**, *282*, 1669–1675.
(24) Katayanagi, K.; Okumura, M.; Morikawa, K. Crystal structure of *Escherichia coli* RNase HI in complex with Mg^{2+} at 2.8 A resolution: proof for a single Mg(2+)-binding site. *Proteins* **1993**, *17*, 337–346.
(25) Nowotny, M.; Gaidamakov, S. A.; Ghirlando, R.; Cerritelli, S. M.; Crouch, R. J.; Yang, W. Structure of human RNase H1 complexed with an RNA/DNA hybrid: insight into HIV reverse transcription. *Mol. Cell* **2007**, *28*, 264–276.
(26) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput.-Aided Mol. Des.* **1997**, *11* (5), 425–45.
(27) Smith, J. S.; Roth, M. J. Purification and characterization of an active human immunodeficiency virus type 1 RNase H domain. *J. Virol.* **1993**, *67*, 4037–4049.
(28) Sarafianos, S. G.; Das, K.; Tantillo, C.; Clark, A. D. Jr.; Ding, J.; Whitcomb, J. M.; Boyer, P. L.; Hughes, S. H.; Arnold, E. Crystal structure of HIV-1 reverse transcriptase in complex with a polypurine tract RNA:DNA. *EMBO J.* **2001**, *20*, 1449–1461.
(29) Schatz, O.; Cromme, F. V.; Gruninger-Leitch, F.; Le Grice, S. F. Point mutations in conserved amino acid residues within the C-terminal domain of HIV-1 reverse transcriptase specifically repress RNase H function. *FEBS Lett.* **1989**, *257*, 311–314.
(30) Nikolenko, G. N.; Delviks-Frankenberry, K. A.; Palmer, S.; Maldarelli, F.; Fivash, M. J.; Coffin, J. M.; Pathak, V. K. Mutations in the connection domain of HIV-1 reverse transcriptase increase 3′-azido-3′-deoxythymidine resistance. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 317−322. Epub 2006 Dec 2019.
(31) Brehm, J. H.; Koontz, D.; Meteer, J. D.; Pathak, V.; Sluis-Cremer, N.; Mellors, J. W. Selection of mutations in the connection and RNase H domains of human immunodeficiency virus type 1 reverse transcriptase that increase resistance to 3′-azido-3′-dideoxythymidine. *J. Virol.* **2007**, *81*, 7852−7859. Epub 2007 May 7816.
(32) Delviks-Frankenberry, K. A.; Nikolenko, G. N.; Barr, R.; Pathak, V. K. Mutations in human immunodeficiency virus type 1 RNase H primer grip enhance 3′-azido-3′-deoxythymidine resistance. *J. Virol.* **2007**, *81*, 6837−6845. Epub 2007 Apr 6811.
(33) Radzio, J.; Sluis-Cremer, N. Efavirenz accelerates HIV-1 reverse transcriptase ribonuclease H cleavage, leading to diminished zidovudine excision. *Mol. Pharmacol.* **2008**, *73*, 601−606. Epub 2007 Nov 2016.
(34) Hirakawa, K.; Midorikawa, K.; Oikawa, S.; Kawanishi, S. Carcinogenic semicarbazide induces sequence-specific DNA damage through the generation of reactive oxygen species and the derived organic radicals. *Mutat. Res.* **2003**, *536*, 91–101.
(35) Kim, B.; Hathaway, T. R.; Loeb, L. A. Human immunodeficiency virus reverse transcriptase. Functional mutants obtained by random mutagenesis coupled with genetic selection in *Escherichia coli*. *J. Biol. Chem.* **1996**, *271*, 4872–4878.
(36) Kim, B.; Hathaway, T. R.; Loeb, L. A. Fidelity of mutant HIV-1 reverse transcriptases: interaction with the single-stranded template influences the accuracy of DNA synthesis. *Biochemistry* **1998**, *37*, 5831–5839.
(37) Dolinsky, T. J.; Nielsen, J. E.; McCammon, J. A.; Baker, N. A. PDB2PQR: an automated pipeline for the setup of Poisson−Boltzmann electrostatics calculations. *Nucleic Acids Res.* **2004**, *32*, W665–W667.
(38) Dolinsky, T. J.; Czodrowski, P.; Li, H.; Nielsen, J. E.; Jensen, J. H.; Klebe, G.; Baker, N. A. PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Res.* **2007**, *35*, W522−W525. Epub 2007 May 2008.
(39) Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. Improved protein−ligand docking using GOLD. *Proteins* **2003**, *52* (4), 609–623.

JM801071M

e-J. Surf. Sci. Nanotech. Vol. 6 (2008) 241-245 Conference - ISSS-5 -

Development of Software Program Predicting the Binding Site and the Binding Mode of Ligands Against a Target Protein*

Hideyoshi Fuji,† Masaaki Suzuki, Saburo Neya, and Tyuji Hoshino‡
Department of Physical Chemistry, Graduate School of Pharmaceutical Sciences, Chiba University, 1-33 Yayoi-cho, Inage-ku, Chiba-shi, Chiba 263-8522, Japan
(Received 16 October 2008; Accepted 31 October 2008; Published 13 November 2008)

Structure-based drug design (SBDD) is one of the attractive and promising methods in drug discovery. In this study, we developed a software program predicting the binding site and the binding mode of ligands against a target protein for SBDD. The prediction method is based on the calculation of the hydrophobic potential energies around a target protein because the hydrophobic interaction is considered to be an important driving force in molecular recognition. Our program was tested with exemplifying 5 kinds of proteins, and was demonstrated to correctly predict the binding site and the binding mode of ligands seen in the experimentally determined structures for the protein-ligand complex. [DOI: 10.1380/ejssnt.2008.241]

Keywords: Computer simulations; Biophysics, medical physics, and biomedical engineering

I. INTRODUCTION

Three-dimensional structures of various proteins have been revealed due to the great advance in experimental determination method like X-ray crystal structure analysis and in theoretical prediction approach like homology modeling. With the development of these techniques, the significance of structure-based drug design (SBDD), which is a technique that accelerates the drug discovery process by utilizing structural information of a target protein, is increasing in the research for developing new drugs. Although information on the ligand-binding site is essentially required for SBDD, the ligand-binding site of target proteins is sometimes unknown. Therefore, the establishment of a computer program which correctly determines the binding site and the binding mode of ligands is important for the success of SBDD.

The purpose of this study is to develop software program named "Hydrophobe" to predict the binding site and the binding mode of ligands against a target protein. We focused on the hydrophobic effects because they are one of the major factors of the driving force for protein-ligand interactions [1]. To predict the binding site of ligands against a target protein, hydrophobic potential around a protein was calculated using an X-ray structure of the protein registered in Protein Data Bank (PDB) [2]. Then, the area showing the highest hydrophobicity was selected by clustering technique and determined as the binding site of ligands. Next, the binding mode of a ligand against the target protein was predicted by superimposing principal axes of inertia of the ligand and the selected hydrophobic area at the predicted binding site. We attempted to predict the binding site and the binding mode of the ligands against 5 kinds of proteins. As a result, the predicted binding site was fitted well to the position of the ligand seen in the X-ray structure in every case. The trials with other kinds of target proteins occasionally showed that the accuracy of the prediction was insufficient for determining the correct binding mode of the ligands. In this paper, we will discuss what type of target proteins are suitable for the prediction by our "Hydrophobe" program and will suggest the advantage of our program in screening of vast numbers of ligands against a target protein at low computational cost.

II. MATERIALS AND METHODS

A. Preparation of the data set

The experimentally determined structures were taken from Protein Data Bank (PDB) [2]. The X-ray structures of human immunodeficiency virus type 1 protease (HIV-1 PR), acetylcholinesterase (AChE), influenza virus neuraminidase (NA), estrogen receptor alpha (ER-α), and cholesterol oxidase (ChO) were used. Their PDB IDs are 1AAQ, 1EVE, 2QWB, 3ERD, and 1COY respectively. Ligands, water, and ions were deleted from the original PDB files.

B. Prediction of the ligand-binding site

Hydrophobic potentials around a target protein were calculated using a method which is based on the study by Yamaotsu *et al.* [3]. They proposed the HydrophoBicity On a Protein (HBOP)/HydrophoBic SITE (HBSITE) technique to swiftly find the substrate-binding site in a protein. First, the grid points of a lattice were generated around the protein surface. Second, probe carbon atoms were put in each grid point, and the hydrophobic energy in each grid point was calculated using the pair interaction free energy function determined by Israelachvili and Pashley [4]. The hydrophobic potential was calculated using only the carbon atoms of hydrophobic residues (Gly, Ala, Val, Leu, Ile, Met, Trp, Phe, and Pro) with excepting the amide carbon. In this calculation, the cut-off distance for the hydrophobic energy was set to 20 Å. Top 100 grid points in the ranking with respect to the calculated hydrophobic potential were selected, and the area showing the highest hydrophobicity was chosen by the clustering

*This paper was presented at International Symposium on Surface Science and Nanotechnology (ISSS-5), Waseda University, Japan, 9-13 November, 2008.

†Corresponding author: fuji@graduate.chiba-u.jp
‡Corresponding author: hoshino@faculty.chiba-u.jp

technique and determined as the binding site of the ligand.

C. Prediction of the ligand-binding mode

The binding mode of the ligand against a target protein was predicted by superimposing the principal axes of inertia of the ligand and the selected hydrophobic area at the predicted binding site. The coordinates of ligands were obtained from PDB data without adding hydrogen atoms. The conformations of the ligands were fixed to the PDB structure, that is, the ligand was used as a rigid body in the binding mode prediction.

D. Calculation of the appearance ratio of hydrophobic amino acids at the ligand-binding site and on the protein surface

The amino acids surrounding each ligand in the experimentally determined structures were examined. If any non-hydrogen atom in an amino acid residue was within 4.5 Å of any non-hydrogen atom of a ligand, it was expected that the amino residue interacted with the ligand. Therefore, 4.5 Å was used as the criteria to judge the presence of amino acid residues at the binding site [5]. The PyMol [6] script originally written in our laboratory was used to count the number of each kind of amino acid residue at the ligand-binding site. Then, the number of hydrophobic amino acid residues was divided by the number of all amino residues in the ligand-binding site to calculate the appearance ratio of hydrophobic amino residues.

The appearance ratio of hydrophobic amino acid residues was also investigated on the protein surface for comparison. The amino acid residues on the surface of the protein were identified by calculating the solvent-accessible surface area (SASA) of each amino residue using a probe sphere with a radius of 1.4 Å. The SASA was estimated with the MSMS program [7]. If the SASA value of an amino residue was higher than zero, the amino residue was regarded as being on the surface of the protein. The number of hydrophobic amino acid residues was divided by the number of all amino residues on the protein surface to calculate the appearance ratio of hydrophobic amino residues.

III. RESULTS AND DISCUSSIONS

We developed our original program, Hydrophobe, for determining the binding site and the binding mode of a ligand against a target protein. Firstly, we tested our program to predict the binding site of a ligand using 5 kinds of proteins. As shown in Fig. 1, our program successfully predicted the binding site in all 5 proteins. The ligand molecules were bound to the region with the highest hydrophobicity for all the proteins. This result is in agreement with the fact that the hydrophobic surface in the binding site is more exposed to solvent than that of the other site on a protein [5, 8, 9]. That is, the rate of appearance of hydrophobic amino acid residues at the binding site is relatively high [5]. Table I provides a summary of the appearance ratio of hydrophobic amino acid residues at the ligand-binding site and on the protein surface for the proteins investigated in this study. The appearance ratio of hydrophobic amino acid residues at the ligand-binding site was higher than the appearance ratio on the protein surface except for NA. The ligand-binding site of NA contains the highly conserved arginine triad (Arg118, Arg292, and Arg371) that forms a salt-bridge with the ligand carboxylic group. Though the Arg292 residue in the experimentally determined structure (PDB ID: 2QWB) was mutated into Lys [10], the appearance ratio of Arg residues at the binding site was 25%. Despite the observation that the ligand-binding site of NA was highly polar, the most hydrophobic area in NA found by Hydrophobe program was identical to the binding site of the protein. This result implies that the ligand association with NA is initiated by getting rid of waters from the binding site because the ligand is more hydrophobic than water molecules. After that, the binding between NA and its ligand is strengthened by electrostatic interactions.

In the case of ChO, the enzyme contains flavin ade-

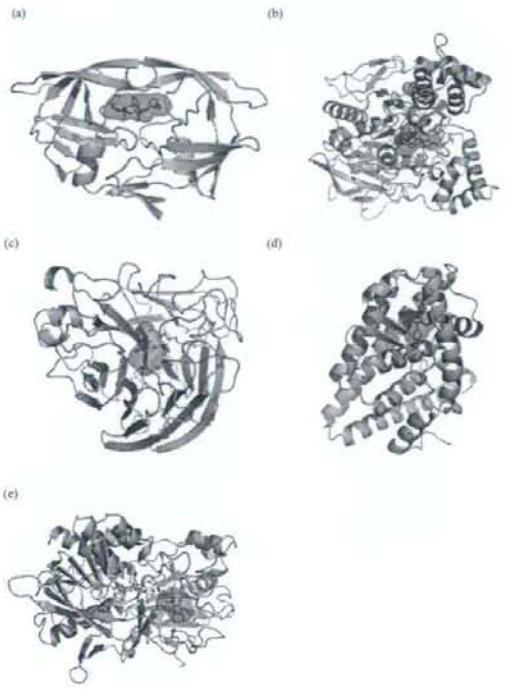

FIG. 1: Prediction of the ligand-binding site for (a) HIV-1 PR, (b) AChE, (c) NA, (d) ER-α, and (e) ChO. The ligand-binding site predicted by Hydrophobe is depicted in the red surface representation. The ligand in X-ray crystal structure is depicted in the stick representation with green carbon, red oxygen, and blue nitrogen atoms. The cofactor molecule FAD is depicted in the stick representation with yellow carbon, red oxygen, blue nitrogen, and orange phosphorus atoms in Fig. 1(e).

TABLE I: Appearance ratio of hydrophobic amino acid residues at the binding site and on the protein surface.

Protein	PDB ID	Appearance ratio[a]	
		binding site	protein surface
HIV-1 protease	1AAQ	70.4% (19/27)	55.5% (96/173)
Acetylcholinesterase	1EVE	52.6% (10/19)	42.3% (170/402)
Neuraminidase	2QWB	25.0% (4/16)	37.5% (107/285)
Estrogen receptor alpha	3ERD	80.0% (16/20)	46.4% (97/209)
Cholesterol oxidase	1COY	75.0% (15/20)	46.7% (188/403)

[a]The values in parenthesis indicate the number of hydrophobic amino acid residues per total amino residues.

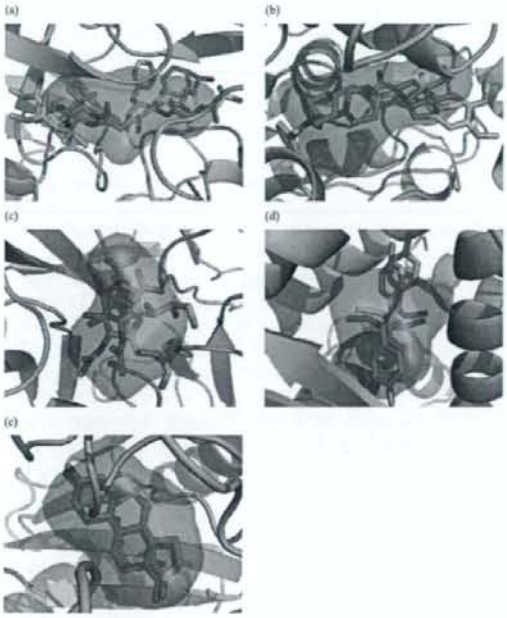

FIG. 2: Prediction of the ligand-binding mode for (a) HIV-1 PR, (b) AChE, (c) NA, (d) ER-α, and (e) ChO. The ligand in the X-ray crystal structure is depicted in a similar manner to Fig. 1. The ligand structure predicted by Hydrophobe is presented in the stick representation with cyan carbon, red oxygen, and blue nitrogen atoms.

nine dinucleotide (FAD) as a cofactor. When the cofactor was deleted before performing the binding site prediction, the predicted binding site matched the FAD-binding site (data not shown). From this result, it is found that hydrophobicity of the cofactor-binding site is also high. Therefore, we left the cofactor to predict the ligand-binding site of ChO, and our Hydrophobe program gave a successful prediction (Fig. 1(e)).

Secondly, we tested our program to predict the binding mode of a ligand against the proteins. The prediction was done only by fitting the principal axes of inertia of a ligand to that of a bundle of grids existing in the area with the highest hydrophobicity determined with Hydrophobe. The RMSD values for ligand heavy atoms between the X-ray structure and the predicted one in HIV-1 PR, AChE, ER-α, and ChO were 3.21 Å, 3.43 Å, 2.81 Å, 1.16 Å, and 0.44 Å, respectively (Fig. 2). After a minimization calculation was performed to remove the steric collision between a protein and a ligand, the RMSD values were improved to 1.32 Å, 1.87 Å, 3.16 Å, 1.27 Å, and 1.19 Å, respectively (data not shown). In general, a prediction of binding mode is usually considered to be successful if the RMSD value is lower than 2.0 Å in docking studies [11–13]. Therefore, it can safely be said that our program successfully predicts the binding mode except for NA. Because the ligand in NA contains many hydrophilic functional groups (5 hydroxy groups, 1 carboxy group, 1 amide group, and 1 ester group), it can be considered that the calculation of hydrophilic interactions, for example hydrogen bonds and salt bridges, are needed to refine the binding mode.

The results showed that Hydrophobe can correctly predict the binding site and the binding mode of ligands against 5 kinds of proteins using the hydrophobic potential and the principal axes fitting method. The hydrophobic interaction is considered to be an important driving force in molecular recognition [1]. Our results support this consideration because the binding site of a ligand was determined by only calculating the hydrophobic potential around a target protein. Surprisingly, the prediction of the ligand-binding mode was successful only by aligning the orientation of the principal axes of inertia between a ligand molecule and the binding site.

There are four possible orientations to be considered unless the direction of the principal axes of inertia is taken into account in alignment of a ligand and the binding site [14] as shown in Fig. 3. Pose1 was the best orientation for the ligands of 5 kinds of proteins tested in this study. Meanwhile, when the binding mode prediction was performed for β-lactamase (β-Lac, PDB ID: 1BLC), pose1 was not the best orientation. Instead, pose2 obtained in the prediction showed more favorable binding mode, in which the RMSD value for ligand heavy atoms between the X-ray structure and the predicted one was 2.90 Å(Fig. 4(a)). In pose1, the carboxylic group of the ligand faces to the opposite direction compared to the experimentally determined structure, and the RMSD value for ligand heavy atoms was 5.02 Å. Because the size of the ligand is small and the shape of the ligand is spherical, the binding mode cannot be simply determined only by aligning the direction of the principal axes of inertia

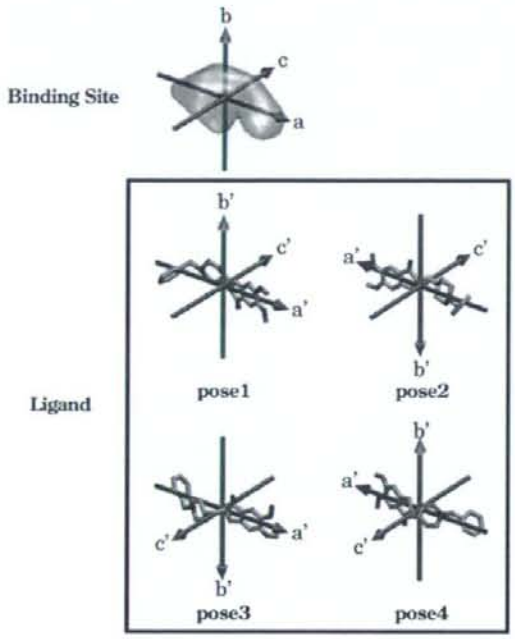

FIG. 3: Four possible orientations by the alignment of the principal axes of inertia between a binding site and a ligand molecule.

between the ligand molecule and the binding site. The appearance ratio of hydrophobic amino acid residues at the ligand-binding site of β-Lac was 20%, and the ligand in β-Lac contains some hydrophilic functional groups. For the same reason as NA, it seems that the refinement of the binding mode is needed by the calculation including hydrophilic interactions between the protein and the ligand. In this study, we carried out minimization calculation for four possible docked structures of β-Lac using the Protein Preparation Wizard Script within Maestro (Schrodinger Inc.), and then the docking scores were calculated by several popular scoring functions (Goldscore [15], Chemscore [13], ASP [16], Glidescore [17], and X-Score [18]). As a result, Goldscore and X-Score selected pose3; Chemscore, ASP, and Glidescore selected pose2 as the best-docked structure (data not shown). Thus, we tried to determine the best orientation of the ligand for β-Lac, and Chemscore, ASP, Glidescore suggested the consistent result with the crystal structure. Although the best orientation should be theoretically selected by considering the docking affinity between the target protein and the docked ligand, some currently available scoring functions failed to give a reliable selection. Accordingly, we are planning to evaluate those scoring functions and introduce an advanced calculation technique to elaborate the ligand-binding modes predicted by Hydrophobe program in our future study.

Another example of failure in the binding mode prediction is Angiotensin-converting enzyme (ACE, PDB ID: 1O86) as shown in Fig. 4(b). A part of the ligand of the

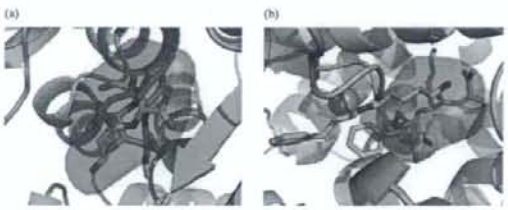

FIG. 4: Examples of failure in prediction of the ligand-binding mode for (a) β-Lac, and (b) ACE. The ligand-binding site predicted by Hydrophobe and the ligand in the X-ray crystal structure is depicted in a similar manner to Fig. 1. The ligand structure of pose1 predicted by Hydrophobe is presented in the stick representation with cyan carbon, red oxygen, and blue nitrogen atoms. The ligand structure of pose2 is presented in the stick representation with white carbon, red oxygen, and blue nitrogen atoms. Zinc ion is depicted in the gray sphere representation.

experimentally determined structure is overlapping with the hydrophobic area determined by Hydrophobe. ACE is a metalloprotein containing one zinc ion at the active site, which strongly mediates the protein-ligand interaction. The binding mode of the ligand of ACE would be mainly determined by the ligand-metal interaction. Therefore, it is naturally understood that the binding mode of the ligand of ACE was not correctly predicted only by the hydrophobic potentials. The binding mode prediction was not successful for ACE, but the hydrophobic area determined by Hydrophobe would be useful information as a guide to increase the binding affinity of the ligand in the process of the optimization of lead chemicals in drug discovery.

The currently available docking programs, such as AutoDock [19], DOCK [20], GOLD [13, 15], and GLIDE [17], require approximately 30-300 seconds in CPU time for the binding mode prediction of one compound, depending on the size of a compound, the parameter settings, and the computer performance. In general, virtual screening is performed to search active compounds from a chemical library that contains a great number of different kinds of compounds. The total entry number of the library is usually from hundreds of thousands to several million. Hence, the above docking programs seem unsuitable for the practical use for virtual screening because it takes over one year to screen a million of compounds. Therefore, simple filters such as Lipinski's rule of five [21], which predicts poor adsorption and permeability of compounds with four parameters (molecular weight, CLogP, the number of H-bond donors, and the number of H-bond acceptors), are applied to eliminate non-prospective compounds before executing the binding mode prediction. In our Hydrophobe program, once the most hydrophobic area is determined, it takes only 0.025 seconds on average in CPU time for predicting the binding mode prediction of one compound on 2.0 GHz Intel Core Duo. Conformation generation of a compound takes approximately 5 seconds in CPU time by OMEGA (OpenEye Scientific Software, Inc.) [22]. Even if 100 conformations of a compound were docked into a target protein by

using Hydrophobe, it takes about 7.5 seconds. When the binding mode prediction was performed for a million of compounds, it will finish within three months. The calculation can be finished in a week if using 13 CPUs in parallel. Accordingly, Hydrophobe will effectively serve the purpose for the initial rapid screening of potent compounds that can fit well into the binding site.

The program developed in this study will be a useful tool for determining the binding site when the binding site of a target protein is unknown. Furthermore, this program enables us to suggest a binding mode of the ligand in the process of the structure-based drug design or the *in silico* screening of compounds.

IV. SUMMARY

We developed a software program named Hydrophobe to predict the binding site and the binding mode of a ligand against a target protein. The method predicting the binding site of a ligand is the HBOP/HBSITE approach previously proposed by Yamaotsu *et al.* [3], in which only one simple function describing hydrophobic potential is employed to search the ligand-binding area. The grid points generated at the most hydrophobic area were used as a spatial query to predict the binding mode of a ligand, in which the principal axes of inertia between the grid points and heavy atoms of a ligand were aligned. The prediction accuracy of Hydrophobe was tested for 5 kinds of proteins, and the results showed that Hydrophobe can correctly predict both the binding site and the binding mode of the ligand against those proteins. This software will be useful for the initial rapid screening of potent compounds stored in the chemical database with a great number of entries.

Acknowledgments

A part of the prediction calculation was performed by using the super computer system at Institute of Media and Information Technology of Chiba University. This work was supported by a grant-in-aid from Japan Society for the Promotion of Science. A part of this work was supported by a Health and Labor Sciences Research Grant for Research on HIV/AIDS from the Ministry of Health, Labor and Welfare of Japan.

[1] T. Young, R. Abel, B. Kim, B. J. Berne, and R. A. Friesner, Proc. Nat. Acad. Sci. U.S.A. **104**, 808 (2007).
[2] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, Nucleic Acids Res. **28**, 235 (2000).
[3] N. Yamaotsu, A. Oda, and S. Hirono, Biol. Pharm. Bull. **31**, 1552 (2008).
[4] J. Israelachvili and R. Pashley, Nature **300**, 341 (1982).
[5] S. Soga, H. Shirai, M. Kobori, and N. Hirayama, J. Chem. Inf. Model. **47**, 400 (2007).
[6] W. L. DeLano, *The PyMOL Molecular Graphics System* (http://www.pymol.org).
[7] M. F. Sanner, A. J. Olson, and J. C. Spehner, Biopolymers **38**, 305 (1996).
[8] M. D. Kelly and R. L. Mancera, J. Med. Chem. **48**, 1069 (2005).
[9] S. Soga, H. Shirai, M. Kobori, and N. Hirayama, J. Chem. Inf. Model. **47**, 2287 (2007).
[10] J. N. Varghese, P. W. Smith, S. L. Sollis, T. J. Blick, A. Sahasrabudhe, J. L. McKimm-Breschkin, and P. M. Colman, Structure **6**, 735 (1998).
[11] J. C. Cole, C. W. Murray, J. W. Nissink, R. D. Taylor, and R. Taylor, Proteins **60**, 325 (2005).
[12] H. Gohlke, M. Hendlich, and G. Klebe, J. Mol. Biol. **295**, 337 (2000).
[13] M. L. Verdonk, J. C. Cole, M. J. Hartshorn, C. W. Murray, and R. D. Taylor, Proteins **52** (2003).
[14] C. M. Venkatachalam, X. Jiang, T. Oldfield, and M. Waldman, J. Mol. Graph. Model. **21**, 289 (2003).
[15] G. Jones, P. Willett, R. C. Glen, A. R. Leach, and R. Taylor, J. Mol. Biol. **267**, 727 (1997).
[16] W. T. Mooij and M. L. Verdonk, Proteins **61**, 272 (2005).
[17] R. A. Friesner, *et al.*, J. Med. Chem. **47**, 1739 (2004).
[18] R. Wang, Y. Lu, and S. Wang, J. Med. Chem. **46**, 2287 (2003).
[19] G. M. Morris, D. S. Goodsell, R. S. Halliday, R. Huey, W. E. Hart, R. K. Belew, and A. J. Olson, J. Comput. Chem. **19**, 1639 (1998).
[20] T. J. Ewing, S. Makino, A. G. Skillman, and I. D. Kuntz, J. Comput. Aided Mol. Des. **15**, 411 (2001).
[21] C. A. Lipinski, F. Lombardo, B. W. Dominy, and P. J. Feeney, Adv. Drug Deliv. Rev. **46**, 3 (2001).
[22] J. Bostrom, J. R. Greenwood, and J. Gottfries, J. Mol. Graph. Model. **21**, 449 (2003).

Ab initio Protein Structure Prediction with Force Field Parameters Derived from Water-Phase Quantum Chemical Calculation

DAISUKE KATAGIRI, HIDEYOSHI FUJI, SABURO NEYA, TYUJI HOSHINO*
Graduate School of Pharmaceutical Sciences, Chiba University, Chiba 263-8522, Japan

Received 28 July 2006; Revised 9 January 2008; Accepted 21 January 2008
DOI 10.1002/jcc.20963
Published online 25 March 2008 in Wiley InterScience (www.interscience.wiley.com).

Abstract: Molecular dynamics (MD) simulations are extensively used in the study of the structures and functions of proteins. *Ab initio* protein structure prediction is one of the most important subjects in computational biology, and many trials have been performed using MD simulation so far. Since the results of MD simulations largely depend on the force field, reliable force field parameters are indispensable for the success of MD simulation. In this work, we have modified atom charges in a standard force field on the basis of water-phase quantum chemical calculations. The modified force field turned out appropriate for *ab initio* protein structure prediction by the MD simulation with the generalized Born method. Detailed analysis was performed in terms of the conformational stability of amino acid residues, the stability of secondary structure of proteins, and the accuracy for prediction of protein tertiary structure, comparing the modified force field with a standard one. The energy balance between α-helix and β-sheet structures was significantly improved by the modification of charge parameters.

 J Comput Chem 29: 1930–1944, 2008

Key words: molecular dynamics; force field; water phase; prediction of protein structure

Introduction

Recent developments in techniques for molecular dynamics (MD) simulation and rapid progress in performance of computers enabled us to execute large-scale calculations for many biomolecules.[1,2] Therefore, the role of computers has become more important in molecular biology. *Ab initio* prediction of protein structure with a computer is one of the most popular subjects attracting great interest in computational biology. Nowadays, a long-time MD simulation becomes possible with the appearance of parallel computer, which raises expectation of predicting the protein folding structures only from sequence data on amino acid residues of proteins. Computer simulation, however, still has a problem to be overcome to achieve the reliable structure prediction. For example, the computationally predicted structure is sometimes unsatisfactory and different from the experimentally determined structure. Accordingly, further improvement in the accuracy of computer simulation, especially of MD simulation, is currently demanded.

The accuracy of protein structure prediction by MD simulation largely depends on the conformation sampling, the force field parameter, the solvent effect, etc. The reliability of conformation sampling has been greatly improved in recent years because of the increase of computation time due to the rapid progress of computer hardware. As for force field, the latest parameter in the AMBER MD simulation program significantly increased the accuracy of simulation results of protein.[3,4] This parameter, ff03 force field, was developed by Duan et al. and is now the most frequently used one in protein simulation. The charges in ff03 force field were derived from quantum chemical (QC) calculations in ether phase ($\varepsilon = 4.335$). This is a remarkable difference from other previous force field parameters. That is, ff03 force field is a parameter explicitly taking account of the solvent effect in atom charge for the first time.

Water molecules surrounding a protein play an important role not only in stability of the protein structure but also in function of proteins such as enzymatic reaction and ligand recognition. As a matter of fact, water molecules are frequently observed at the active sites of proteins in the structures determined by X-ray crystallographic analysis.[5–8] The influence of water molecules will be more serious for small proteins because they have a relatively large solvent accessible surface area for

This article contains supplementary material available via the Internet at http://www.interscience.wiley.com/jpages/0192-8651/suppmat.

*Also belongs to: PRESTO, Japan Science and Technology Agency, Saitama 322-0012, Japan

Correspondence to: T. Hoshino; e-mail: hoshino@faculty.chiba-u.jp

Contract/grant sponsors: JSPS Research Fellowship for Young Scientists and Japan Science and Technology Agency

their volume. Membrane proteins may also be affected by water molecules because membrane proteins often regulate their functions through the binding of cytokines at the loop regions on the outside of membrane and cause the subsequent change in their conformation. Since the loop regions are exposed to many water molecules, the influence of water molecules cannot be ignored. Our previous QC studies demonstrated that the β-sheet structure was dominantly stabilized by water molecules[9] and that the end of the helix structure was also stabilized by water molecules.[10] Another QC study on the conformational stability of alanine dipeptide indicated that the most stable conformation in water phase was different from that in gas phase or ether phase.[11]

To incorporate the solvent effect, calculation models used in the ordinary MD simulation explicitly contain water molecules around a protein. In contrast, the generalized Born (GB) method, in which the solvent effect is implicitly considered, is often used for the purpose of saving simulation time. The solvent effect in the GB method has been refined by the recent studies on the GB calculation theory.[12–17] The GB method is quite suitable for protein prediction because protein conformation drastically changes during the simulation. Hence, we executed MD simulation with the GB method, using ff03 force field in the preliminary trials for *ab initio* prediction of protein folding. Our trials, however, barely succeeded. The results suggested that the solvent effect by the GB method or ff03 force field was insufficient because the accuracy of prediction was poor especially at β-sheet and loop regions. Judging from detailed inspection of our trials, ff03 force field seemed insufficient in stability of β-sheet and loop structures and to have a strong helical tendency. Hence, a modification leading a good valance in energetic stability between α-helix and β-sheet structures is needed for reliable prediction of protein folding with MD simulation.

In this study, a force field modified from ff03 was applied to *ab initio* protein structure prediction. The atom charges of this force field was determined from the results of water-phase QC calculations for the purpose of obtaining more large stability of β-sheet and loop conformations because the charge parameters are the most effective in force field and the electrostatic energy calculated from the atom charges has the largest contribution in total energy of MD simulation among the terms for bond, angle, torsion energies, and van der Waals, electrostatic interaction energies. Evaluation of this new force field was performed in terms of amino acid conformations, secondary structure, and tertiary structure of proteins. The improvement will be seen in the energy balance between α-helix and β-sheet structures.

Methods

Determination of Atom Charges with Water-Phase QC Calculations

In MD simulation, atom charges largely affect the total energy through the electrostatic potential. Therefore, only the charges of ff03 force field were modified. In the development of ff03 force field by Duan et al., the effective charges were obtained by fitting to the ether-phase electrostatic potentials calculated at the B3LYP/cc-pVTZ//HF/6-31G** level by the RESP method.[18] In our study, the calculation condition for the solvent effect was changed. The new charges were obtained by electrostatic potentials derived from the QC calculations with water phase at the B3LYP/cc-pVTZ//HF/6-31G** level. All QC calculations were carried out using the Gaussian03 program.[19] The IEFPCM continuum solvent model[20,21] implemented in Gaussian03 was applied to mimic a water solvent environment. The atom radii in the united-atom topological parameters[22] were used to construct the solute cavities.

In QC calculations, the electrostatic potentials of each dipeptide model were calculated for two optimized conformations of C5 and αR. The main chain torsion angles of $(\phi, \psi) = (180, 180)$ were set for the C5 conformation and $(-60, -40)$ for the αR conformation as the initial structures. Only the PRO residue required the initial structure of $(\phi, \psi) = (-65, -150)$ for the C5 conformation and $(-61, -35)$ for the αR conformation.[23] The electrostatic potentials of the two conformations were used in the charge fitting in the RESP method, in which effective charges were obtained by the two-stage fitting procedure. In the first stage, the charges for the respective atoms were calculated from the combined electrostatic potential. As for the terminal groups, the charges of acetic acid (ACE) plus ammoniummethylate (NME) were set to 0.0. In the second stage, the chemically equivalent atoms were set to have the same charges. Cα atom, H atom bonding to Cα atom, the chemically equivalent H atoms, and the heavy atoms bonding to the chemically equivalent H atoms were set to be changeable. However, the charges for the rest of atoms were fixed. For example, in the case of the ACE-ALA-NME model, three H atoms bonding to Cβ atom were set to have the same charges, and the charges of the Cα atom, H atom bonding to Cα atom, Cβ atom, and three equivalent H atoms were allowed to change during the charge fitting procedure. The charges of N atom, H atom bonding to N atom, C atom, and O atom of ALA were fixed. The charges of ACE and NME of the terminal groups were also fixed. Finally, the charges of ACE and NME were determined by combining electrostatic potentials of all amino acids. To derive the charges of the charged N- and C-terminal amino acid residues, additional QM calculations were performed under the same condition. These atom charges were calculated according to the method of Cieplak et al.[23]

Protein Structure Prediction by Using the Modified Force Field

Ab initio protein structure prediction was carried out for four proteins: 1J4M,[24] 1LE3,[25] 1L2Y,[26] and 1VII.[27] MD simulations were performed for all proteins on the same machine (CPU: Intel® Xeon™ 1.70 GHz 2CPU, OS: Red Hat Linux 7.3.2). The Sander module in the AMBER 8 package[28] was used for the simulation. The AMBER 8 package was compiled using the Intel Fortran Compiler for IA-32 version 8.1 with Intel Math Kernel Library version 7.2. MPICH version 1.2.6 was used for parallel computation. All initial structures were set to a straight strand form that was constructed with the Leap module in AMBER8. The calculation was started by energy minimization with the steepest descent method in 100 steps. The temperature was elevated to 375 K for 80 ps and kept at 375 K for the

purpose of accelerating the motion of atoms. The MD simulations were performed for 30 ns. The coordinates were stored every 1 ps, and totally 30,080 structures were acquired. The 10 structures that had the most stable E_{GBTOT} were selected from the 30,000 structures excluding the 80 structures in the heating process. Principal component analysis was executed for the 30,080 structures. The structure most distant from the initial structure with respect to principal components among the selected 10 structures was adopted as the final prediction structure. As for the solvent effects, IGB = 5[16,17] (GBSA = 1) and the Langevin dynamics (NTT = 3) with a collision frequency of 1.0 ps^{-1} (GAMMA_LN = 1.0)[29–31] were selected, and *mbondi2* atom radii[16] was used. The integration time step was 1.0 fs. The SHAKE method was not used. Both a cutoff value and RGBMAX value were set to 200 Å for the nonbonded interaction. The RGBMAX parameter concerns the maximum distance between atom pairs involved in the summation appearing in calculation of the effective Born radii. The slowly varying force is evaluated every two steps in the GB method (NRESPA = 2). Default values were used for all other options.

Evaluation of the Reliability of the Modified Force Field in Terms of the Amino Acid Conformations and the Secondary Structure of Proteins

First, the reliability of the modified force field was evaluated in terms of amino acid conformations. The potential energy differences between C5 and αR conformations were computed by QC calculation using the optimized structure for each amino acid that was obtained through the calculation described in the "Determination of Atom Charges with Water-Phase QC Calculations." The potential energy differences were also calculated by MD simulation using the modified force field. We constructed dipeptide models for all amino acids: ALA, ARG, ASN, ASP, CYS, GLN, GLU, GLY, HID, ILE, LEU, LYS, MET, PHE, PRO, SER, THR, TRP, TYR, and VAL. The respective amino acid was inserted between ACE and NME terminating groups (ACE-XXX-NME). These dipeptide models were formed by the "sequence" command in the Leap module. Energy minimization with the steepest descent method was carried out 10,000 steps for all dipeptide models. The minimized structures were heated up to 300 K for 20 ps, and the 6 ns MD simulation was executed at 300 K. Other simulation conditions were the same as those used in the procedure described in the section "Protein Structure Prediction by Using the Modified Force Field."

Atom coordinates were stored every 1 ps, and totally 6000 structures were acquired in the 6 ns MD simulation. The ϕ and ψ values of the main chain torsion angles were obtained for all structures. The structure whose ϕ and ψ values (ϕ : ψ) = (−110 - −30 : −90 - −40), (−180 - −110, 160 - 180 : −180 - −150, 110 - 180), (−110 - −50 : 40 - 110), (40 - 90 : −140 - −30), (−110 - −40 : 110 - 180, −180 - −150), (−180 - −110 : −20 - 50), and (30 - 80 : 10 - 80) are assigned to αR, C5, C7eq, C7ax, β, β2, and αL conformations, respectively. Energy differences of all conformations relative to C7eq (ΔE_{Conf}) were estimated from the appearance probability of the respective conformation, assuming that the probability follows the Boltzmann distribution:

$$\Delta E_{Conf} = E_{Conf} - E_{C7eq} \quad (1)$$

$$E_{Conf} = -kN_A T \log(N_{Conf}) \times \frac{1.0 \times 10^{-3}}{J} \quad (2)$$

$$E_{C7eq} = -kN_A T \log(N_{C7eq}) \times \frac{1.0 \times 10^{-3}}{J} \quad (3)$$

where E_{Conf} is the energy of each conformation, and E_{C7eq} is the energy of C7eq conformation in kcal/mol units. k is the Boltzmann constant: 1.38066 × 10^{-23} J/K. N_A is the Avogadro constant: 6.02214 × 10^{23} mol^{-1}. T is the temperature, which is set to 300 K. J is a reduced constant: 4.184 cal/J. N_{Conf} or N_{C7eq} is the appearance probability calculated from the count of each conformation in the acquired 6000 structures.

Second, the reliability of the modified force field was evaluated in terms of secondary structure of proteins. MD simulation was performed under the condition of IGB = 5 (GBSA = 1) for six small proteins, whose Protein Data Bank (PDB) codes are 1B03,[32] 1J4M,[24] 1LE0,[25] 1LE1,[25] 1LE3,[25] and 1NIZ.[33] 1B03, 1J4M, 1LE0, 1LE1, 1LE3, and 1NIZ have 18, 14, 13, 13, 17, and 16 amino residues, respectively. These six proteins satisfy the following eight conditions: (1) a protein consists of amino acid peptides (not DNA or RNA), (2) the number of residues is less than 30, (3) the protein is a monomer, (4) the protein contains a β-sheet region and more than 30% of residues are assigned to be in the β-sheet region, (5) the protein contains no helix structure, (6) the number of β-sheet regions is one or two, (7) the protein structure is available in PDB as a single molecule (not complex structure), and (8) there is no disulfide bond. 1NIZ has a sequence of "YNKRKRIHIGPGRAFYTTKNIIG" but contains no structural information on the "YNKR" region at the N-terminus and "KNIIG" at the C-terminus. Therefore, ACE and NME are attached to the N-terminus and C-terminus, respectively. HIS residue is assigned to be in the HIP state in which two H atoms are bound to δ and ε positions. The keyword "SALTCON = 0.01" is applied in the Sander module because the structure of 1NIZ was obtained by NMR measurement under the condition of "pH = 5" and "ionic strength = 10 mM." Other proteins are simulated in the neutral condition.

We constructed two kinds of computational models for all proteins: "Native structure" and "All-helix structure," and potential energy differences between the Native structure and the All-helix structure were calculated with MM-PBSA/GBSA calculations. The Native structure was extracted from the original PDB structure. All main chain torsion angles were changed to (ϕ, ψ) = (−60, −40) for the All-helix structure. The positions of side chains of the All-helix structures were modified with the Leap module. Energy minimization with the steepest descent method was executed 500 steps for these two models using the Sander module, followed by 5 ps MD simulation for equilibration. In the equilibration calculation, the initial velocities of atoms were assigned from the Maxwellian distribution in which the total kinetic energy is equivalent to 100 K. In the simulation for equilibration, the temperature was kept constant at 300 K. Atom coordinates were stored every 1 ps, and five structures were obtained after the equilibration. Other simulation conditions were the same as those used in the procedure described in the section "Protein Structure Prediction by Using the Modified Force Field."

The MM-PBSA/GBSA calculations were performed for all five snapshot structures obtained from the above simulations. The MM-PBSA energy contains the following terms:

$$E_{PBTOT} = E_{bond} + E_{angle} + E_{dihedral} + E_{1\text{-}4NB} + E_{1\text{-}4EEL} + E_{vdw} + E_{elec} + E_{PBSURF} + E_{PBCAL} \quad (4)$$

where E_{bond}, E_{angle}, and $E_{dihedral}$ are the bonded terms, $E_{1\text{-}4NB}$, $E_{1\text{-}4EEL}$, E_{vdw}, and E_{elec} are the nonbonded terms, E_{PBSUR} is the hydrophobic interaction in the solvent-free condition, and E_{PBCAL} is the charge-dependent energy computed by the Poisson-Boltzmann (PB) method. In our analysis, the PB total energy was calculated for five snapshot structures and their averaged value was regarded as the computationally evaluated potential energy E_{PBTOT}. The dielectric constant for protein was set to 1.0 and that for the surrounding solvent was set to 80.0. The lattice space was set to 2.0 Å, and the maximum number of iterations to solve the linear PB equation was set to 500. The solvent probe radius for calculating the solvent accessible surface area was set to 1.6 Å. The surface tension was set to 0.005 kcal/mol/Å.[2] The MM-GBSA energy function contains the following terms:

$$E_{GBTOT} = E_{bond} + E_{angle} + E_{dihedral} + E_{1\text{-}4NB} + E_{1\text{-}4EEL} + E_{vdw} + E_{elec} + E_{GBSURF} + E_{GBCAL} \quad (5)$$

where E_{GBSUR} is the hydrophobic interaction in the solvent-free condition and E_{GBCAL} is the charge-dependent energy computed by the GB method. E_{GBSURF} is calculated only if the GBSA = 1 keyword is set on. The average of the GB total energy was calculated from five snapshot structures. Consequently, the energies from the PB and the GB calculations were obtained for each of the Native and the All-helix structures of six proteins.

Finally, to compare the performance of the modified force field with the standard force field, the reliability of ff03 force field was also evaluated in terms of amino acid conformations and secondary structure of proteins as the above methods. Additional comparisons were performed by including other solvent conditions: IGB = 1[12–14] (GBSA = 0), IGB = 1 (GBSA = 1), IGB = 2[15,16] (GBSA = 0), IGB = 2 (GBSA = 1), and IGB = 5 (GBSA = 0) in addition to IGB = 5 (GBSA = 1). Thus, totally six types of GB calculations were performed in the MD simulations with ff03 force field.

In the case of IGB = 1 (GBSA = 0, 1) and IGB = 2 (GBSA = 0, 1), Berendsen's weak coupling scheme[34] (NTT = 1) was applied to keep the temperature constant at 300 K. On the other hand, Langevin dynamics (NTT = 3) was applied with a collision frequency of 1.0 ps^{-1} (GAMMA_LN = 1.0)[29–31] for IGB = 5 (GBSA = 0, 1). The IGB = 1 (GBSA = 0, 1) calculations were executed with the atom radii prepared by Tsui and Case in 2001,[14] which was set by the Leap module. The IGB = 2 (GBSA = 0, 1) and IGB = 5 (GBSA = 0, 1) calculations were executed with the *mbondi2* atom radii.[16]

Execution of MD Simulation with the Explicitly Water-Generated Models

MD simulation for the alanine dipeptide model (ACE-ALA-NME) was performed under the condition that solvent waters were explicitly generated with the TIP3P model.[35] The minimum distance from the peptide to the boundary of the solvated box was set to 10 Å, which resulted in generation of 630 water molecules. The calculation was started from energy minimization with the steepest descent method in 10,000 cycles. The temperature was elevated to 300 K for 20 ps and was maintained at 300 K. The MD simulation was executed for 6 ns. Atom coordinates were stored every 1 ps, and totally 6000 structures were acquired. Energy differences of all conformations against C7eq (ΔE_{Conf}) were estimated from eqs. (3)–(5). The particle mesh Ewald[36–39] method was used to evaluate the long-range interaction. The short-range Coulomb and Lennard-Jones interactions were truncated at 8.0 Å. The pressure was maintained at 1.0 bar using the Berendsen algorithm, and the periodic boundary condition was applied. The integration time step was 1.0 fs. The SHAKE method was used for the covalent bond involving H atoms.

For the purpose of examining the structural stability, 5 ns MD simulations were executed for four mini-proteins and two enzymatic proteins in the section "Stability of Protein Structure." All calculation models were constructed from the respective PDB structure. Each model was solvated with TIP3P waters,[35] using a box periodic boundary condition. To neutralize the charges of the models, Na^+ or Cl^- were generated as counter ions. The particle mesh Ewald method was used for the long-range electrostatic interaction.[36–39] The cutoff distance for the long-range electrostatic and the van der Waals energy terms was set at 12.0 Å. All covalent bonds to hydrogen atoms were constrained using the SHAKE algorithm. Energy minimization was achieved in two steps. First, the movement was allowed only for the water molecules and the ions. Next, all atoms were allowed to move freely. In each step, energy minimization was executed by the steepest descent method for the earlier 2500 steps and the conjugated gradient method for the later 2500 steps. After 80.0 ps heating calculation until 310 K using NVT ensemble, 5.0 ns equilibrating calculation was executed at 1.0 bar and at 310 K using NPT ensemble, with an integration time step of 1.0 fs.

Results

Force Field Parameter Derived from Water-Phase QC Calculations and its Reliability in Terms of Amino Acid Conformations and Secondary Structure of Proteins

Effective charges were obtained by the RESP method with fitting to the electrostatic potential that was obtained by Gaussian03 calculations of the dipeptide models in water phase. The newly derived charges are shown in Table S1 of Supplementary Materials. The conformational stabilities of 20 amino acids were evaluated from the force field modified with the new charge set. The conformational stabilities in MD simulations with the IGB = 5 (GBSA = 1) solvent model are shown in Table 1. The C5 conformation is the most stable in 12 amino acids: ALA, ARG, ASN, ASP, GLN, GLU, GLY, LYS, MET, PHE, TRP, and TYR. The β conformation is the most favorable in seven amino acids: CYS, HID, ILE, LEU, PRO, SER, and THR. VAL is stabilized in the αR conformation. As for the two conformations of

Table 1. Conformational Stabilities of All Amino Acids Obtained by the MD Simulation with the Modified Force Field.

Conformation	C7eq	C5	αR	β	C7ax	β2	αL
ALA	0.00	−2.06	−1.54	−1.90	2.63	−0.12	2.63
ARG	0.00	−1.84	−1.36	−1.65	2.84	−0.12	2.84
ASN	0.00	−1.84	−0.44	−1.80	1.46	1.31	2.89
ASP	0.00	−2.18	−0.37	−2.13	2.58	1.27	2.58
CYS	0.00	−1.41	−1.61	−1.72	2.90	0.00	2.90
GLN	0.00	−1.93	−1.07	−1.80	0.93	0.25	2.80
GLU	0.00	−1.94	−1.44	−1.92	2.70	0.11	2.70
GLY	0.00	−1.55	−0.21	−1.55	−0.44	0.76	0.69
HID	0.00	−1.81	−0.89	−1.81	2.87	0.55	2.87
ILE	0.00	−1.77	−1.21	−1.98	2.77	−0.02	2.77
LEU	0.00	−1.40	−1.07	−1.69	3.06	0.91	3.06
LYS	0.00	−2.28	−1.06	−2.02	2.54	0.76	2.54
MET	0.00	−2.13	−1.08	−1.94	2.65	0.41	2.65
PHE	0.00	−2.11	−1.24	−2.04	2.61	0.36	2.60
PRO	0.00	2.24	−1.89	−2.83	2.24	2.24	2.24
SER	0.00	−1.89	−1.72	−1.89	2.64	0.02	2.64
THR	0.00	−1.56	−1.15	−1.72	2.97	0.52	2.97
TRP	0.00	−2.23	−2.10	−2.17	2.33	−0.65	2.33
TYR	0.00	−2.38	−1.78	−2.12	2.37	−0.49	2.37
VAL	0.00	−1.01	−1.75	−1.30	3.03	−0.44	3.03

The energy is relative to the C7eq conformation in unit of kcal/mol.

C5 and αR, the C5 conformation is more stable than the αR conformation for all amino acids except CYS and VAL.

To evaluate the conformational stability for αR and C5 conformations more precisely, QC calculations were performed in water phase. The main chain torsion angles of the optimized structures and the potential energy differences between C5 and αR conformations are shown in Table 2. In PRO residue, the αR conformation is compared with the β conformation instead of the C5 conformation. All optimized dipeptides retained the initial secondary conformations. The C5 conformation is more stable than the αR conformation for all amino acids except GLY, MET, and THR at the HF/6-31G** level. As for PRO residue, the β conformation is more stable than the αR conformation. The energy differences were re-estimated at the B3LYP/cc-pVTZ level. The C5 conformation is more stable than the αR conformation for all amino acids except MET and THR, and PRO residue is stabilized in the β conformation. As for THR residue, geometry optimization of the β conformation was also performed at the HF/6-31G** level in water phase. The main chain torsion angles were optimized to be $(\phi, \psi) = (-83.135, 122.953)$. The potential energies of C5, αR, and β conformations were estimated at the B3LYP/6-311++g(2d,2p) level. The energy calculation shows that the β conformation is the most stable among the three, and the energy differences between β and αR and between β and C5 are 0.29 kcal/mol and 1.15 kcal/mol, respectively.

The conformational stability deduced from MD simulation with the modified force field was compared with the results of QC calculation. Figure 1 shows the energy differences between C5 and αR conformations using the MD simulation with the IGB = 5 (GBSA = 1) solvent model and the QC calculation.

Table 2. Main Chain Torsion Angles for C5 and αR Conformations and Their Energy Differences Obtained by QC Calculation.

	C5 conformation	αR conformation	ΔE [kcal/mol]
	Φ value	Φ value	HF/6-31G*,*
	Ψ value	Ψ value	B3LYP/cc-pVTZ
ALA	−156.430	−77.965	−0.25
	149.488	−26.212	−0.68
ARG	−152.719	−73.613	−0.75
	130.685	−33.650	−1.27
ASN	−160.502	−79.121	−0.05
	168.992	−25.435	−0.13
ASP	−153.605	−71.835	−1.12
	142.315	−35.688	−1.94
CYS	−152.351	−74.142	−2.93
	116.591	−35.896	−3.36
GLN	−151.871	−73.267	−0.06
	130.591	−34.039	−0.40
GLU	−149.831	−71.670	−0.50
	132.786	−37.715	−0.94
GLY	179.448	−80.579	0.06
	−179.391	−20.353	−0.15
HID	−146.811	−71.139	−3.04
	122.969	−40.912	−3.69
ILE	−151.031	−76.511	−0.93
	137.940	−24.881	−1.38
LEU	−153.976	−71.125	−1.28
	125.770	−37.837	−1.60
LYS	−152.572	−75.245	−0.92
	129.944	−33.186	−1.65
MET	−151.899	−80.757	0.31
	130.993	−24.086	0.01
PHE	−155.532	−68.481	−4.06
	130.682	−41.515	−3.74
PRO	−61.015*	−66.478	−0.73
	147.301*	−30.665	−1.15
SER	−152.127	−73.366	−1.96
	129.351	−43.409	−2.01
THR	−119.944	−94.441	1.20
	165.437	−6.633	0.93
TRP	−153.274	−64.146	−0.40
	126.771	−43.007	−0.50
TYR	−154.148	−70.202	−1.44
	128.923	−40.402	−1.66
VAL	−151.855	−75.300	−0.88
	139.083	−24.666	−1.53

ΔE indicates the energy differences between C5 and αR conformations: $\Delta E = E(\text{C5}) - E(\alpha\text{R})$.
*β-conformation is employed instead of C5 conformation only for PRO.

A fine compatibility is seen between the results of the MD simulation and the QC calculation for all amino acids except four: CYS, MET, THR, and VAL (see Fig. 1). In PRO and THR residues, the β conformation is the most stable in the MD simulation, and this result is consistent with the QC calculation.

The reliability of the modified force field was evaluated using the secondary structure of proteins. The energy differen-

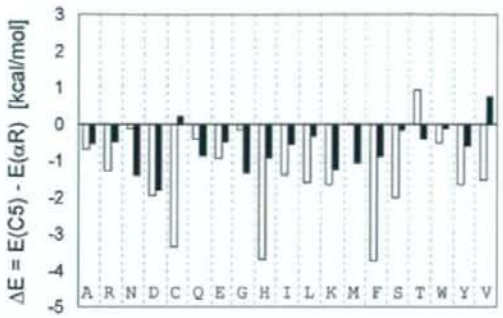

Figure 1. Comparison between QM and MM energies of all dipeptide models except for PRO. Open columns indicate the QM energy differences between C5 and αR conformations in the B3LYP/cc-pVTZ level. Solid columns indicate the energy difference estimated by MD simulation with the modified force field in the IGB = 5 (GBSA = 1) solvent method. All amino acids are expressed by the one-letter code in the horizontal axis.

ces between the Native structure and the All-helix structure of six proteins in MD simulations with the IGB = 5 (GBSA = 1) solvent model are shown in Table 3. ΔE_{GBTOT}, which is the energy difference between the Native and the All-helix structures calculated by the MM-GBSA method [$E_{GBTOT(Native)} - E_{GBTOT(All\text{-}helix)}$], shows a negative value for six proteins: 1B03, 1J4M, 1LE0, 1LE1, 1LE3, and 1NIZ except for 1NIZ. In the evaluation from the energy differences calculated by the PB method, ΔE_{PBTOT} [$E_{PBTOT(Native)} - E_{PBTOT(All\text{-}helix)}$], the Native structure is more stable in five proteins and the All-helix structure is preferred only in 1NIZ. That is, the Native structure is more stable than the All-helix structure for all proteins except 1NIZ. The difference between ΔE_{GBTOT} and ΔE_{PBTOT} is less than 5 kcal/mol for three proteins: 1LE0, 1LE1, and 1NIZ.

The root mean square deviation (RMSD) between the initial and the five snapshot structures for N, Cα, and C atoms of main chain ranges from 1.5 to 2.2 Å for the Native structure of 1B03. The RMSD is considerably small for 1LE0, 1LE1, and 1LE3 (Table 3). Accordingly, the initial secondary conformation was retained during the MD simulation for the Native structure. The RMSD of the All-helix structures for 1B03 ranges from 1.6 to 1.8 Å. The RMSD is relatively large for 1J4M and 1LE3, whereas the other proteins have small values and retain the initial secondary conformation. In 1J4M protein, there are three large side chains of ARG1, LYS3, and TRP4 at the N-terminus, and some atoms of these residues were excessively close to one another in the initial conformation of the All-helix structure. The large RMSD value of 1J4M is due to the relaxation of this initial conformation. The initial secondary conformation is retained in other residues except for ARG1, LYS3, and TRP4. In 1LE3 protein, there are also three large side chains of TRP12, TRP14, and GLU16 at the C-terminus, and the main reason for the large RMSD value is the relaxation of TRP12, TRP14, and GLU16 residues.

Examination of the Reliability of the Modified Force Field, Evaluated in Terms of the Tertiary Structure Obtained from a Trial for the Ab Initio *Protein Structure Prediction*

Ab initio protein structure predictions were performed for four proteins: 1J4M, 1LE3, 1L2Y, and 1VII. 1J4M consists of 14 residues of "RGKWTYNGITYEGR" and forms a β-sheet structure. 1LE3 consists of "GEWTWDDATKTWTWTE-NH2" (16 amino acids and one NH2 group) and also forms a β-sheet structure. π–π interactions are observed among four TRP residues. In contrast, both 1L2Y and 1VII have helix structures. These two proteins are important to examine if the modified force field can predict β-sheet structure as well as helix structure. 1L2Y consists of 20 amino acids of "NLYIQWLKDGGPSSGRPPPS," and 1VII consists of 36 amino acids of "MLSDEDFKAVFGMTRSAFANLPLWKQQNLKKEKGLF."

In the case of 1J4M, 30,000 snapshot structures were acquired through 30 ns MD simulation. The top 10 structures with respect to the energetic stability estimated by E_{GBTOT} were selected from the 30,000 structures. Four elements, i.e., simulation time, energy (E_{GBTOT}), difference from the initial structure, and similarity of the predicted structure to the PDB one for N, Cα, and C atoms of the main chain (RMSD (N, Cα, C)), were examined as shown in Table S2 of Supplementary Materials. Those 10 structures appeared at the simulation times of 6689, 17094, 29747, 9349, 7376, 16041, 16523, 28360, 15922, and 6300 ps in order from the lowest E_{GBTOT}. The difference from the initial structure is derived from principal component analysis for all 30,080 structures. The distance of the component coordinates between the initial structure and the predicted one is defined as the structural difference between them. Since the structure at 9349 ps shows the largest distance of 80.2 among the 10 structures, this structure is regarded as the most different structure from the initial structure. In this study, the structure at 9394 ps is assigned as the final predicted structure. The final predicted structure and the PDB one are shown in Figure 2a. The predicted structure is similar to the PDB one except for the incompleteness in the β-sheet formation. The RMSD for the N, Cα, and C atoms between the predicted and the PDB structures, RMSD(N, Cα, C), is 3.5 Å. This value, 3.5 Å, is the second

Table 3. Comparisons of Energy Differences and RMSD Between the Native Structure and the All-Helix Structure in MD Simulations with the Modified Force Field.

	ΔE_{GB} [kcal/mol]	ΔE_{PB} [kcal/mol]	RMSD native [Å]	RMSD all-helix [Å]
1B03	−7.60	−25.72	1.5–2.2	1.6–1.8
1J4M	−2.01	−18.27	1.1–1.3	2.4–3.3
1LE0	−23.66	−22.66	0.5–0.7	1.1–1.5
1LE1	−1.53	−5.49	0.9–1.1	0.8–0.9
1LE3	−15.80	−10.71	0.6–1.0	2.8–3.5
1NIZ	20.77	15.93	2.0–2.2	1.2–1.4

Energy differences are given by $E_{(Native)} - E_{(All-helix)}$ in both the MM-GBSA and the MM-PBSA methods.
RMSD values represent the deviation range of 5 snap-shot structures from the initial structure for N, Cα, and C atoms of main chain.

smallest among the 10 structures and relatively small among the 30,000 structures (Fig. 3a).

In 1LE3, the most stable 10 structures were selected by the same method described above and four elements were examined: simulation time, energy (E_{GBTOT}), difference from the initial structure, and RMSD (N, Cα, C) (Table S2). Those structures appeared at 9698, 22954, 25384, 3149, 11128, 23917, 1876, 28573, 1898, and 10075 ps in order of E_{GBTOT}. The difference from the initial structure is the largest at 22954 ps. Hence, the structure at 22954 ps is adopted as the final predicted structure. The predicted structure and the PDB one are shown in Figure 2b. The predicted structure is also similar to the PDB one despite the fact that the β-sheet formation is not completed. The RMSD(N, Cα, C) is 3.7 Å, which is the smallest among the 10 structures and is considerably small among the 30,000 structures (Fig. 3b).

In 1L2Y, the most stable 10 structures were also obtained (Table S2), and they appeared at 10288, 29995, 4199, 10647, 28966, 13828, 13535, 10913, 29745, and 12416 ps. The difference from the initial structure is the largest at 29995 ps. Hence, the structure at 29995 ps is adopted as the final predicted structure. The predicted structure and the PDB one are shown in Figure 2c. The helix formation in the PDB structure is adequately reproduced in the predicted one, and the predicted structure is similar to the PDB one in the whole region of the protein. The RMSD(N, Cα, C) is 2.3 Å, which is the smallest among the 10 structures and relatively small among the 30,000 structures (Fig. 3c).

In 1VII, the most stable 10 structures were obtained (Table S2), and they appear at 25207, 26541, 24341, 24067, 25388, 22135, 29431, 25345, 23898, and 28356 ps. The difference from the initial structure is the largest at 25207 ps. Hence, the structure at 25207 ps is adopted as the final predicted structure as shown in Figure 2d. Three helix regions in the PDB structure are reproduced in the predicted structure. The position of the first helix region against the second helix region is deviated from the PDB structure. Therefore, the RMSD(N, Cα, C) shows a large value of 6.4 Å. The most similar three structures appearing at 25207, 25388, and 25343 ps are also the most different three structures from the initial one, and the distance values of these three structures, 271.6, 271.1 and 270.2, are very close to each other. The RMSD(N, Cα, C) of 6.4 Å is relatively small among the 30000 structures (Fig. 3d).

Evaluation of a Standard Force Field

The energetic stability on the respective conformation was surveyed for the normal 20 amino acids when MD simulation is executed with ff03 force field in the IGB = 1 (GBSA = 0) solvent model as shown in Table S3(a) of Supplementary Materials. The β conformation is the most stable for six amino acids and the αR conformation is the most favorable for 11 amino acids. Three amino acids, GLN, GLU, and VAL, are stabilized in the C5 conformation. A comparison of the two conformations of C5 and αR for all amino acids except PRO, which is obviously disadvantageous for the C5 conformation, shows that 11 amino acids prefer the αR conformation. The C5 conformation is more stable than αR in the other eight amino acids. In the IGB = 1 (GBSA = 1) solvent model [Table S3(b)], a comparison of conformational stability shows that the αR conformation is more stable than C5 in 14 amino acids. The αR conformation is more stable than C5 in 15 amino acids in the IGB = 2 (GBSA = 0) solvent model [Table S3(c)], and 16 amino acids prefer the αR conformation than C5 in conformational stability in the IGB = 2 (GBSA = 1) solvent model [Table S3(d)]. In the IGB = 5 (GBSA = 0) and IGB = 5 (GBSA = 1) solvent model [Table S3(e) and Table S3(f)], 16 and 13 amino acids prefer the αR conformation than C5, respectively. Consequently, the αR conformation has turned out to be energetically favorable in all solvent models of MD simulation with ff03 force field.

The conformational stability deduced from MD simulation with ff03 force field was compared with the results of QC calculation for αR and C5 conformations. Figure S1a shows the energy differences between the C5 and αR conformations using the MD simulation with the IGB = 1 (GBSA = 0, 1) solvent model and the QC calculation. In the case of GBSA = 0, the conformational stability in MD simulation conflicts with the QC calculation for nine amino acids. These nine amino acids are stabilized in the αR conformation in MD simulation, whereas the C5 conformation is more stable in QC calculation. In the case of GBSA = 1, the αR conformation is more stable than the C5 conformation in 11 amino acids. The energy differences between the C5 and αR conformations using the IGB = 2 (GBSA = 0, 1) solvent model is shown in Figure S1b. Thirteen amino acids in the case of GBSA = 0 and 15 amino acids in the case of GBSA = 1 are incompatible with the results of QC calculation. In the case of IGB = 5 (GBSA = 0, 1) solvent model (Fig. S1c), 14 amino acids and 11 amino acids are incompatible with the results of QC calculation, respectively. The MD simulation was shown to be inconsistent with the QC calculation for the conformational stability of most of the amino acids.

To examine the influence of GB method, the MD simulation explicitly including water molecules was executed only for the ACE-ALA-NME dipeptide model with the calculation condition described in the section "Execution of MD Simulation with the Explicitly Water Generated Models." The conformational stability is shown in Table S4 of Supplementary Materials. The αR conformation is more stable than the C5 conformation by 0.4 kcal/mol in MD simulation, but the C5 conformation is more stable in QC calculation. Therefore, no improvement is seen in the compatibility between the results of MD simulation and QC calculation even though explicit water molecules are included in the calculation model.

The energy differences between the Native and the All-helix structures of all proteins in six solvent models are shown in Table S5 of Supplementary Materials. In the IGB = 5 (GBSA = 1) solvent model, ΔE_{GBTOT} is 1.84 kcal/mol for 1B03. This means that the All-helix structure is more stable than the Native structure. ΔE_{GBTOT} also shows a positive value for the proteins 1J4M, 1LE1, 1LE3, and 1NIZ. That is, all proteins except for 1LE0 are stable in the All-helix structure. ΔE_{GBTOT} also shows a positive value except for 1B03 and 1LE0. As for the IGB = 5 (GBSA = 0) solvent model, the Native structure is more stable in three proteins in ΔE_{GBTOT} and two proteins in ΔE_{PBTOT}.

In ΔE_{GBTOT} and ΔE_{PBTOT} of the IGB = 2 (GBSA = 1) solvent model, the Native structure is more stable in four proteins.

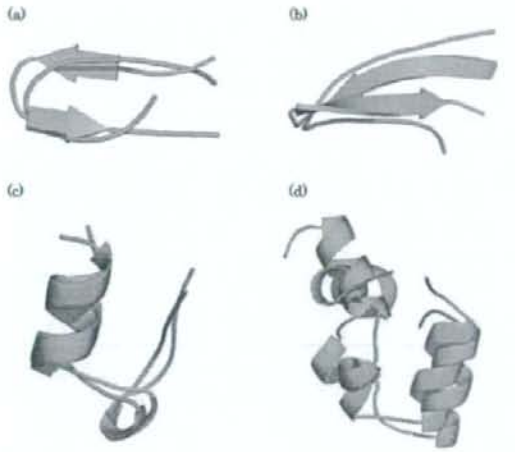

Figure 2. Predicted structures (cyan) and PDB structures (green) represented in cartoon style. Protein structure predictions were performed for four proteins, (a) 1J4M, (b) 1LE3, (c) 1L2Y, and (d) 1VII, using the modified force field.

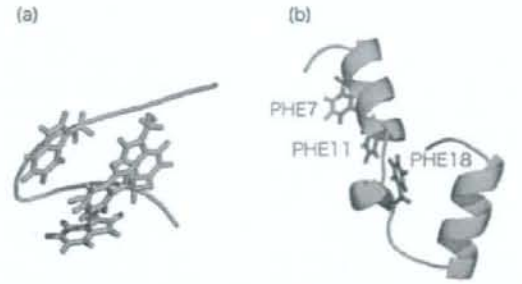

Figure 4. Predicted structures for (a) 1LE3 and (b) 1VII. Four TRP residues barely form obvious π-π interactions among them in (a). PHE18 makes a π–π interaction with PHE11 but does not interact with PHE7 in (b).

Every initial secondary conformation except for the N- and C-terminus of 1J4M and 1LE3 is retained during the MD simulation. In the IGB = 2 (GBSA = 0) solvent model, three proteins, 1B03, 1J4M, and 1NIZ, have large RMSD values for the Native structure and form a completely different structure from the initial β-sheet structure. In the All-helix structure, all proteins

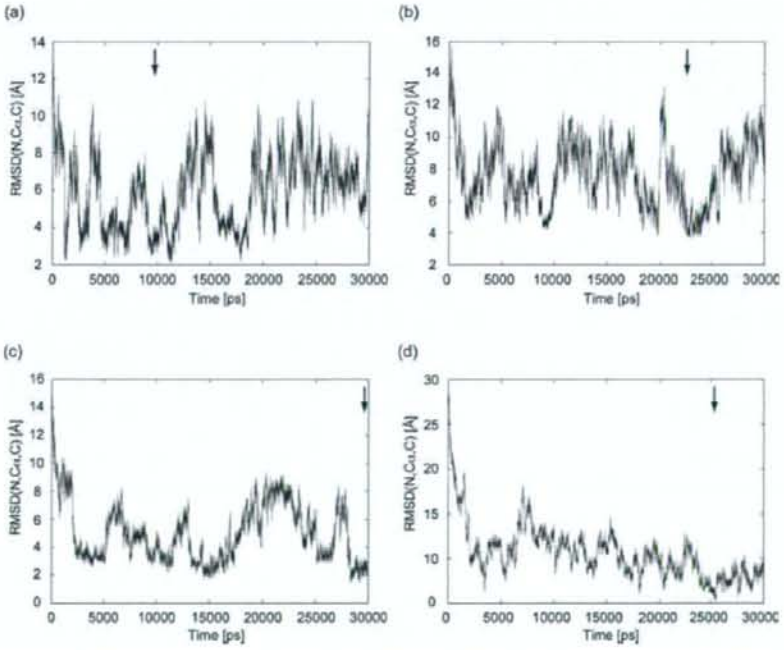

Figure 3. RMSD (N, Cα, C) plot relative to the PDB structure for all acquired structures during the simulations for (a) 1J4M, (b) 1LE3, (c) 1L2Y, and (d) 1VII. The arrow indicates the time point that the predicted structure was selected from.

except for 1LE3 have large RMSD values and form a completely different structure from the initial helix conformation. Hence, the energies were compared only for the 1LE3 protein. The Native structure is more stable than the All-helix structure in 1LE3. For the other proteins that could not retain the initial conformation, a sudden structure change caused by the electrostatic repulsion of charged amino acid residues is observed in the heating process. In the IGB = 1 (GBSA = 1) solvent model, the Native structure of 1J4M and the All-helix structure of 1B03, 1LE1, and 1NIZ form a completely different structure from the initial secondary conformation. Therefore, energy comparison between the helix and the β-sheet structures was performed only for 1LE0 and 1LE3 proteins. The Native structures of 1LE0 and 1LE3 are more stable than the All-helix structure. In the IGB = 1 (GBSA = 0) solvent model, energy comparison was performed only for 1LE1 protein, in which the Native structure was preferred.

Discussion

Modified Force Field

Reliability Evaluated from Amino Acid Conformations and Secondary Structure of Proteins

As for the difference in stability between αR and C5 conformations, the results of MD simulation are compatible with those of QC calculation for all amino acids except CYS, MET, THR, and VAL. The problem of excessive tendency to form a helix structure of ff03 force field has been modified. As long as THR residue, the β conformation is the most stable in both MD simulation and QC calculation. Therefore, the force field parameter of THR will also be acceptable.

The Native structure is more stable than the All-helix structure for all proteins except 1NIZ. Accordingly, the modified force field will be acceptable in terms of the secondary structure. 1NIZ is a part of the V3 loop region of gp120 of HIV-1. The full sequence of this region is "CTRPNYNKRKRIHIGPGRAFYTTKNIIGTIRQAHC" and it corresponds to the 301st to 335th residues of gp120 protein. A disulfide bond is formed between CYS301 and CYS335. Whereas 1NIZ is registered as a single molecule in PDB, this structure was determined by NMR spectroscopy in a complex with an HIV-1 neutralizing antibody.[33] Therefore, the partial model of ACE-KRIHIGPGRAFYTT-NME will not reflect the NMR experimental condition of 1NIZ.

The protein 1CE4, which has an amino acid sequence very similar to that of 1NIZ, was synthesized by a peptide synthesizer and was determined by NMR spectroscopy as a single molecule.[40] 1CE4 forms a disulfide bond like the V3 loop region and consists of bend, turn, and helix structures instead of the β-sheet structure seen in 1NIZ. Thus, there is no β-sheet structure in 1CE4. Furthermore, the C-terminal half side of 1CE4 has α helix structure (Fig. S2 in Supplementary Materials). Some amino acids on the N-terminal side also show a helix structure. The disulfide bond of 1CE4 strongly assists the folding of the protein and makes a twist at the center region of the protein. From this point of view, if there is no disulfide bond, most parts of 1CE4 will be stabilized in the helix structure. Accordingly, it is plausible that the All-helix structure is more stable than the Native structure in 1NIZ. In contrast, the proteins of 1B03, 1J4M, 1LE0, 1LE1, and 1LE3 were synthesized by a peptide synthesizer. Their structures were determined by NMR spectroscopy as a single molecule, and contains many β-sheet and β-turn conformations. As a result, the modified force field is confirmed to accurately describe the β-sheet structure, and the helix structure will be also reproducible as shown in the results for 1NIZ and 1CE4.

Reliability Evaluated in Terms of Tertiary Structure of Proteins and Application to Protein Structure Prediction

Protein structure prediction was performed for 1J4M and 1LE3 molecules containing the β-sheet structure and also for 1L2Y and 1VII molecules containing the α-helix structure. The accuracy of the structural prediction was satisfactory for all molecules. In the case of 1LE3, the side chains of four TRP residues gather in a row due to their π-π interactions in the PDB structure. However, the π-π interactions are not observed in the prediction structure (Fig. 4a). Since the kinetic energy is large because of the MD simulation at 375 K, the weak π-π interaction disappears and will not be observed in the predicted structure. The PDB structure of 1VII protein also shows two π-π interactions of PHE7-PHE18 and PHE11-PHE18 between the first helix and the second helix regions. These π-π interactions are not observed in the predicted structure of 1VII (Fig. 4b). For more precise protein prediction, the proteins containing weak π-π interaction should be simulated again at a lower temperature starting from the structures obtained in the present prediction, or using the simulation technique explicitly implementing the effect of π-π interaction.[41] For 1J4M, the whole structure was accurately predicted, but the β-sheet structure is still loose. 1J4M also needs to be simulated again at a lower temperature or by the π-π interaction implemented method.

Figure 5. Residue-averaged B-factors derived from the calculation with the modified force field (blue), with ff03 force field (green), and the experimental measurement by X-ray crystallography (red). The B-factors for β-lactamase computed in the explicit and implicit water conditions are shown in (a) and (a'), respectively. Those for HIV-1 protease in the explicit and implicit water conditions are in (b) and (b').

Figure 6. RMSD plot for main chain atoms during MD simulations with the modified force field (blue thick line) and the ff03 force field (green broken line). Simulations were executed at 310 K under the explicitly water-generated condition both for mini-proteins (a–d) and enzymatic proteins (e–f).

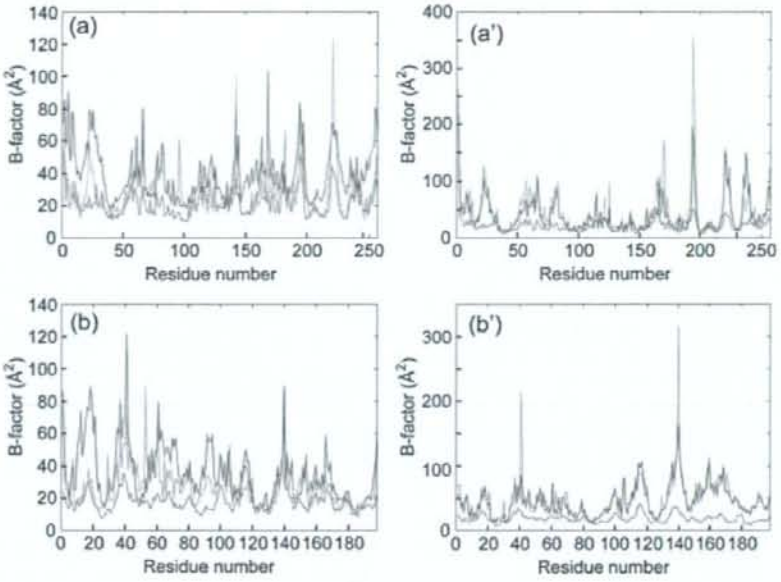

Figure 5.

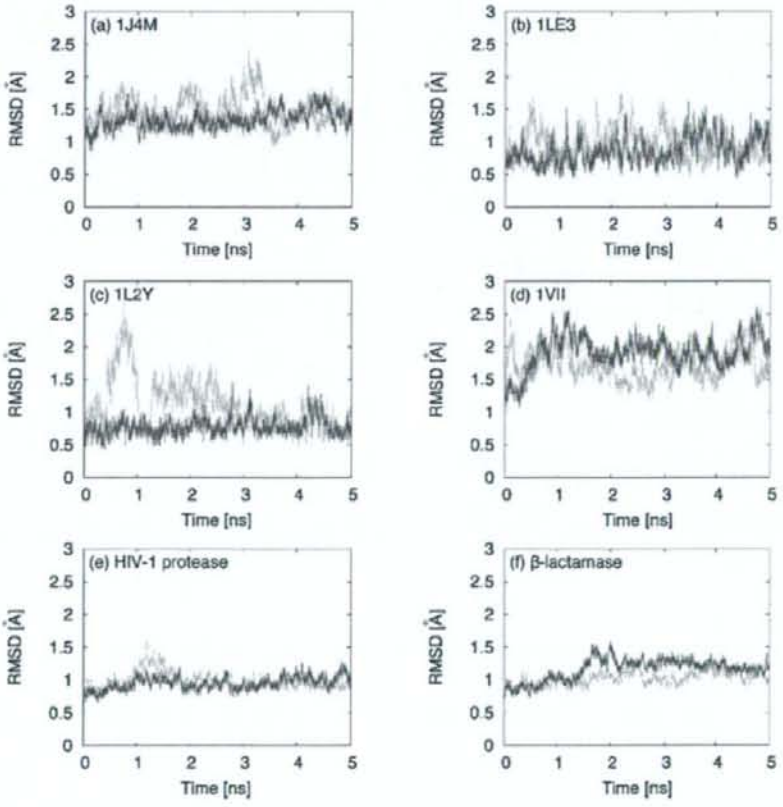

Figure 6.

The predicted structure of 1L2Y was accurate enough. Many previous studies have shown the success of structure prediction for 1L2Y protein with the conventional force field.[42–45] In the PDB structure of 1L2Y, the first helix region consists of eight residues of LEU2, TYR3, ILE4, GLU5, TRP6, LEU7, LYS8, and ASP9, the turn region consists of two residues of GLY10 and GLY11, the second helix region consists of three residues of PRO12, SER13, and SER14, and the strand region consists of five residues of GLY15, ARG16, PRO17, PRO18, and PRO19. Eight residues of the first helix region form an αR conformation and are stabilized as the helix structure. The main chain of GLY residue comprising the turn region is flexible and its conformational degree of freedom is the largest among the normal 20 amino acids. Three residues of the second helix region form the αR conformation and are stabilized as the 3-10 helix structure. In the strand region, GLY and ARG residues form the C5 conformation and three PRO residues form the β conformation. Considering the conformational stability of all amino acids with ff03 force field [Table S3(a–f)], five amino acids, ASP, ILE, LEU, TRP, and TYR, included in the first helix region are stabilized in the αR conformation in almost all solvent models. That is, the residues that stabilize the helix structure are fortunately assigned at the helix region of this protein. PRO and SER in the second helix region are stabilized in the β and αR conformations in ff03 force field, respectively. Hence, this second helix region has an excessive tendency to form the helix structure by SER. GLY, ARG, and PRO residues at the strand region are stabilized in the β conformation. Three PRO residues prefer the β conformation in ff03 force field. GLY and ARG are stabilized in the C5 conformation in almost all cases in Table S3. The residues whose parameters stabilize the strand structure are also fortunately located in the strand region of the protein. PRO is always stabilized in the β conformation because of its structural peculiarity of main chain. Accordingly, the structure of 1L2Y can be predicted easily although the force field that has the excessive helical tendency is used, which results in many reports on the success of 1L2Y prediction.[42–45] The modified force field gives an accurately predicted structure for 1L2Y without any excessive helical tendency in parameter. This suggests that the modified force field can stabilize the helix structure from the hydrogen bond formations.

Flexibility of Proteins

To evaluate the flexibility of proteins those consist of 200–300 amino acid residues, B-factors of β-lactamase from *Staphylococcus aureus* PC-1 (PDB code : 3BLM),[46,47] and human immunodeficiency virus type 1 (HIV-1) protease (PDB code : 1OHR)[48,49] have been compared among three methods: the calculation with the modified force field, the calculation with ff03 force field, and the measurement in X-ray crystallography (see Fig. 5). The B-factors from the calculations with the modified and ff03 force fields were obtained by using the Ptraj program after the 10 ns MD simulations both with the implicit and the explicit water conditions. The B-factors were estimated in the average for the last 1 ns MD simulation. The keywords of IGB = 5 and GBSA = 1 were employed in the simulation with the implicit water condition. The simulation with the explicit water condition was performed according to the Ross Walkers method.[50] The B-factors derived from the modified force field were similar to that from ff03 force field. As for the MD simulation with the implicit water condition (Figs. 5a' and b'), a large B-factor value in ff03 force field, that was over 300 Å^2 at the 194th residue of β-lactamase, was improved in the modified force field. The B-factor values of HIV-1 protease were also improved at the 41st and 140th residues by using the modified force field. The simulated B-factors tend to be large compared with the B-factors observed in X-ray crystallography as seen in the previous studies.[47,49,51] The results on the B-factor suggest that the modified force field can simulate and analyze protein function as effective as ff03 force field from the viewpoint of flexibility of proteins.

Stability of Protein Structure

In this study, atom charges were determined from QC calculations in water phase. Atom charges in the standard force field had been derived from the vacuum- or ether-phase condition.[3] Hence, our modified parameters will be more suitable for describing the structure of such regions that are exposed to solvent. This advantage, however, may cause a drawback for overemphasizing hydrophilic effect and relatively underestimating hydrophobic one. The energy balance between hydrophilic and hydrophobic interactions is critically important to keep an appropriate protein structure in computer simulation. To assess the stability of protein structure in MD simulation with our charge-modified force field, MD simulations were performed and RMSD during the simulation was evaluated for all mini-proteins exemplified in the section "Evaluation of a Standard Force Field" and two enzymatic proteins instanced in the previous section.

MD simulations of 5 ns at 310 K were executed for all mini-proteins and enzymatic ones. All calculations were performed under the calculation condition described in the section "Execution of MD Simulation with the Explicitly Water-Generated Models." The change of RMSD values during the simulation are shown in Figure 6. The fluctuation in RMSD becomes small after 3.5 ns for 1J4M, 1LE3, and 1L2Y. It is confirmed from principal component analysis of Figure S3 in Supplementary Materials that the structure is satisfactorily equilibrated after 4.0 ns. For 1LE3 and 1L2Y, the averaged RMSD values are no longer than 1 Å measured from the PDB structure. For 1J4M, the averaged RMSD value is slightly large, 1.5 Å. In contrast, 1VII shows a certain degree of difficulty in equilibration as seen in both of RMSD plot and PCA map. This is due to the flexibility of N- and C- terminus of 1VII. In every case of mini-proteins, the fluctuation in structure during MD simulation is almost in the similar level between the charge-modified force field and the standard ff03 force field. As long as mini-proteins, no noticeable problem is found in the presently proposed force field from the viewpoints of structural stability.

To examine the structural stability using nonartificial proteins larger than mini-proteins, we executed MD simulation for two enzymatic proteins: HIV-1 protease and β-lactamase. HIV-1

Table 4. Comparison of Number Surface Atoms with That of Buried Ones for Several Proteins.

PDB code	Number of residues	Surface atoms/buried atoms	
2HBO	157	632/405	Thioesterase superfamily protein
1AJ6	219	834/625	DNA gyrase B
7HVP	99 * 2	878/669	HIV-1 protease
1A2F	291	1260/1091	Cytochrome C peroxidase
1OG5	208 * 2	1713/1567	Human cytochrome p450
1A7T	232 * 2	1805/1683	Metallo β-lactamase
117E	282 * 2	2270/2196	Pyrophosphatases
1A88	275 * 3	3017/3298	Chloroperoxidase L
12E8	214 * 4	3591/3071	Antibody Fab fragment
1ARZ	273 * 4	4201/3677	Dihydrodipicolinate reductase

Asterisk indicates that the protein structure is obtained in dimer, trimer, or tetramer.

protease consists of 198 amino acid residues and β-lactamase, class A β-lactamase in this calculation, contains 257 residues. In both cases, the change in RMSD value during MD simulation is fairly small compared with that for mini-proteins. A comparison of RMSD values between the charge-modified force field and the standard ff03 shows no significant difference in both cases of HIV-1 protease and β-lactamase. The RMSD values were also confirmed to be on a similar level to the previous computational studies on these proteins.[52–55] As far as the simulations for enzymatic proteins are concerned, the charge-modified force field leads to no serious error from the viewpoints of structural stability. Judging from the fluctuation in structure during MD simulations for mini-proteins and enzymatic proteins, the presently proposed force field satisfactorily provides the stably equilibrated structure and hardly induces serious error owing to the overestimation of solvent effect.

Importance of Solvent Effects of Waters

The dipeptide model is very useful for investigating the stability of the main chain torsion angles and the reliability of the force field parameters. For the ACE-ALA-NME model, detailed investigation has been performed on the conformational stability with gas-, ether- and water-phase QC calculations by Wang and Duan[56] According to their study, the dipeptide model was stabilized in the C7eq conformation in both gas- and ether phase and the αR, C5, and β conformations were less stable. In contrast, the αR, C5, and β conformations are more stable than the C7eq conformation in water phase. The C5 conformation is the most stable among these three conformations. The ϕ and ψ values of the main chain torsion angles are $(\phi, \psi) = (-70.5, -32.1)$ and $(-156.4, 143.8)$ in the optimized structures for αR and C5 conformations in water-phase calculation at the MP2/6-31G** level by Duan et al. In our QC calculations, the αR conformation is located at $(-77.9, -26.2)$ and the C5 conformation is located at $(-156.4, 149.5)$ at the HF/6-31G** level. Our results reproduced the results of Duan et al. in spite of our calculation level of HF/6-31G** being lower than theirs.

The gas- and ether-phase QC calculations by Wang and Duan [56] suggest that the protein folding structure cannot be correctly generated in a low-permittivity condition because electrostatic interaction is emphasized in a low-permittivity condition. In their results, the dipeptide model is strongly stabilized in the C7eq conformation, which forms a hydrogen bond between the O atom of ACE and the H atom bonding to N atom of NME. Therefore, it is reasonable to assume that the helix structure is easily stabilized in the condition of low permittivity because the intramolecular interaction is effective compared to the intermolecular interaction like hydrogen bonds between protein and water. Indeed, there are some proteins that have a strong tendency to form a helix structure in the low-permittivity condition.[57–59] In contrast, the C7eq conformation is less stable than the αR, C5, and β conformations in water phase. In particular, the C5 and β conformations are important in the condition of high permittivity since these conformations correspond to the β-sheet structure and make an intermolecular interaction with water. According to the results of the previous studies,[9,10] water molecules are indispensable for correct protein folding. Therefore, the protein folding structure will be accurately predictable by using the force field parameters reflecting the water-phase QC calculation, especially for the proteins such as mini-protein and the proteins containing β-sheet region. Accordingly, charge parameter modification is one of the promising approaches to precisely describe the solvent effect.

Water molecules play an important role in protein folding. In particular, the β-sheet structure is known to be significantly stabilized by the presence of water molecules.[9–10] In the present work, the reliability of protein structure prediction has increased by adopting the modified atom charges derived from the water-phase QM calculations. This implies that a fairly large number of residues in proteins are exposed to solvent water and located in the high-permittivity environment. For the purpose of examining the access of waters to the residues in proteins, we have calculated the ratio of the number of surface atoms to that of buried ones as shown in Table 4. Several globular proteins, whose number of residues is more than 150, are selected and the number of surface atoms are calculated by using a software program computing the solvent accessible surface area, GETAREA 1.1.[60] The calculated result indicated that the number of surface

atoms is almost comparable with that of buried ones or larger. That is, many atoms in proteins can be directly influenced by solvent waters. The permittivity will be fairly high due to the influence of waters although it may be true that the permittivity in the core of proteins is about 4.0 like in ether. Accordingly it is reasonable that the charge modification has improved the reliability in computation on the folding structure of soluble proteins.

Effective Solvent Model in GB Method

Six kinds of methods describing solvent effect in GB calculation were examined from the secondary conformations of proteins. In the three solvent models, IGB = 1 (GBSA = 0), IGB = 1 (GBSA = 1), and IGB = 2 (GBSA = 0), serious changes in protein structures were often observed. The initial structures of five proteins (1B03, 1J4M, 1LE0, 1LE1, 1NIZ) for IGB = 2 (GBSA = 0), three proteins (1B03, 1LE1, 1NIZ) for IGB = 1 (GBSA = 1), and four proteins (1B03, 1LE0, 1LE3, 1NIZ) for IGB = 1 (GBSA = 0) were drastically changed only for 5 ps MD simulation, and the RMSD values became large. 1B03 protein includes five positively charged amino acids: ARG1, LYS2, ARG5, ARG8, and ARG12. Strong repulsion forces among these five amino acids caused a sudden structural change. Five proteins, 1J4M, 1LE0, 1LE1, 1LE3 and 1NIZ, include several positively and negatively charged amino acids (Table S6 in Supplementary Materials). Sudden structural change caused by the electrostatic interaction has been frequently observed in these proteins. This suggests that the electrostatic interaction is excessively estimated in these three solvent models and that these solvent models are inadequate when used in combination with ff03 force field. In contrast, there was no sudden structural change in three solvent models: IGB = 2 (GBSA = 1), IGB = 5 (GBSA = 0), and IGB = 5 (GBSA = 1). In the IGB = 2 (GBSA = 1) solvent model, three proteins show a small energy difference less than 5 kcal/mol between ΔE_{GBTOT} and ΔE_{PBTOT}. The energy difference is also small for two proteins in the IGB = 5 (GBSA = 0) model and for four proteins in the IGB = 5 (GBSA = 1) model. These results suggest that the solvent model IGB = 5 (GBSA = 1) is the most favorable among the six methods describing solvent effect from the viewpoint of energy and structure.

Protein Structure Calculated with ff03 Force Field

The charge parameters of ff03 force field were reported to be obtained by QC calculations in ether phase (ε = 4.335). The charge parameters provided before ff03 force field had been obtained by QC calculation in gas phase. In MD simulation, the solvent effects had been incorporated only by using dielectric cavity methods[12–17] or explicit water molecules.[35,61–64] Therefore, the advantage of ff03 force field is to include the solvent effect in itself. This advantage will appear especially in the improvement of the main chain torsion angles of proteins. According to the ΔE_{GBTOT} calculation with the IGB = 5 (GBSA = 1) solvent model, the All-helix structure is more stable than the Native structure except for 1LE0 protein in the calculation with ff03 force field. The solvent models with IGB = 2 (GBSA = 1) and IGB = 5 (GBSA = 0) also show energetic preference for the All-helix structure in some proteins. These results clearly show that MD simulation using ff03 force field still cannot simulate the proteins accurately enough, especially when water molecules strongly influence the protein structure.

Both αR and β conformations are often observed as the most stable conformation in protein peptides because the ϕ angle of main chain is greatly stabilized at $-60°$. The C5 conformation, ϕ angle of which is around 180°, is rather unstable. The helix structure is normally stabilized by hydrogen bonds of main chain atoms between each residue and its fourth neighboring. On the other hand, the dipeptide model consists of only one amino acid residue, NME, and ACE. Hence, this model cannot form a hydrogen bond like the helix structure. However, the dipeptide model is stabilized in the αR conformation in MD simulations for many amino acids with every solvent method as shown in Table S3(a–f). MD simulation was also performed under the condition of explicit generation of water molecules for the ACE-ALA-NME dipeptide model. This dipeptide model resulted in stabilization of the αR conformation, and ϕ angle of main chain was also greatly stabilized at $-60°$. That is, no improvement was seen in the results of MD simulations in spite of the explicit generation of water molecules. These results suggest that ff03 force field seems intrinsically to have a tendency to lead the helix structure excessively in spite of the choice of solvent models or the incorporation of explicit waters.

For comparison, we have examined the conformational stability of secondary structures in MD simulation with ff99 force field using the ACE-ALA-NME dipeptide model. The ff99 force field was one of the most broadly used parameters before the release of ff03. The energies of each conformation estimated through MD simulation with ff99 force field are shown in Table S7 of Supplementary Materials. As long as ALA, the αR conformation is always the most energetically stable irrespective of the choice of GB calculation method. This indicates that ff99 force field also has a tendency to lead the α-helix structure excessively. A comparison of conformational stability between ff99 (Table S7) and ff03 (Table S3) force fields suggests that the tendency for helix is more serious in ff99. Hence, it is true that ff03 force field improved the overestimation of helix stability to a certain extent. In the conversion from ff99 to ff03 force field, a new atom type, H0, which is the hydrogen atom bonding to Cα atom of GLY, was introduced. As far as bonded terms are concerned, only the dihedral parameters for ϕ and ψ angles for main chain and the dihedral parameters concerning Cβ and three main chain atoms were modified. In our present study, the parameter values for the bonded terms are unchanged from ff03 force field, but only the atom charges are modified. This charge modification was demonstrated to diminish the tendency for helix conformation and to be effective for the prediction of protein folding structure.

Conclusions

Currently one of the most widely accepted force fields, ff03, still tends to overestimate the stabilization energy of the helix structure. Hence, a mini-protein structure or β-sheet structure, whose