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

| Protein                  | PDB ID | Appearance ratio         |                  |
|--------------------------|--------|--------------------------|------------------|
| 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)  |

*aThe 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.

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 spheri-cal, the binding mode cannot be simply determined only by aligning the direction of the principal axes of inertia.

http://www.sssj.org/ejssnt (J-Stage: http://www.jstage.jst.go.jp/browse/ejssnt/)
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 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. Thus, we tried to determine the best orientation of the ligand for β-Lac, and Chemscore, ASP, and Glidescore suggested the consistent result with the crystal structure. The best orientation is defined by theoretical selection by considering docking affinity between the target protein and the docked ligand. However, in practice, 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 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 was 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 \textit{in silico} screening of compounds.

\section*{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 \textit{et al.} \cite{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.

\section*{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.

\section*{References}

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