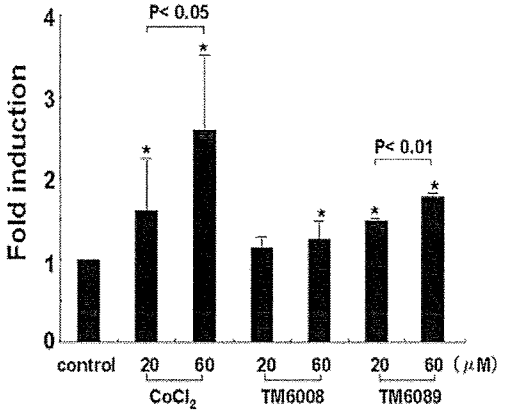

Figure 1. Stimulation of HIF-dependent luciferase reporter gene expression. IRPTC expressing the 7xHRE/Luc plasmid were incubated with the tested compounds. Cobalt chloride was used as a positive control. Results from 3 independent experiments are averaged and shown as fold increase above unstimulated control cells. *P<0.05 vs control.

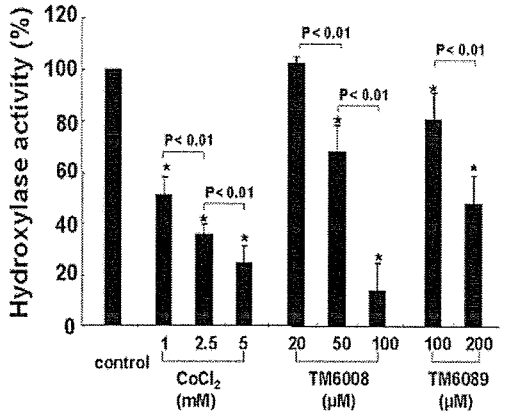

Figure 2. Inhibition of PHD activity. Results from 3 independent experiments are averaged and shown as the percentage of inhibition of ODD-dependent hydroxylase activity. *P<0.01 vs control.

Docking Simulations

The X-ray crystal structure of human PHD2 was obtained from the Protein Data Bank[17] (PDB code: 2HBT). Throughout the present study, the software system MOE (Molecular Operating Environment, version 2005.06) and the MMFF94s force field[18] were used. Binding sites were characterized using the alpha site finder function[19] in MOE. The docking of small molecules and the target sites was performed by the program Ph4Dock.[20]

PHD Activity

PHD activity was determined as described by Kaule et al.[21] In brief, mitochondrial fraction of IRPTC homogenates was reacted with the tested compounds and ODD peptide of HIF-1α. ODD-dependent hydroxylase activity was assessed by counting the radioactivity of [1-^{14}C]-succinate converted from [5-^{14}C]-2-OG by PHD.

Transition Metal Chelation

The chelating activity of the tested compounds for transition metal ions was measured by the method of Price et al[22] with some modifications.

Capillary Network Formation

Capillary network formation was examined by Matrigel assays (BD Biosciences) as described previously.[23]

Sponge Assays

Sponge angiogenesis assays were performed as described previously.[24]

Hypoxia-Sensing Transgenic Rat

Stimulation of the HIF-HRE system by systemic administration of TM6008 or TM6089 was evaluated using the hypoxia-sensing transgenic rat strain.[24] Expression of the hypoxia-responsive luciferase gene was estimated by semiquantitative RT-PCR as described previously.[25]

Cerebral Ischemic Injury Model

Transient global ischemia of Mongolian gerbils was achieved by bilateral carotid occlusion.[26] Animals were then randomly divided into 3 experimental groups: Groups 1 (TM6008) and 2 (vehicle) animals underwent transient global ischemia. Group 3 animals were sham-operated and served as controls.

We also measured cortical microperfusion by laser-Doppler flowmetry in gerbil forebrain ischemia treated with TM6008 or vehicle.

Statistics

Differences among groups were assessed by Kruskal-Wallis test or ANOVA. The statistical significance was determined by 2-tailed Mann–Whitney U test or Student t test. Data are expressed as means±SD. Values are considered significant at P<0.05.

Results

Identification of Novel HIF-Stimulating Compounds

Thirty-seven compounds that have structural similarities to FG-0041, a previously reported PHD inhibitor supposedly acting through iron chelation,[27] were selected from a chemical database. The chemical structures of these compounds are shown in supplemental Figure I. Their HIF-stimulating activity was tested by an in vitro screening assay which used cells expressing luciferase controlled by hypoxia responsive element (HRE) (supplemental Figure II). Cobalt, a well known chemical mimicker of hypoxia by stabilizing HIF-α subunit,[28] was used as a positive control. Two derivatives, TM6008 and TM6089, exhibited strong HIF-stimulating activities (Figure 1). TM6008 is 6-amino-1, 3-di methyl-5-(2-pyridin-2-yl-quinoline-4-carbonyl)-1H-pyrimidine-2, 4-dione, and TM6089 is 6-amino-1,3-di-methyl-5-[2-(pyridin-2-ylsulfanyl)-acetyl]-1H-pyrimidine-2,4-dione.

PHD Inhibition

The inhibitory effect of our compounds on the oxygen-dependent hydroxylation of HIF-α subunit by PHD was evaluated. All tested compounds inhibited PHD activity in a dose-dependent manner (Figure 2). TM6008 was the most effective, exceeding cobalt chloride.

In Vitro Transition Metal Chelation of PHD Inhibitors

Previously reported PHD inhibitors, such as 3,4-DHB,[29] S956711,[29] and FG-0041,[27] share an iron chelating motif. Although chemical structures of TM6008 and TM6089 differ significantly from previous PHD inhibitors, TM6008 also share this motif. By contrast, TM6089 lacks this motif.

We therefore evaluated their abilities to chelate transition metals in vitro by copper-catalyzed oxidation of ascorbic acid. 3,4-DHB, S965711, and TM6008 chelated transition

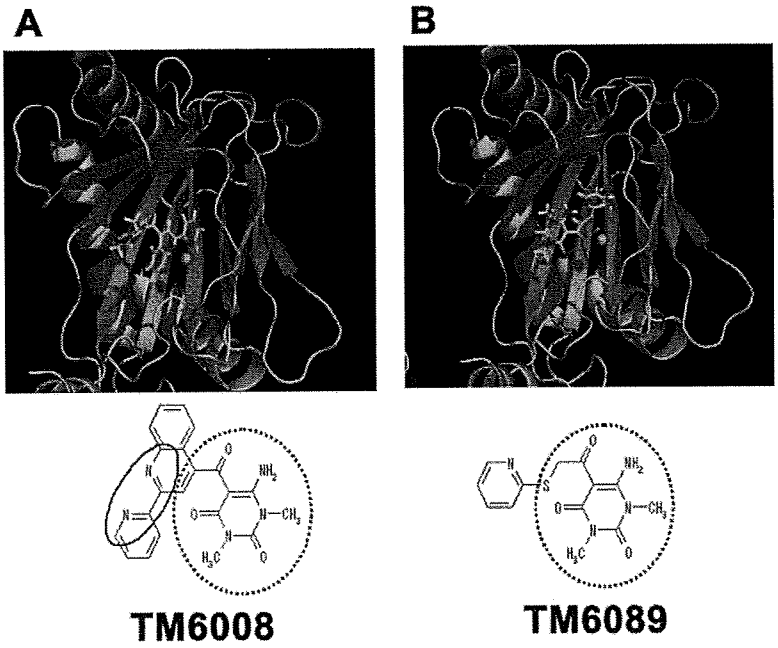

Figure 3. The predicted binding modes of TM6008 (A) and TM6089 (B) in PHD2. TM6008 and TM6089 are drawn by stick models. Sulfur, oxygen, nitrogen, carbon, and hydrogen atoms are shown in orange, red, blue, green, and white, respectively. Fe(II) is shown by an orange sphere. Figures were drawn by the software PyMOL version 0.97 (DeLano Scientific LLC).

metal (copper) and inhibited the autoxidation of ascorbic acid in a dose-dependent manner (IC 50 values were 330, 31.4, and 0.57 μmol/L, respectively). By contrast, TM6089 did not chelate transition metal even at the concentration of 100 μmol/L.

Binding Mode to Human PHD

PHD produces trans-4-hydroxyproline from 2-OG and L-proline (Pro) in the presence of Fe(II). The crystal structure of the catalytic domain of human PHD2, an important prolyl-4-hydroxylase in the human hypoxia response in normal cells, has been recently reported.[30] Based on the 3-dimensional structure of this PHD, we undertook docking simulations between our 2 PHD inhibitors and human PHD2. The docking modes of TM6008 and TM6089 are shown in Figure 3. TM6008 binds to the active site of PHD2 by chelating 2 nitrogen atoms with the iron atom. By contrast, TM6089 binds to the active site by nonchelating mechanism. The sulfur and 1 carbonyl oxygen atom of TM6089 point to the iron atom. The disposition of these 3 atoms, however, is unfavorable to form coordinate bonds. The binding mode of TM6089 demonstrates that TM6089 is a unique inhibitor without iron chelating affinity.

Toxicity and Pharmacokinetics

TM6008 and TM6089 did not exhibit cytotoxicity at the tested concentrations (up to 100 μmol/L). No acute toxicity was observed in mice up to 2 weeks after a single oral dose of 2000 mg/kg for TM6008, whereas the 50% lethal dose of TM6089 was 500 mg/kg. Pharmacokinetics studies in rats given an oral dose of 50 mg/kg of each compound disclosed plasma Tmax, Cmax, and T1/2 values of 3.5 hour, 0.9 μg/mL and 1.5 hour for TM6008, and 1.0 hour, 0.5 μg/mL, and 0.6 hour for TM6089.

Demonstration of the In Vivo Effectiveness

As VEGF is regulated by the HIF-HRE system, we examined whether TM6008 and TM6089 stimulate angiogenesis.

Firstly, we examined whether local injection of our compounds stimulates angiogenesis in vivo. For this purpose, we introduced small sponges under the skin of mice and measured their hemoglobin contents and vessel numbers after 10 days to estimate the stimulation of the HIF-HRE system. Injection of TM6008 increased angiogenesis as demonstrated by an increase of the hemoglobin content, and by an increased vessel number on immunohistochemical evaluation of the sponges. TM6089 also enhanced angiogenesis in the sponge assays (Figure 4A through 4C).

To investigate whether systemic administration of TM6008 and TM6089 stimulates in vivo the HIF-HRE system in various organs, we used the hypoxia-sensing transgenic rats. In the kidney expression of the reporter gene was not detected under basal conditions (amplification of 40 cycles), but expression of the reporter gene was obviously induced after a single oral dose 100 mg/kg of TM6008 and TM6089 (detected at 31.0±0.85 cycles and 31.0±2.05 cycles, respectively). In the liver, expression of the reporter gene, which was undetectable under basal conditions, was also induced after TM6008 administration (detected at 32.3±0.35 cycles), whereas TM6089 was ineffective. In the heart, the reporter gene was detected under basal conditions and both TM6008 and TM6089 increased its expression (1.37±1.00 and 6.69±5.45 fold increase, respectively). No attempt was made to evaluate the expression of the transgene in the brain because the pharmacokinetics studies showed that neither of the tested compounds crossed the blood-brain barrier.

Next, we evaluated capillary network formation by Matrigel assays. When endothelial cells were seeded onto Matrigel at subconfluent density, they developed tube-like structures at 9 hours. Quantification of capillary network formation by measuring the tube length revealed promotion of capillary network formation by TM6008, confirming the results of the sponge assays (Figure 4D).

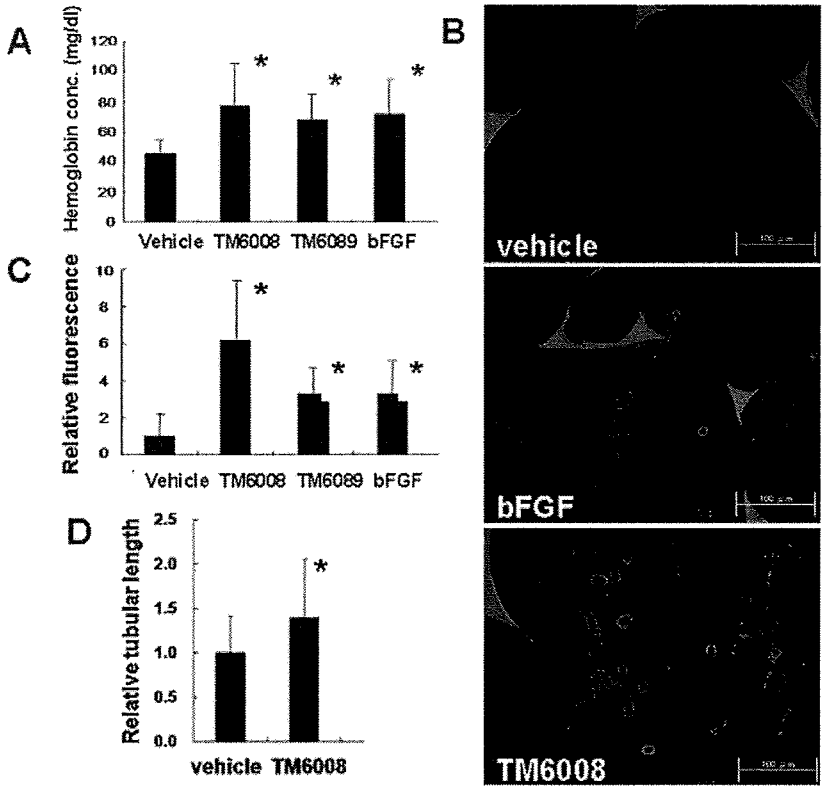

Figure 4. Stimulation of angiogenesis in the mouse sponge model and in the Matrigel assay. To assess the degree of angiogenesis, we measured the hemoglobin content in the sponge (A) and stained for the endothelial cell marker CD31 (B and C). B, Representative immunostaining of CD31 (×200); C, The average number of vessels. D, Capillary network formation on the Matrigel. *$P<0.05$ vs vehicle.

Prevention of Neuronal Cell Death Induced by Hypoxia

PHD inhibitors might protect cells against hypoxic damage. To test this hypothesis we used the delayed neuronal death model in gerbil. Nontoxic TM6008 (100 mg/kg/d) was given orally for 7 days in gerbils after a 5-minute transient global cerebral ischemia. The pathological outcome of neuronal cells was examined after 7 day administration of TM6008 in CA1 hippocampus with light microscopy.

In contrast with nonischemic gerbils (Figure 5A), gerbils subjected to ischemia and given vehicle alone (Figure 5B) exhibited in most pyramidal neurons ischemic cell damage, characterized by shrunken, darkly stained cytoplasm, and pyknotic nuclei with accumulation of glial cells. In the TM6008-treated animals, only a few neurons showed ischemic changes (Figure 5C). The number of viable neurons in the CA1 hippocampus, was higher in the TM6008-treated animals than in the vehicle-treated gerbils (166±73 versus

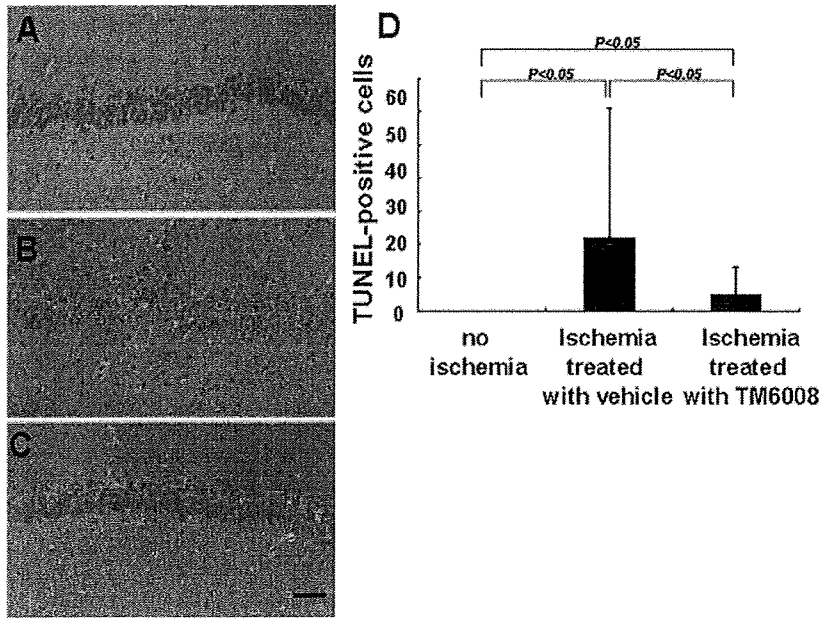

Figure 5. Prevention of hypoxic-induced neuronal cell death. HE staining in CA1 in a nonischemia animal (A), an animal subjected to ischemia and treated with vehicle alone (B), and an animal subjected to ischemia and treated with TM6008 (C). Scale bar=0.1 mm. D, TUNEL-positive cells in CA1.

61±55, $P<0.05$). The number of viable neurons in the CA1 hippocampus of the TM6008-treated animals was not statistically different from that observed in the nonischemia control group (227±50). Further, treatment with TM6008 decreased the number of apoptotic cells (Figure 5D). The final plasma concentration of TM6008 in these experiments was 7.8±2.9 μg/mL. Thus, TM6008 clearly protected against hypoxia-induced apoptotic neuronal death.

Next, we examined whether the protective effect of TM6008 against delayed neuronal death was attributable to enhanced angiogenesis. There was no statistically significant difference of the number of VEGF-positive cells between TM6008- and vehicle-treated groups (10.17±5.02 versus 9.12±1.55, respectively). Further, there was no significant difference of the value of cortical microperfusion at 7 days after occlusion between TM6008- and vehicle-treated groups (25.0±9.3 versus 29.4±8.7, respectively).

To clarify a neuroprotective mechanism of TM6008 in global ischemia models, we immunohistochemically stained with EPO, GLUT-1, and GLUT-3. The number of GLUT-3–positive cells in the CA1 hippocampus was significantly higher in TM6008 treated than in the vehicle-treated gerbils (26.9±7.5 versus 15.3±8.6, $P<0.05$). However, there was no statistically significant difference in EPO or GLUT-1–positive cells in the CA1 hippocampus between TM6008- and vehicle-treated groups (2.9±2.9 versus 3.3±1.7; and 5.5±2.1 versus 6.4±1.8, respectively).

Discussion

We identified novel molecules able to inhibit PHD activity and stabilize HIF. Our docking simulation studies based on the 3 dimensional structure of PHD2 have disclosed the molecular events required to inhibit PHD and therefore stabilize HIF. The target of these PHD inhibitors is the PHD active site.

Most of the PHD inhibitors reported so far, eg, 3,4-DHB, S956711 and FG-0041, are believed to inhibit the enzyme by iron chelating mechanism.[27,29] Iron chelating compounds could have nonspecific binding affinity to the iron containing proteins or iron ions and may not be desirable from the therapeutic point of view because iron is an essential cofactor for a host of important cellular functions, including oxidative phosphorylation and arachidonic acid signaling. To our surprise, the docking simulations demonstrated that TM6089 could preferentially bind to the active site of PHD2 without chelating to the iron atom. Indeed, TM6089 is devoid of iron chelating activity in vitro. Thus, iron chelation is not a necessary intermediate of PHD inhibition. According to our knowledge, TM6089 is the first unique PHD inhibitor which stimulates HIF activity without iron chelation.

The in vivo relevance of our novel PHD inhibitors was first demonstrated by the sponge assay in mice. Previous reports have shown that the hemoglobin contents of the sponge implants and the surrounding granuloma tissue correlated with the degree of angiogenesis.[31] Accordingly, both 3,4-DHB and S956711 were shown to raise the number of vessels in the sponge. In this study, we demonstrated not only an augmented number of vessels by immunohistochemistry but also an increased hemoglobin content in the sponge after local administration of TM6008 and TM6089.

Of great interest, these effects of TM6008 and TM6089 are not restricted locally but extend to several organs. To reach this conclusion, we used a hypoxia-sensing transgenic rat expressing a hypoxia-responsive reporter vector using a HRE of the 5′ VEGF untranslated region.[25] These transgenic rats have the unique asset to allow a sensitive and specific evaluation of HIF stimulation. As a consequence of systemic administration of TM6008 and TM6089 to these rats, expression of the reporter gene was considerably upregulated in the kidney, liver, and heart.

To extend these findings, we used less toxic TM6008 and obtained therapeutically relevant results in studies using gerbils. In gerbils, transient brain ischemia followed by reperfusion results in neuronal death in selectively vulnerable brain regions such as the hippocampal CA1 sector and caudate-putamen. The discovery that, in this model, TM6008 rescued neurons from apoptotic cell death in the CA1 hippocampus is noteworthy. Whereas TM6008 did not cross the blood-brain barrier, TM6008 protected the brain in a model of global cerebral ischemia. This is likely attributable to an increase in permeability of the blood-brain barrier, as previous reports showed that ischemic injury in this model destroys the blood-brain barrier and allows passage of compounds which do not penetrate the barrier under normal conditions.[32]

Mechanisms of neuroprotection by TM6008 can theoretically be multifactorial because HIF regulates a wide range of protective genes such as those involved in erythropoiesis (EPO, transferring, and hepcidin), angiogensis (VEGF), antioxidative stress (HO-1), glycolysis (Glut-1, Glut-3, and aldolase A), and so on. Angiogenic effects of TM6008 shown by the Matrigel assays and sponge assays stimulated us to study whether enhanced angiogenesis played a role in neuronal protection in our model. However, we could not find enhanced angiogenesis in the brain of gerbils treated with TM6008 by counting VEGF-positive vessels or measuring blood flow by laser Doppler flowmetry. Therefore, it is unlikely that the protective effect of TM6008 was related to angiogenesis in the gerbil forebrain ischemia model. This may be explained by different concentrations of TM6008 among the assays. Although we could not measure the local concentrations of TM6008 in the damaged brain, it is likely that the concentration of the gerbil forebrain treated with TM6008 p.o. is lower than those obtained by local administration such as sponge assays and Matrigel assay.

We next focused on effects of TM6008 on neuronal apoptosis. Our terminal deoxynucleotidyl transferase-mediated dUTP nick end-labeling (TUNEL) assays demonstrated that TM6008 decreased the number of apoptotic cells in the brain, and other potential neuroprotective mechanisms by TM6008 include antiapoptotic effects mediated by other HIF-regulated genes such as EPO,[33] VEGF,[34] and glucose transporters.[35] EPO is a pleiotropic cytokine[36] and induces neuroprotection via the antiapoptotic signaling cascades like Bcl-X_L through direct binding to the Bcl-X promoter.[37] Antiapoptotic effects of VEGF contribute to reduction of ischemic brain damage in addition to its angiogenic effects.[38]

The glucose transporter GLUT-1 is also positively regulated through HIF-1α, and the microinfusion of virus vectors bearing the GLUT-1 isoform into the brain tissue reduced seizure-induced[39] and ischemic neuronal damage in vivo.[40] However, our immunohistochemical analysis could not demonstrate upregulation of these genes. In contrast, we observed upregulation of Glut-3. Glut-3 is also regulated by HIF,[41] and recent studies suggested a critical role of Glut-3 in protecting against a decline in brain glucose uptake under ischemic conditions.[42]

These results fit with the observations collected during various therapeutic strategies related to HIF target genes. For instance, cobalt chloride has been used as a conventional HIF stabilizer. It is generally believed to replace the iron present in PHD, but recent studies demonstrated that cobalt also depletes intracellular ascorbate,[28] a substrate of PHD. Cobalt is effective in a variety of hypoxia-related disorders including cerebrovascular disease.[23,43,44] In addition to PHD, there are other factors regulating the HIF stability/activity. Factor-inhibiting-HIF (FIH) hydroxylates regulates HIF activation via controlling CBP/p300 recruitment. The phosphoinositide 3-kinase (PI3K)/Akt pathway and the protein kinase C signaling have also been implicated in the regulation of HIF-α. Whether these pathways can be a good target for therapeutic approaches is a future subject to be pursed.

The protective effect of TM6008 against ischemia-induced cerebral lesions suggested its potential usefulness in other ischemic disorders such as cardiac or kidney diseases. It should not be forgotten that HIF stimulation acts as a general switch for several proteins such as VEGF, erythropoietin, etc. Although these proteins are protective under hypoxic conditions, recent demonstration that both erythropoietin and VEGF accelerates diabetic retinopathy independently[45] should call for caution. Their administration during several months warrants long-term experimental studies before concluding to its safety. On the other hand, the short-term use of PHD inhibitors for acute hypoxic damage will probably prove safe.

Sources of Funding

This study was supported by grants from the Program for Promotion of Fundamental Studies in Health Sciences of the Pharmaceuticals and Medical Devices Agency (PMDA) and from the Japan Society for the Promotion of Science for Scientific Research.

Disclosures

None.

References

1. Marx J. How cells endure low oxygen. *Science*. 2004;303:1454–1456.
2. Semenza GL, Wang GL. A nuclear factor induced by hypoxia via de novo protein synthesis binds to the human erythropoietin gene enhancer at a site required for transcriptional activation. *Mol Cell Biol*. 1992;12:5447–5454.
3. Wang GL, Semenza GL. General involvement of hypoxia-inducible factor 1 in transcriptional response to hypoxia. *Proc Natl Acad Sci U S A*. 1993;90:4304–4308.
4. Epstein AC, Gleadle JM, McNeill LA, Hewitson KS, O'Rourke J, Mole DR, Mukherji M, Metzen E, Wilson MI, Dhanda A, Tian YM, Masson N, Hamilton DL, Jaakkola P, Barstead R, Hodgkin J, Maxwell PH, Pugh CW, Schofield CJ, Ratcliffe PJ. C. elegans EGL-9 and mammalian homologs define a family of dioxygenases that regulate HIF by prolyl hydroxylation. *Cell*. 2001;107:43–54.
5. Schofield CJ, Ratcliffe PJ. Oxygen sensing by HIF hydroxylases. *Nat Rev Mol Cell Biol*. 2004;5:343–354.
6. Semenza GL. Hydroxylation of HIF-1: oxygen sensing at the molecular level. *Physiology (Bethesda)*. 2004;19:176–182.
7. Masson N, Ratcliffe PJ. HIF prolyl and asparaginyl hydroxylases in the biological response to intracellular O(2) levels. *J Cell Sci*. 2003;116:3041–3049.
8. Bruick RK, McKnight SL. A conserved family of prolyl-4-hydroxylases that modify HIF. *Science*. 2001;294:1337–1340.
9. Semenza GL. HIF-1, O(2), and the 3 PHDs: how animal cells signal hypoxia to the nucleus. *Cell*. 2001;107:1–3.
10. Hon WC, Wilson MI, Harlos K, Claridge TDW, Schofield CJ, Pugh CW, Maxwell PH, Ratcliffe PJ, Stuart DI, Jones EY. Structural basis for the recognition of hydroxyproline in HIF-1 alpha by pVHL. *Nature*. 2002;417:975–978.
11. Ivan M, Kondo K, Yang H, Kim W, Valiando J, Ohh M, Salic A, Asara JM, Lane WS, Kaelin WG. HIFa targeted for VHL-mediated destruction by proline hydroxylation: implications for O_2 sensing. *Science*. 2001;292:464–468.
12. Maxwell PH, Wiesener MS, Chang GW, Clifford SC, Vaux EC, Cockman ME, Wykoff CC, Pugh CW, Maher ER, Ratcliffe RJ. The tumour suppressor protein VHL targets hypoxia-inducible factors for oxygen-dependent proteolysis. *Nature*. 1999;399:271–275.
13. Ohh M, Park CW, Ivan M, Hoffman MA, Kim TY, Huang LE, Pavletich N, Chau V, Kaelin WG. Ubiquitination of hypoxia-inducible factor requires direct binding to the beta-domain of the von Hippel-Lindau protein. *Nat Cell Biol*. 2000;2:423–427.
14. Pugh CW, Ratcliffe PJ. Regulation of angiogenesis by hypoxia: role of the HIF system. *Nat Med*. 2003;9:677–684.
15. Giaccia A, Siim BG, Johnson RS. HIF-1 as a target for drug development. *Nat Rev Drug Discov*. 2003;2:803–811.
16. Hewitson KS, Schofield CJ. The HIF pathway as a therapeutic target. *Drug Discov Today*. 2004;9:704–711.
17. Bernstein FC, Koetzle TF, Williams GJ, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. The Protein Data Bank: a computer-based archival file for macromolecular structures. *J Mol Biol*. 1977;112:535–542.
18. Halgren TA. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J Comp Chem*. 1996;17:490–519.
19. Edelsbrunner H, Facello M, Fu R, Liang J. Measuring proteins and voids in proteins. *Proceedings of the 28th Hawaii International Conference on Systems Science*. 1995; 256–264.
20. Goto J, Kataoka R, Hirayama N. Ph4Dock-Pharmacophore-based protein-ligand docking. *J Med Chem*. 2004;47:6804–6811.
21. Kaule G, Gunzler V. Assay for 2-oxoglutarate decarboxylating enzymes based on the determination of [1-14C] succinate: Application to prolyl 4-hydroxylase. *Anal Biochem*. 1990;184:291–297.
22. Price DL, Rhett PM, Thorpe SR, Baynes JW. Chelating activity of advanced glycation end-product (AGE) inhibitors. *J Biol Chem*. 2001;276:48967–48972.
23. Tanaka T, Kojima I, Ohse T, Ingelfinger JR, Adler S, Fujita T, Nangaku M. Cobalt promotes angiogenesis via hypoxia-inducible factors and protects ischemic tubulointerstitium in the remnant kidney. *Lab Invest*. 2005;85:1292–1307.
24. Muramatsu M, Katada J, Hayashi I, Majima M. Chymase as a proangiogenic factor. A possible involvement of chymase-angiotensin-dependent pathway in the hamster sponge angiogenesis model. *J Biol Chem*. 2000;275:5545–5552.
25. Tanaka T, Miyata T, Inagi R, Fujita T, Nangaku M. Hypoxia in renal disease with proteinuria and/or glomerular hypertension. *Am J Pathol*. 2004;165:1979–1992.
26. Kirino T. Delayed neuronal death in the gerbil hippocampus following ischemia. *Brain Res*. 1982;239:57–69.
27. Ivan M, Haberberger T, Gervasi DC, Michelson KS, Gunzler V, Kondo K, Yang H, Sorokina I, Conaway RC, Conaway JW, Kaelin WG Jr. Biochemical purification and pharmacological inhibition of a mammalian prolyl hydroxylase acting on hypoxia-inducible factor. *Proc Natl Acad Sci U S A*. 2002;99:13459–13464.
28. Salnikow K, Donald SP, Bruick RK, Zhitkovich A, Phang JM, Kasprzak KS. Depletion of intracellular ascorbate by the carcinogenic metals nickel and cobalt results in the induction of hypoxic stress. *J Biol Chem*. 2004;279:40337–40344.
29. Warnecke C, Griethe W, Weidemann A, Jurgensen JS, Willam C, Bachmann S, Ivashchenko Y, Wagner I, Frei U, Wiesener M, Eckardt

KU. Activation of the hypoxia-inducible factor-pathway and stimulation of angiogenesis by application of prolyl hydroxylase inhibitors. *FASEB J.* 2003;17:1186–1188.
30. McDonough MA, Li V, Flashman E, Chowdhury R, Mohr C, Lienard BM, Zondlo J, Oldham NJ, Clifton IJ, Lewis J, McNeill LA, Kurzeja RJ, Hewitson KS, Yang E, Jordan S, Syed RS, Schofield CJ. Cellular oxygen sensing: Crystal structure of hypoxia-inducible factor prolyl hydroxylase (PHD2). *Proc Natl Acad Sci U S A.* 2006;103:9814–9819.
31. Majima M, Isono M, Ikeda Y, Hayashi I, Hatanaka K, Harada Y, Katsumata O, Yamashina S, Katori M, Yamamoto S. Significant roles of inducible cyclooxygenase (COX)-2 in angiogenesis in rat sponge implants. *Jpn J Pharmacol.* 1997;75:105–114.
32. Picozzi P, Todd NV, Crockard HA. Regional blood-brain barrier permeability changes after restoration of blood flow in postischemic gerbil brains: a quantitative study. *J Cereb Blood Flow Metab.* 1985;5:10–16.
33. Zaman K, Ryu H, Hall D, O'Donovan K, Lin KI, Miller MP, Marquis JC, Baraban JM, Semenza GL, Ratan RR. Protection from oxidative stress–induced apoptosis in cortical neuronal cultures by iron chelators is associated with enhanced dna binding of hypoxia-inducible factor-1 and atf-1/creb and increased expression of glycolytic enzymes, $p21^{waf1/cip1}$, and erythropoietin. *J Neurosci.* 1999;15:9821–9830.
34. Hayashi T, Abe K, Itoyama Y. Reduction of ischemic damage by application of vascular endothelial growth factor in rat brain after transient ischemia. *J Cereb Blood Flow Metab.* 1998;18:887–895.
35. Vannucci SJ, Clark RR, Koehler-Stec E, Li K, Smith CB, Davies P, Maher F, Simpson IA. Glucose transporter expression in brain: Relationship to cerebral glucose utilization. *Dev Neurosci.* 1998;20:369–379.
36. Liu J, Narasimhan P, Yu F, Chan PH. Neuroprotection by hypoxic preconditioning involves oxidative stress-mediated expression of hypoxia-inducible factor and erythropoietin. *Stroke.* 2005;36:1264–1269.
37. Wen TC, Sadamoto Y, Tanaka J, Zhu PX, Nakata K, Ma YJ, Hata R, Sakanaka M. Erythropoietin protects neurons against chemical hypoxia and cerebral ischemic injury by up-regulating Bcl-xL expression. *J Neurosci Res.* 2002;67:795–803.
38. Sun FY, Guo X. Molecular and cellular mechanisms of neuroprotection by vascular endothelial growth factor. *J Neurosci Res.* 2005;79: 180–4.
39. McLaughlin J, Roozendaal B, Dumas T, Gupta A, Ajilore O, Hsieh J, Ho D, Lawrence M, McGaugh JL, Sapolsky R. Sparing of neuronal function postseizure with gene therapy. *Proc Natl Acad Sci.* 2000;97: 12804–12809.
40. Lawrence MS, Sun GH, Kunis DM, Saydam TC, Dash R, Ho DY, Sapolsky RM, Steinberg GK. Overexpression of the glucose transporter gene with a herpes simplex viral vector protects striatal neurons against stroke. *J Cereb Blood Flow Metab.* 1996;16:181–185.
41. O'Rourke JF, Pugh CW, Bartlett SM, Ratcliffe PJ. Identification of hypoxically inducible mRNAs in HeLa cells using differential-display PCR. Role of hypoxia-inducible factor-1. *Eur J Biochem.* 1996;241: 403–410.
42. Zovein A, Flowers-Ziegler J, Thamotharan S, Shin D, Sankar R, Nguyen K, Gambhir S, Devaskar SU. Postnatal hypoxic-ischemic brain injury alters mechanisms mediating neuronal glucose transport. *Am J Physiol Regul Integr Comp Physiol.* 2004;286:R273–R282.
43. Bergeron M, Gidday JM, Yu AY, Semenza GL, Ferriero DM, Sharp FR. Role of hypoxia-inducible factor-1 in hypoxia-induced ischemic tolerance in neonatal rat brain. *Ann Neurol.* 2000;48:285–296.
44. Matsumoto M, Makino Y, Tanaka T, Tanaka H, Ishizaka N, Noiri E, Fujita T, Nangaku M. Induction of renoprotective gene expression by cobalt ameliorates ischemic injury of the kidney in rats. *J Am Soc Nephrol.* 2003;14:1825–1832.
45. Watanabe D, Suzuma K, Matsui S, Kurimoto M, Kiryu J, Kita M, Suzuma I, Ohashi H, Ojima T, Murakami T, Kobayashi T, Masuda S, Nagao M, Yoshimura N, Takagi H. Erythropoietin as a retinal angiogenic factor in proliferative diabetic retinopathy. *N Engl J Med.* 2005;353: 782–792.

Use of Amino Acid Composition to Predict Ligand-Binding Sites

Shinji Soga,† Hiroki Shirai,† Masato Kobori,† and Noriaki Hirayama*,‡

Molecular Medicine Research Laboratories, Drug Discovery Research, Astellas Pharma Inc., 21 Miyukigaoka, Tsukuba, Ibaraki 305-8585, Japan, and Basic Medical Science and Molecular Medicine, Tokai University School of Medicine, Boseidai, Isehara, Kanagawa 259-1193, Japan

Received May 30, 2006

A novel method for predicting the binding sites for druglike compounds on the surface of proteins was developed on the basis of the specific amino acid composition observed at the ligand-binding sites of ligand−protein complexes determined by X-ray analysis. A profile representing the preference of each of the 20 standard amino acids at the binding sites of druglike molecules was obtained for a small set of high-quality complex structures. An index termed propensity for ligand binding (PLB) was created from these profiles. The PLB index was used to predict the propensity of binding for 804 ligands at all potential binding sites on the proteins whose structures were determined by X-ray analysis. If the sites with the first two highest PLB indices are taken into consideration, the successfully predicted sites reached a high percentage of 86. The PLB prediction is relatively simple, but the validation study showed that it is both fast and accurate to detect ligand-binding sites, especially the binding sites of druglike molecules. Therefore, the PLB index can be used to predict the ligand-binding sites of uncharacterized protein structures and also to identify novel drug-binding sites of known drug targets.

1. INTRODUCTION

The specific binding of a ligand to a therapeutic target molecule is the key to drug action. Each ligand binds preferentially to a specific site on the surface of the target molecule. The binding site is usually located in a concavity on the surface of the target protein whose configuration, in most cases, is highly specialized for the binding of a particular group of compounds.

Identification of the ligand-binding site for each specific protein molecule is crucially important when trying to find a suitable drug molecule for the target, but it is also important to understand the function of the protein.

Recently, the number of X-ray structures of ligand−protein complexes has increased markedly. A sufficient number of reliable ligand−protein complex structures are now available in the Protein Data Bank (PDB)[1] for use in the systematic study of the common characteristics of molecular interactions between ligands and proteins.

Since the binding site of a drug is considered to be highly specific to its molecular characteristics, the binding site must have a distinct character significantly different from those of similar concavities on the target molecule. Although each binding site is made up of specific amino acid residues from mostly noncontiguous regions of the protein, the site is expected to be richer in certain specific amino acid residues and poorer in others.

Several algorithms for predicting ligand-binding sites have been published over the past 10 years. They can be divided into three major categories: (i) geometric algorithms,[2] (ii) probe-mapping algorithms,[3] and (iii) physical potential algorithms.[4] Although these algorithms have each achieved a measure of success, none of them take the clustering of specific amino acids at the ligand-binding sites into account, which is an important effect that needs to be considered.

The compositions of the amino acids at the ligand-binding sites of the ligand−protein complexes whose structures were determined by X-ray crystallography were examined in this study. It is of particular interest to study the frequency of appearance at the binding sites for each of the 20 standard amino acids and check for specific amino acid compositions. Attempts to use this information for the prediction of ligand-binding sites were also made.

In this study, only ligand−protein complexes with drug or druglike compounds were considered. A detailed analysis of the amino acid compositions around the binding sites for these compounds revealed that there were clear propensities for the presence of specific amino acids at the binding site of each drug or druglike compound. A novel propensity for ligand binding (PLB) index was developed on the basis of the idiosyncratic amino acid profile of each ligand-binding site. The PLB index's ability to predict ligand-binding sites was tested on a systematically collected data set. Validation of this method indicated that it is both fast and accurate.

2. MATERIALS AND METHODS

In this study, the amino acid compositions of the binding sites of druglike compounds were analyzed using a training data set that consisted of the most accurate and diverse complex structures from the PDB. The PLB index was defined in terms of the specific profile of the amino acid composition observed at each binding site of the druglike compounds. The predictive power of the PLB index was then evaluated using a test data set consisting of a different set of accurate complex structures from the PDB (none of the

* Corresponding author tel.: +81 463 93 1121; e-mail: hirayama@is.icc.u-tokai.ac.jp.
† Astellas Pharma Inc.
‡ Tokai University School of Medicine.

10.1021/ci6002202 CCC: $37.00
Published on Web 01/23/2007

ones used in the training data set were included). The two data sets were compiled using the PDB data downloaded on July 20, 2005.

2.1. Identification of Concavities. Small organic molecules, especially drug molecules, typically bind at concavities on the surface of the protein. In this study, a program named Alpha Site Finder[5] implemented in the software system MOE[6] was used to detect concavities on the surface of proteins. Concavities identified by Alpha Site Finder are characterized by a cluster of small spheres called "α-spheres." The positions and characteristics of the spheres are calculated on the basis of the geometry and character of the protein surface. The cluster of α-spheres represents the shape and size of the concavity. Since Alpha Site Finder can map every depression on the protein surface, it is ideal for the purposes of this study. Alpha Site Finder usually identifies multiple concavities on each protein molecule; thus, the goal is to identify which of these is the most probable drug binding site.

2.2. Training Data Set. Since the training data set must represent typical drug−protein complexes, the candidates were carefully selected using the criteria described below.

2.2.1. High-Quality X-ray Structures. The amino acids at the binding sites are identified on the basis of the coordinates of ligand atoms other than hydrogen. Reliable positions of non-hydrogen atoms are essential for this study. For the training data set, in particular, the highest-quality X-ray structures were selected in the interests of accuracy.

In order to obtain the crystal structures in which all non-hydrogen atoms were unambiguously determined, structures were selected using the following criteria: a R_{free} value of less than 0.24, a resolution value of less than or equal to 2.5 Å, occupancy factors of 1.0 for all non-hydrogen atoms, and atomic displacement parameters of less than 30 Å^2 for all non-hydrogen atoms. If a protein was multimeric, only a monomer with the smallest atomic displacement parameters for their non-hydrogen atoms was considered.

2.2.2. Complexes with Druglike Ligands. The drug-likeness profile of a molecule, which comprises multiple molecular descriptors, is useful for determining how much like a drug the molecule is.[7] The ranges of values of the 14 descriptors given in Table 1 cover 85% of the drugs used clinically in Japan now. These values were used to eliminate nondruglike ligands. When 12 of the descriptor values for a particular ligand fell within these ranges, the ligand was considered to be a druglike molecule, and the corresponding complexes were taken into consideration.

There are many ligands with multiple phosphorus atoms in the PDB. However, from the drug-likeness viewpoint, they are not suitable. Therefore, all ligands containing more than one phosphorus atom per molecule were eliminated.

2.2.3. Nonredundant Structures. When a complex contained multiple identical ligands, only the one with the smallest average atomic displacement factor was considered. If a ligand formed complexes with multiple homologous proteins, only the complex structure with the smallest R_{free} was considered. The proteins in the training data set are considered to be sufficiently diverse since their maximum percent identity is 48%.

These selection criteria resulted in a training data set that includes 41 complex structures. The chemical structures of the ligands contained in the data set are listed in Figure 1, together with the PDB codes. As Figure 1 shows, the structures of ligands are chemically diverse.

Table 1. Value Distributions for 14 Molecular Descriptors That Apply to 85% of the Drugs Used Clinically in Japan[a]

descriptor	ranges	
weight	165	555
SlogP	−1.18	5.30
SMR	4.34	14.46
TPSA	13.0	165
density	0.73	0.99
vdw_area	165	497
vdw_vol	181	623
a_acc	1	7
a_don	0	6
a_hyd	6	26
KierA1	7.82	26.3
KierA2	3.13	11.8
KierA3	1.48	7.32
KierFlex	1.68	8.82

[a] weight: molecular weight; vdw_area: area of van der Waals surface calculated using a connection table approximation; vdw_vol: van der Waals volume calculated using a connection table approximation; density: molecular mass density (molecular weight divided by vdw_vol); a_acc: number of hydrogen-bond acceptor atoms (not counting acidic atoms but counting atoms that are both hydrogen-bond donors and acceptors such as −OH); a_don: number of hydrogen-bond donor atoms (not counting basic atoms but counting atoms that are both hydrogen-bond donors and acceptors such as −OH); a_hyd: number of hydrophobic atoms; SlogP:[9] calculated hydrophobicity by Crippen; SMR:[9] calculated molar refractivity by Crippen; TPSA:[10] topological polar surface area; KierA1, KierA2, KierA3, and KierFlex:[11] molecular connectivity indices.

2.3. Test Data Set. 2.3.1. Data Selection. A test data set of the complex structures was constructed in order to evaluate the predictive power of the PLB index. The proteins that are homologous to ones in the training data set were not included in the test data set. Sequence similarity was judged by a Basic Local Alignment Search Tool[8] search of which E-value threshold was below 1.0. The minimum value of percent identity between similar sequences was 48. By relaxing four of the training data set selection conditions, the proteins for the test data set were further narrowed down. First, the occupancy factors of the non-hydrogen atoms in a protein could be less than 1.0. Second, the atomic displacement factors of the non-hydrogen atoms in a ligand could be greater than 30 Å^2. Third, only the range of molecular weight was used to select molecules. Finally, ligands were allowed to contain multiple phosphorus atoms. If a protein was multimeric, only a monomer with the largest occupancy factors and the smallest atomic displacement parameters for their non-hydrogen atoms was considered. The test data set consisted of 756 complex structures with 804 ligands. Although the selection conditions were relaxed, the X-ray structures were still high-quality and many druglike molecules were included, making the test data set suitable for evaluating the PLB index.

2.3.2. Calculation of Concavities. All concavities in the protein structures were identified and catalogued. Alpha Site Finder found 15 892 concavities in 756 protein structures. Although the number of binding sites was 778, 804 of the ligands were bound. This means that multiple ligands were bound in some proteins. Since it was likely that some of the concavities were too small to accommodate ligands, they were not considered as possible binding sites. The number

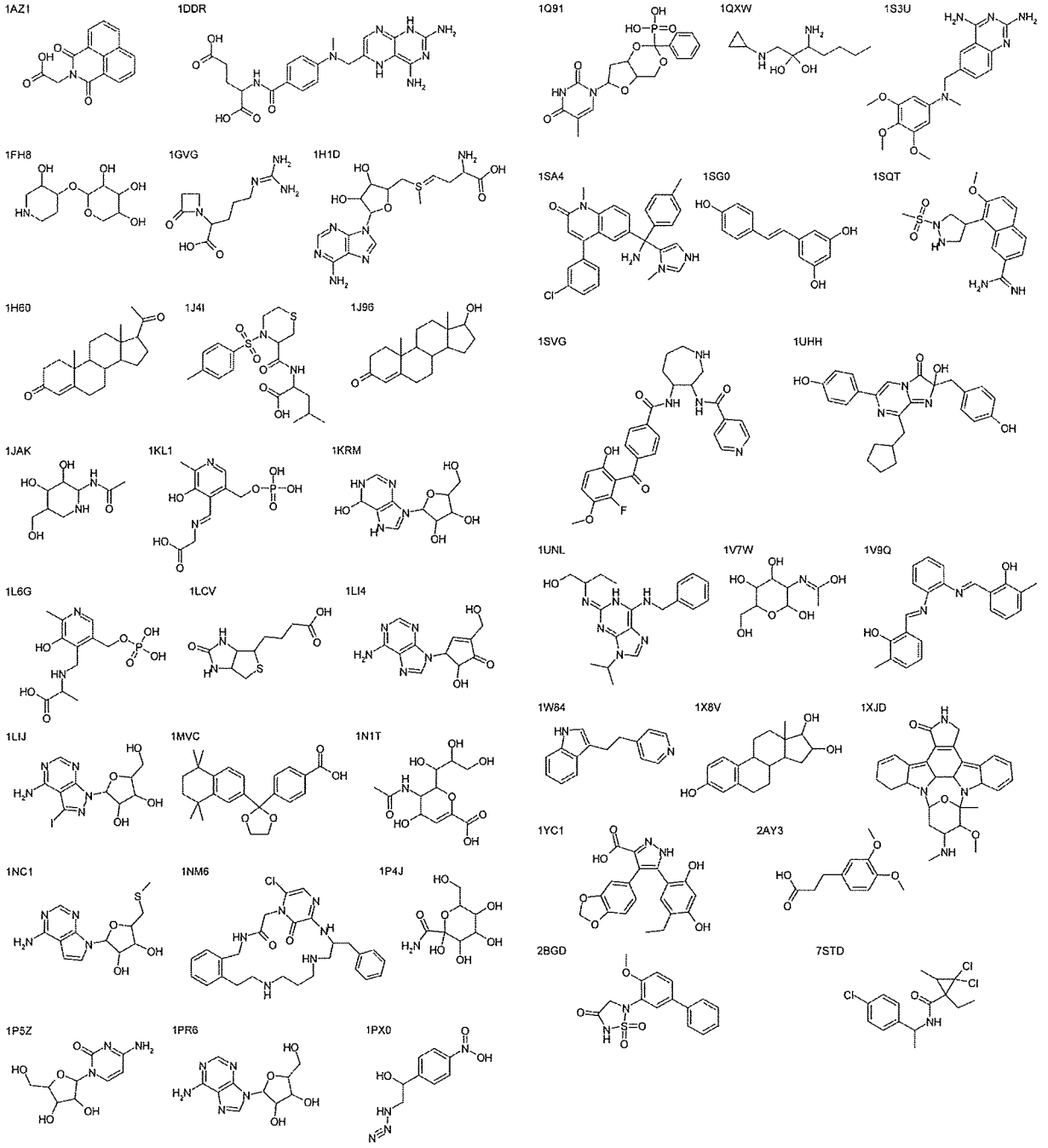

Figure 1. Chemical structures of the ligands in the training data set together with their PDB codes.

of α-spheres per concavity was used to judge size. Those concavities smaller than or equal to the minimum size of the smallest concavity in the training data set were considered trivial and eliminated. The number of concavities of realistic size for ligand binding was determined to be 15 232.

2.4. Specific Amino Acid Composition at Ligand-Binding Site. The amino acids surrounding each ligand in the training data set were examined. If any non-hydrogen atom in an amino acid was within 4.5 Å of any non-hydrogen atom of a ligand, it was expected that the amino acid would interact with the ligand. Therefore, 4.5 Å was used as the cutoff value to determine the amino acid composition of the binding site. MOE was used to calculate this distance. If the number of amino acids of type x at a binding site is represented as N_x, the composition of the amino acid of type x at the binding site, CA(x), is defined as

$$\mathrm{CA}(x) = N_x / \sum_{y=1}^{20} N_y \qquad (1)$$

Here, N_y denotes the number of amino acids of type y at the binding site. The denominator in the above equation is the total number of amino acid residues at the binding site.

Table 2. Amino Acid Compositions Normalized Using the 20 Standard Amino Acids[a]

x	CA(x)	SA(x)	RA(x)
A	0.050	0.072	0.701
C	0.021	0.013	1.650
D	0.065	0.064	1.015
E	0.061	0.064	0.956
F	0.080	0.041	1.952
G	0.058	0.073	0.788
H	0.056	0.025	2.286
I	0.050	0.050	1.006
K	0.028	0.060	0.468
L	0.087	0.084	1.045
M	0.037	0.020	1.894
N	0.041	0.051	0.811
P	0.010	0.049	0.212
Q	0.027	0.040	0.669
R	0.047	0.052	0.916
S	0.056	0.064	0.883
T	0.041	0.057	0.730
V	0.056	0.064	0.884
W	0.058	0.019	3.084
Y	0.068	0.041	1.672

[a] CA(x) denotes the composition of amino acid of type x at the ligand-biding sites of the proteins in the training data set. SA(x) denotes the composition of amino acid of type x on the surface of the proteins in the test data set. RA(x) denotes the ratio of CA(x) to SA(x).

The incidence of amino acid residues on the protein surface was also investigated. The amino acids on the surface of the protein were identified by calculating the solvent-accessible surface of each amino acid using a probe sphere with a radius of 1.4 Å. The solvent-accessible surface was calculated using MOE. If the probe came into contact with a non-hydrogen atom in a residue, the residue was regarded as being on the surface of the protein. If the number of amino acids of type x on the surface of a protein is N_{sx}, the rate of occurrence of amino acids of type x on the surface of the protein, SA(x), is defined as

$$\mathrm{SA}(x) = N_{sx} / \sum_{y=1}^{20} N_{sy} \tag{2}$$

Here, N_{sy} denotes the number of amino acids of type y on the surface of the proteins. The denominator is the total number of all amino acids on the surface of the protein. SA(x) was determined using all protein structures in the test data set. The normalized values of CA and SA using the 20 standard amino acids are given in Table 2.

The ratio of CA(x) to SA(x), designated as RA(x),

$$\mathrm{RA}(x) = \mathrm{CA}(x)/\mathrm{SA}(x) \tag{3}$$

is the rate of occurrence of an amino acid of type x at the ligand-binding site. RA(x) is also considered to be the preference factor for an amino acid of type x at the binding site. By using a linear combination of the RAs for all 20 standard amino acids, an index of PLB can be defined with respect to a particular concavity, i, as follows:

$$\mathrm{PLB}_i = \sum_{x=1}^{20} N_{i,x}\mathrm{RA}(x) \tag{4}$$

The weighting-factor, $N_{i,x}$, denotes the number of amino acids of type x found in concavity i.

In order to distinguish the binding site with the most potential from the other concavities on a protein, the PLB values should be normalized by all concavities in the protein. Z-scored PLBs are used for this purpose. If there are M concavities in a protein, the Z-scored PLB for concavity i is calculated as follows:

$$Z_{\mathrm{PLB}_i} = \frac{\mathrm{PLB}_i - \mu}{\sigma} \tag{5}$$

Here,

$$\mu = \frac{\sum_{i=1}^{M} \mathrm{PLB}_i}{M} \tag{6}$$

and

$$\sigma = \sqrt{\frac{\sum_{i=1}^{M} (\mathrm{PLB}_i - \mu)^2}{M}} \tag{7}$$

Hereafter, Z_{PLB_i} is designated simply as PLB. By use of the PLB index, it is possible to judge the probability of ligand binding for a given concavity. The concavity with the highest PLB index in a protein is the most probable site for ligand binding. Accordingly, a smaller PLB index implies that ligand binding is less probable.

3. RESULTS AND DISCUSSION

3.1. Frequency of Amino Acid Presence at the Ligand-Binding Sites and on the Surface of Proteins. The CAs and SAs for the 20 standard amino acids were calculated for all proteins in the training and test data sets, respectively. The results are shown in Figure 2.

The CA values for many of the amino acids are markedly different from their corresponding SA values. The RA values, shown in Figure 3, illustrate these differences clearly.

The amino acids are sorted in ascending order of RA in this figure. The effect is striking, and the implications are particularly intriguing. Since the frequency rates of the aromatic residues and Met at the ligand-binding sites were high, these residues can be considered binding-site-philic residues. By the same token, Pro, Lys, Gln, and Ala can be considered binding-site-phobic since they were not often found. These profiles clearly show that the amino acid composition at each ligand-binding site is highly specific. It follows, then, that these characteristic profiles could be used to predict ligand-binding site locations on the basis of the amino acid compositions around the concavities.

3.2. Prediction of Ligand-Binding Sites Using the PLB Index. The PLB indices were calculated for 15 232 concavities found in 756 proteins. Since most proteins have multiple concavities, the PLB index would be an ideal tool if it narrows down the options. The concavity with the highest PLB index can be considered to be the most probable ligand-binding site. There were 611 cases where the concavity with the highest PLB index corresponded to the true binding site, which is 79% of the true ligand-binding sites in the test data set. If the sites with the first two highest PLB indices are

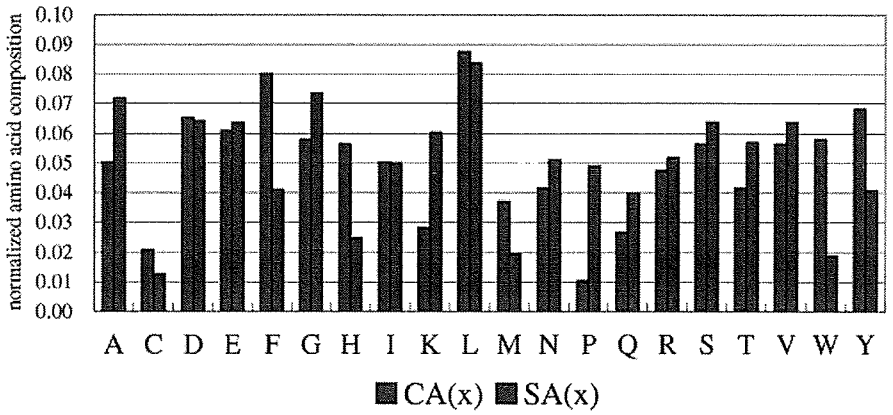

Figure 2. Normalized composition of the 20 standard amino acids. Blue: composition at the ligand-binding sites of the proteins in the training data set (CA). Red: composition on the surface of the proteins in the test data set (SA).

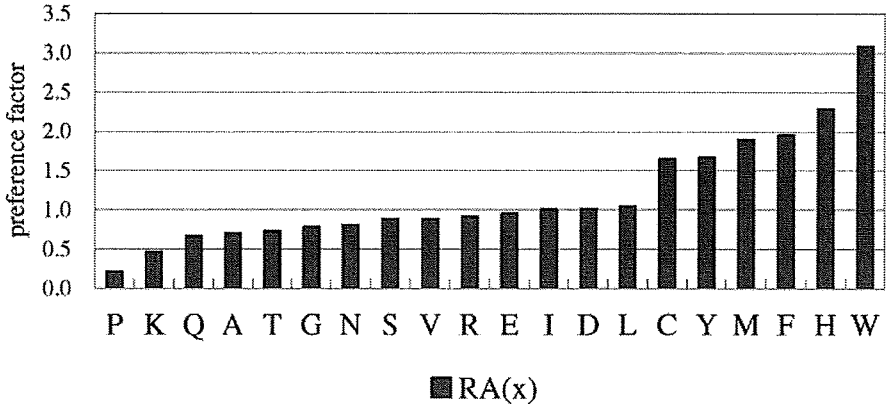

Figure 3. Preference factors for the 20 standard amino acids (RA = CA/SA).

taken into consideration, the successfully predicted sites reached to the higher percentage of 86. From a practical point of view, the latter selection is reasonable.

There were 1255 concavities with a PLB index greater than or equal to 1.2. Of these, 649 were true concavities. Hence, the PLB index prediction enrichment rate was calculated to be 649/1255/(778/15 232) = 10.12. This indicates that the PLB index is useful for the identification of ligand-binding sites.

Many nondruglike ligands were included in the test data set. Since the PLB index was derived from a training data set that contained many drug and druglike molecules, it is of great interest to see how well the PLB index predicted the binding sites of druglike compounds in the test data set. Each molecule's degree of "druglikeness" was judged using the druglikeness profile mentioned above. The number of true ligand-binding sites where druglike compounds bind is 126. The highest PLB index predicted 110 binding sites, which is 87% of the true binding sites. If the concavities with the top two PLB indices are taken into consideration, 120 binding sites can be identified. It covers as much as 95% of the true binding sites.

The results clearly show that the PLB index is useful for detecting not only small-molecule binding sites in general, but also binding sites specific to druglike molecules.

3.3. Two Typical Examples of Prediction. The following two examples illustrate how the PLB index is a more effective way to distinguish the true binding site from other concavities in a protein.

A total of 17 concavities were detected in the protein structure of a protein−ligand complex of carbonic anhydrase II (PDB code: 1OKL),[12] the eight largest of which are shown in Figure 4a. The concavities are represented by a cluster of α-spheres. Each α-sphere is classified as either hydrophilic or hydrophobic (red or white, respectively), depending on whether it is in a location conducive to hydrogen bonding or not.

The volume of a concavity can be expressed using several indices, such as the number of the α-spheres in the concavity, the number of protein atoms in contact with the α-spheres, and the number of amino acids in contact with the α-spheres. These indices were calculated for the 17 concavities and are listed in Table 3.

The concavity surrounded by the green circle is the largest with respect to the number of contact atoms. A closeup of the landscape around the concavity is shown in Figure 4b. The size and shape indicate that it could be a ligand-binding site. The PLB index of the concavity is, however, only 0.39. This value is significantly small, and it turns out that, indeed, the concavity was not used by the ligand. The highest PLB index, 2.01, was assigned to the concavity indicated by the red circle. The other indices for this concavity were not the highest, as is shown in Table 3. X-ray analysis revealed that the inhibitor was in actuality bound in this concavity, as

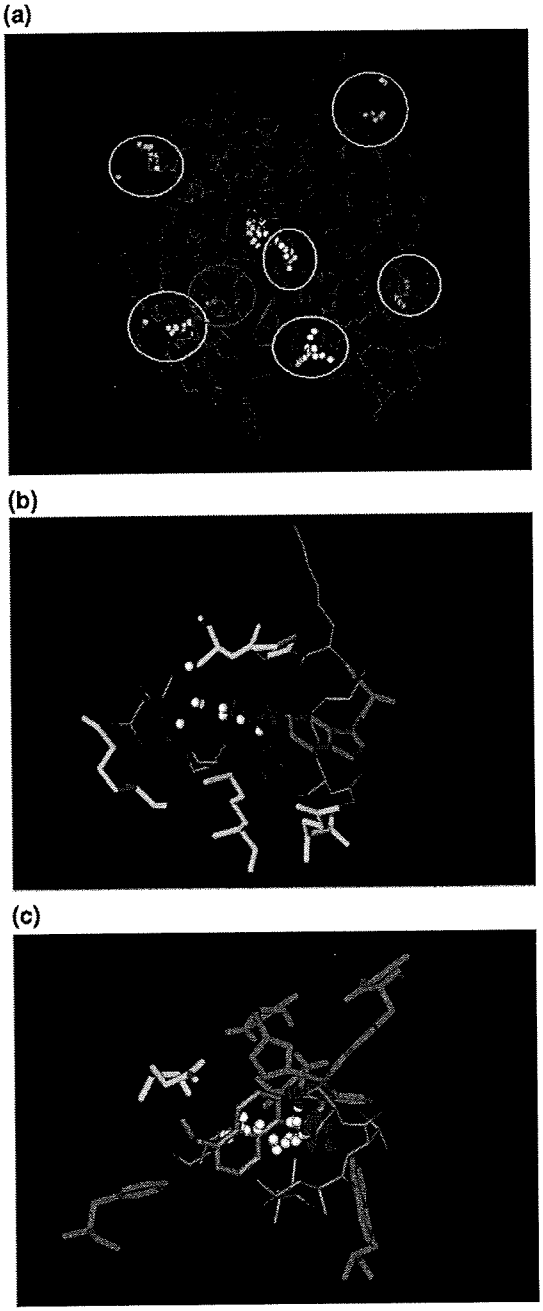

Figure 4. (a) The eight main concavities located in the protein structure of the complex between carbonic anhydrase II and a ligand of 5-(dimethylamino)-1-naphthalenesulfonamide (PDB code: 1OKL). The red and white spheres denote α-spheres (hydrophilic and hydrophobic, respectively). The concavities are circled. The carbon, nitrogen, oxygen, and sulfur atoms are gray, blue, red, and yellow, respectively. (b) A closeup of the largest concavity with respect to the number of atoms in contact with the α-spheres (circled in green in Figure 4a). In this figure, the red and white spheres are α-spheres, the binding-site-philic aromatic residues and Met are light blue, and the binding-site-phobic residues of Pro, Lys, Gln, and Ala are yellow. The PLB index of this concavity is 0.39. (c) A closeup of the true binding site circled in red in part a. The ligand is displayed by a thick stick model in the center of this figure. The red and white spheres denote α-spheres, the binding-site-philic aromatic residues and Met are light blue, and the binding-site-phobic residues of Pro, Lys, Gln, and Ala are yellow. The PLB index of this concavity is 2.01.

Table 3. Indices Characterizing the Concavity Sites Located in the Protein Structure of 1OKL

site	number of contact atoms around the concavity	number of amino acids around the concavity	number of α spheres in the concavity	PLB index
1	57	12	24	0.39
2	51	13	29	1.09
3	50	9	32	−0.59
4	50	12	42	1.88
5	47	11	45	1.35
6	47	10	27	0.25
7	45	10	31	−0.71
8	44	11	44	2.01 true binding site
9	39	8	20	−0.64
10	38	6	13	−0.63
11	30	7	15	−0.73
12	30	7	10	−0.80
13	28	7	11	−1.03
14	26	6	13	−1.34
15	25	8	12	0.33
16	24	6	11	−0.17
17	21	8	12	−0.64

depicted in Figure 4c. This result is indicative of the PLB index's usefulness for distinguishing the correct concavity among multiple possibilities.

Two distinct concavities were found in the protein structure of a ligand−protein complex of retinoic acid receptor γ-1 (PDB code: 1FCZ).[13] The size and shape of one concavity, shown in Figure 5a, made it look like a promising ligand-binding site.

The PLB index of 1.84 for this concavity, however, was appreciably smaller than the corresponding value (3.04) of the concavity shown in Figure 5b. Although there were some binding-site-phobic residues present in both concavities, the former contained five, including two Pro's, which are slightly more binding-site-phobic than the other three, and the latter contained two Lys's. In addition, there were eight binding-site-philic residues in the latter concavity compared with four in the former. The characteristics of the two concavities were distinctively different, and the PLB index indicated that the latter was much more likely to be a ligand-binding site. The inhibitor BMS181156 was actually bound at this site in the crystal structure of the complex.

Laskowski et al.[14] reported that ligand-binding sites tend to be associated with the largest clefts in the surface of a protein. However, the examples above clearly show that binding sites cannot be determined on the basis of size alone. The chemical properties resulting from amino acid composition significantly contribute to the determination of ligand-binding sites. The PLB index takes both the chemical properties and size of the concavity into consideration (eq 4), and its high rate of successful prediction reflects this.

4. CONCLUSIONS

Identification of the potential binding sites of small molecules on a particular target molecule is an important issue in drug discovery. On the basis of the assumption that drug-binding sites on target molecules have specific amino acid compositions, the compositions around the binding sites of drug or druglike compounds were determined by examining the X-ray structures of ligand-protein complexes. As expected, the amino acid compositions around the ligand-

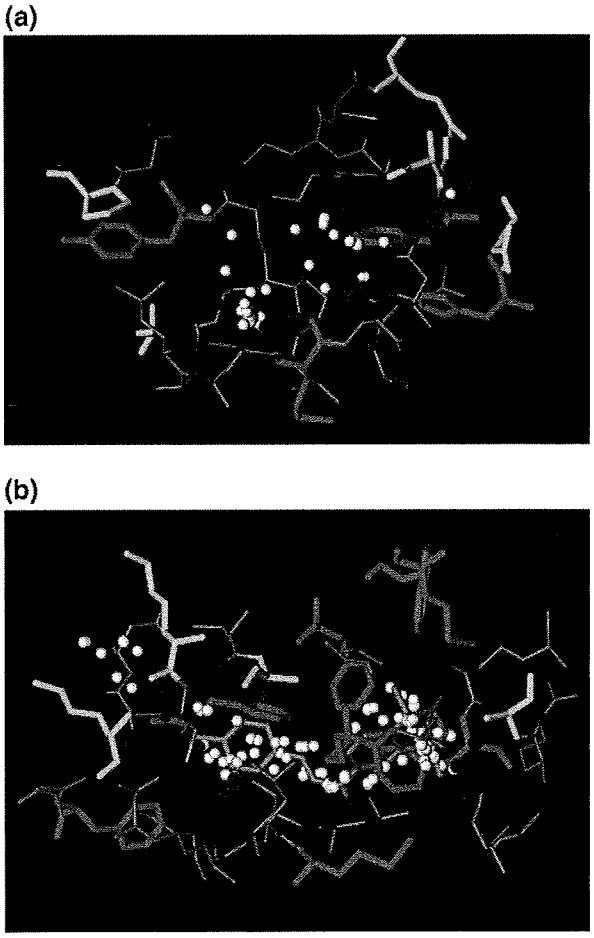

Figure 5. (a) One of the two distinct concavities located in the protein structure of the complex between retinoic acid receptor γ-1 and the ligand BMS181156 (PDB code: 1FCZ). The red and white spheres denote α-spheres; the binding-site-philic aromatic residues and Met are light blue, and the binding-site-phobic residues of Pro, Lys, Gln, and Ala are yellow. The PLB index of this concavity is 1.84. (b) One of the two distinct concavities with the largest PLB index, 3.04. This concavity corresponds to the true binding site of the ligand. The ligand is displayed by a thick stick model in the center of this figure. The red and white spheres denote α-spheres; the binding-site-philic aromatic residues and Met are light blue, and the binding-site-phobic residues of Pro, Lys, Gln, and Ala are yellow.

binding sites were markedly different from those on the surfaces of the proteins. The specific amino acid composition at each ligand-binding site was used to create a novel PLB index. The results of this study show that the PLB index is a good predictor of ligand-binding sites. It is also clear that the molecular interplay between the amino acids at a given site is crucial to the creation of the ligand-binding sites.

It is particularly interesting that the binding sites can be predicted accurately by the specific amino acid composition surrounding the concavities on the surface of proteins. From a practical point of view, this prediction method is useful because the accurate determination of the positions of the amino acids within the concavities would not necessarily be required. The method may be applicable to relatively low-resolution X-ray structures and those constructed using homology modeling. The PLB index would also be useful for identifying ligand-binding sites on novel target molecules with unknown ligands. It is clear that using the PLB index to predict ligand-binding sites could become a useful research tool for various aspects of drug discovery.

REFERENCES AND NOTES

(1) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(2) Laskowski, R. A. SURFNET: A Program for Visualizing Molecular Surfaces, Cavities, and Intermolecular Interactions. *J. Mol. Graphics* **1995**, *13*, 323−330.
(3) Dennis, S.; Kortvelyesi, T.; Vajda, S. Computational Mapping Identifies the Binding Sites of Organic Solvents on Proteins. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 4290−4295.
(4) An, J.; Totrov, M.; Abagyan, R. Comprehensive Identification of "Druggable" Protein Ligand Binding Sites. *Genome Inf.* **2004**, *15*, 31−41.
(5) Edelsbrunner, H.; Facello, M.; Fu, R.; Liang, J. Measuring Proteins and Voids in Proteins. *Proceedings of the 28th Annual Hawaii International Conference on Systems Science*; Publisher: Place of Publication, 1995; pp 256−264.
(6) *MOE (Molecular Operating Environment)*, version 2005.06; Chemical Computing Group Inc.: Montreal, Quebec, Canada, 2006.
(7) Horio, K.; Goto, J.; Hirayama, N. A Simple Method To Improve the Odds in Finding 'Lead-Like' Compounds from a Chemical Library. *Chem. Pharm. Bull.* Submitted.
(8) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic Local Alignment Search Tool. *J. Mol. Biol.* **1990**, *215*, 403−410.
(9) Labute, P. A Widely Applicable Set of Descriptors. *J. Mol. Graphics Modell.* **2000**, *18*, 464−477.
(10) Ertl, P.; Rohde, B.; Selzer, P. Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties. *J. Med. Chem.* **2000**, *43*, 3714−3717.
(11) Hall, L. H.; Kier, L. B. The Molecular Connectivity Chi Indexes and Kappa Shape Indexes in Structure-Property Modeling. *Rev. Comput. Chem.* **1991**, *2*, 367−442.
(12) Nair, S. K.; Elbaum, D.; Christianson, D. W. Unexpected Binding Mode of the Sulfonamide Fluorophore 5-Dimethylamino-1-naphthalene Sulfonamide to Human Carbonic Anhydrase II. Implications for the Development of a Zinc Biosensor. *J. Biol. Chem.* **1996**, *271*, 1003−1007.
(13) Klaholz, B. P.; Mitschler, A.; Moras, D. Structural Basis for Isotype Selectivity of the Human Retinoic Acid Nuclear Receptor. *J. Mol. Biol.* **2000**, *302*, 155−170.
(14) Laskowski, R. A.; Luscombe, N. M.; Swindells, M. B.; Thornton, J. M. Protein Clefts in Molecular Recognition and Function. *Protein Sci.* **1996**, *5*, 2438−2452.

CI6002202

Identification of the Druggable Concavity in Homology Models Using the PLB Index

Shinji Soga,† Hiroki Shirai,† Masato Kobori,† and Noriaki Hirayama*,‡

Molecular Medicine Research Laboratories, Drug Discovery Research, Astellas Pharma Inc., 21 Miyukigaoka, Tsukuba, Ibaraki 305-8585, Japan, and Basic Medical Science and Molecular Medicine, Tokai University School of Medicine, 143 Shimokasuya, Isehara, Kanagawa 259-1143, Japan

Received July 3, 2007

Identification of the druggable concavity, in which drug-like molecules are highly inclined to bind, is an important step in structure-based drug design. We previously proposed an index named PLB (propensity for ligand binding), which is based on the amino acid composition characteristically observed at the small molecule binding sites in the X-ray structures of the complexes between proteins and drug-like small molecules. The PLB index was proven to be useful in identifying the druggable concavities in the quality X-ray structures of proteins. Here, we apply the PLB to predicting the druggable concavity in target proteins using the structures of homologous proteins constructed by homology modeling. In this study, we assembled a set of reference proteins that were accurately determined by X-ray analysis in forms of complexes with drug-like small molecules. Homology models for the reference protein were constructed using multiple homologous proteins as templates. The PLB index was then used to predict the druggable concavity. If the template protein in a complex with a drug-like small molecule was used, the druggable concavity was predicted well, with a prediction rate of 78%. When only the apo protein was available as the template, the practical prediction rate was 71%. Interestingly, even when the percent sequence identity between the reference and template proteins was lower than 30, the PLB index could successfully identify the druggable concavity in some cases. This study demonstrates the practical value of applying the PLB index to identifying the drugabble concavity in the homology model.

1. INTRODUCTION

The specific interaction between a drug molecule and its corresponding target protein is a crucial event in the action of the drug. The surface of target proteins usually contains multiple concavities to which small molecules may bind. Generally, however, a particular drug molecule will bind at a very specific concavity to initiate the subsequent drug action. Identification and characterization of such specific concavities, defined as druggable concavities in this paper, are indispensable to structure-based drug discovery.

We have investigated the binding sites, i.e., concavities, of drug-like molecules in various target proteins whose structures were accurately determined by X-ray analysis. Results showed that the amino acids clustering at these concavities are highly specific, leading us to develop a simple discrimination index named PLB (propensity for ligand binding).[1] Validation studies using relatively high-quality X-ray structures demonstrated that the PLB index is a suitable means of identifying the specific concavity to which a drug-like molecule should bind.

Many drug discovery projects still suffer a lack of suitable high-quality X-ray structures of the target molecule. Against this background, however, the rapid increase in high-quality X-ray structures in the Protein Data Bank (PDB)[2] has markedly facilitated the identification of proteins homologous to the target protein from among the thousands-strong pool of proteins in the PDB. Further support is provided by the homology modeling method, which is also now a reliable means[3] of constructing a reasonable three-dimensional structure of an object protein by the best use of sequence similarity and X-ray structure of a template protein. Given the opportunities offered by these advances, the ability to predict or identify druggable concavities in structures constructed by the homology modeling method would be useful.

We were particularly interested in examining whether the PLB index can be applied to this challenging problem. Here, we investigated the possibility of predicting druggable concavities in homology models using the PLB index.

2. MATERIALS AND METHODS

Figure 1 shows a schematic diagram of the flow of druggable concavity prediction based on homology modeling.

2.1. Use of Quality Protein Structures. A set of 13 416 quality protein structures in the PDB was downloaded on Feb. 21, 2007 and termed the quality dataset. Although not compared in detail in this study, the highest quality structures were used to avoid various expected ambiguities caused by low-quality structures. Quality dataset structures were required to meet two criteria, an X-ray diffraction resolution of less than or equal to 2.5 Å, to ensure that the data were derived from quality structures, and an R_{free} value of less than or equal to 0.24, to ensure adequate concordance between the diffraction data and the structure.

2.2. Selection of Reference Structures. A set of complexes between nonredundant proteins and small molecules

* Corresponding author phone: +81 463 93 1121; e-mail: hirayama@is.icc.u-tokai.ac.jp.
† Astellas Pharma Inc.
‡ Tokai University School of Medicine.

10.1021/ci7002363 CCC: $37.00
Published on Web 10/13/2007

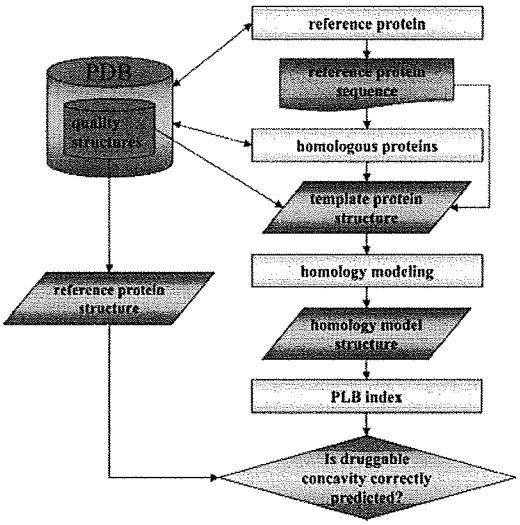

Figure 1. Flow chart of druggable concavity prediction in the homology model.

from the quality dataset was selected and used as reference structures throughout the study. The following four criteria were used to extract the reference structures. First, the small molecule bound in the reference protein had to be a drug-like molecule, as determined using criteria reported previously.[4] Second, occupancy factors and atomic displacement factors of all non-hydrogen atoms of drug-like molecules had to be 1.0 and less than 30 Å^2, respectively. This ensured that the atomic positions of the drug-like molecules were unambiguously determined and, hence, that, on the basis of the atomic positions, the amino acids around the concavity where the drug-like molecules were bound could be clearly identified. Third, the proteins had to be nonredundant, as determined using the "Non-redundant PDB chain set (NR-PDB)" resource (http://www.ncbi.nlm.nih.gov/Structure/VAST/nrpdb.html): this resource clusters all PDB chains into groups of similar chains by sequence similarity and assigns a specific "Group ID" to each group, which was then used to judge nonredundancy. A p-value of 10×10^{-7} was used to judge sequence similarity by BLAST.[5] Where there were multiple proteins with the same group ID, we selected structures which included more drug-like ligand. Fourth, each reference protein had to have multiple homologous proteins in the quality dataset with greater than 80% sequence length alignment with the reference protein. In addition, at least one such protein must be a complex with a small molecule. The strict requirements of the second criterion were dropped for the atomic parameters of the small molecule in the homologous proteins.

On examination, only 15 reference structures were found to fulfill the above requirements (Table 1). The number of homologous proteins for each reference protein is given in Table 2, in which the PDB codes are arranged according to the secondary structure content in ascending order. Table 2 also gives the number of homologous proteins complexed with small molecules. The number of homologous proteins without small molecules is given in Table 4. Since these homologous proteins were used as templates to construct the homology models corresponding to the reference structures, they are named as template proteins hereafter. As shown in Tables 2 and 4, various homologues were selected with percent identities from 20s to 90s.

2.3. Homology Modeling. The homology model corresponding to the reference protein was constructed by the homology modeling method using the X-ray coordinates of a template protein and the amino acid sequence of the reference protein. We used the homology modeling algorithms implemented in the software system MOE.[6] The effect of small molecules bound in the template protein was not taken into account when the model structure was built. Since there are multiple template proteins with different sequence similarities, the degree of concordance between a reference structure and its homology model should vary.

2.4. PLB and R Indices. Specific amino acids tend to cluster at a druggable concavity in a protein. In the previous study,[1] we found that aromatic residues and Met appeared highly frequently at the druggable concavities. On the contrary, the occurrence rates of Pro, Lys, Gln, and Ala at the druggable concavities were significantly low. On the basis of this observation, we introduce the PLB index, which identifies the druggable concavity among other concavities in the protein. The PLB index is defined as follows:

$$\mathrm{PLB}_i = \sum_{x=1}^{20} N_i(x)\,\mathrm{RA}(x) \qquad (1)$$

RA(x) is the ratio of the occurrence rate of amino acid x at the druggable concavities to the occurrence rate of amino acid x on the surface of the proteins. $N(x)$ denotes the number of amino acid x found in a concavity i. Values are summed for all 20 standard amino acids. To distinguish the most druggable concavity from other concavities on a protein surface, the PLB values should be normalized by all concavities in the protein. Z-scored PLBs are used for this purpose. Hereafter Z-scored PLB is designated simply as PLB. Use of the PLB index allows the druggability of each concavity in the protein to be estimated. In this study, the PLB index was applied to identify the druggable concavity in the homology model in order to check how the PLB index can be applied to the homology models, starting from multiple template structures with different sequence similarity.

If the degree of sequence homology between a template and its reference protein is small, the structural discrepancy between the homology model and the reference structure is expected to be large. This means that the concavity in the homology model is to some extent deformed compared to the druggable concavity in the reference structure. This in turn requires evaluation of how accurately the druggable concavity in the reference structure was reconstructed in the homology model. A simple method is to compare the amino acids occurring around these concavities: if there are n amino acids at the druggable concavity in a reference protein and m corresponding amino acids at the relevant concavity in a homology model, the ratio of m/n indicates the goodness of modeling. This ratio, designated as the R index, was used to evaluate the goodness of concavities formed in the homology model, using amino acids located within 4.5 Å of non-hydrogen atoms of the drug-like molecule.

Table 1. Fifteen Reference Proteins Extracted from the Protein Data Bank[2]

PDB code	chain ID	group ID	protein name	enzyme code	ligand name[a]	drug-likeness[b]
1ZUA	X	2	aldo−keto reductase family 1 member B10	1.1.1.−	TOL	14
1E0X	A	33	endo-1,4-β -xylanase A precursor	3.2.1.8	X2F-XYS	13
1BK9		57	phospholipase A2, acidic	3.1.1.4	PBP	12
1TU6	A	59	cathepsin K precursor	3.4.22.38	FSP	14
1W4P	A	73	ribonuclease pancreatic precursor	3.1.27.5	UM3	13
1JZF	A	78	azurin precursor		RTB	13
1YMS	A	126	β-lactamase CTX-M-9a		NBF	14
2WEA		128	penicillopepsin	3.4.23.20	PP6	12
1HEE	A	174	carboxypeptidase A1 precursor	3.4.17.1	ZN-LHY	13
1WBI	A	198	avidin-related protein 2 precursor		BTN	14
1CXV	A	218	collagenase 3 precursor	3.4.24.−	CBP	14
1H4G	A	237	glycoside hydrolase		FXP	13
1TT1	A	473	glutamate receptor, ionotropic kainate 2 precursor		KAI	12
2CYB	A	478	tyrosyl-tRNA synthetase	6.1.1.1	TYR	13
1H60	A	678	pentaerythritol tetranitrate reductase		FMN	14

[a] Abbreviations for ligands used in the PDB. [b] The drug-likeness of small molecules complexed with the proteins was judged using the 14 descriptors in a previous paper.[4] The ranges of these descriptors were calculated to cover 85% of all drugs now used clinically in Japan. The number of the descriptors whose values were within the relevant ranges was used as an index of drug-likeness. For example, a drug-likeness index of 12 means that 12 of 14 descriptors had values within the above ranges.

Table 2. Number of Homologous Proteins Complexed with Small Molecules[a]

% identity[b]	1CXV	1TU6	1JZF	1ZUA	1H60	1W4P	1HEE	1WBI	2WEA	1E0X	1BK9	1TT1	1YMS	1H4G	2CYB	total
20%s	0	2	0	0	3	0	0	6	8	0	0	0	0	0	1	20
30%s	0	0	0	5	4	1	0	4	1	3	0	0	2	0	0	20
40%s	0	7	0	11	5	1	0	0	0	1	3	0	4	3	1	36
50%s	6	4	0	1	1	0	0	1	3	1	2	4	0	1	1	25
60%s	3	0	0	0	0	0	0	1	1	0	0	0	1	0	0	6
70%s	0	0	0	8	0	1	0	0	0	0	0	0	0	0	0	9
80%s	0	0	0	0	0	0	0	2	0	0	0	1	1	0	0	4
90%s	1	2	1	0	6	0	2	0	1	5	0	0	3	0	0	21
total	10	15	1	25	19	3	2	14	14	10	5	5	11	4	3	141

[a] The PDB code of the reference proteins is written at the top of the table. PDB codes were arranged according to the secondary structure content of the structure in ascending order. [b] Value of the % identity is the range of percent identity between a reference and a homologous protein. 20%s means the range between 20% and 30%.

Table 3. Prediction Results of Druggable Concavities in Homology Models Constructed from Template Structures with Small Molecules[a]

% identity[b]	1CXV	1TU6	1JZF	1ZUA	1H60	1W4P	1HEE	1WBI	2WEA	1E0X	1BK9	1TT1	1YMS	1H4G	2CYB	av
20%s		1.00			0.67			0.83	1.00						1.00	0.90
30%s				0.80	1.00	1.00		0.50	1.00	0.67			0.50			0.75
40%s		0.71		1.00	1.00	1.00				1.00	1.00		0.25	0.67	1.00	0.83
50%s	0.83	1.00		1.00	1.00			1.00	1.00	1.00	1.00	0.75		1.00	1.00	0.92
60%s	1.00							1.00	1.00				1.00			1.00
70%s				1.00		1.00										1.00
80%s								1.00				1.00	1.00			1.00
90%s	1.00	1.00	1.00		1.00		1.00		1.00	1.00			1.00			1.00
av	0.90	0.87	1.00	0.96	0.95	1.00	1.00	0.79	1.00	0.90	1.00	0.80	0.64	0.75	1.00	

[a] Values indicate the success rate of prediction. For 1CXV, there were six homology models with a 50%s identity, of which the druggable concavities of five were predicted successfully, giving a success rate of 0.83. The average means the successful prediction rate averaged for each row or column. [b] See footnote *b* of Table 2.

3. RESULTS AND DISCUSSION

3.1. Identification of the Druggable Concavity in Homology Models Constructed from Template Proteins with Small Molecules. Using a template protein that was homologous to a reference protein, a model structure was constructed using the homology modeling method. The concavity formed in the homology model was then compared to that in the reference structure using the R and PLB indices. If we set $R \geq 0.5$ and PLB ≥ 1.2, the druggable concavities of reference protein structures are adequately predicted. These thresholds were therefore used in this study. Evaluation was conducted on the following basis. If the template structure contained a small molecule in the relevant concavity, the protein was expected to be modeled more or less to optimally recognize the small molecule. In the homology model constructed from such a template structure, the druggable concavity should be reconstructed better than that modeled from the template structure without a bound small molecule. It was therefore expected that the latter case would produce a significantly worse prediction rate.

Table 4. Number of Homologous Proteins in the Apo State

% identity[a]	1CXV	1TU6	1JZF	1ZUA	1H60	1W4P	1HEE	1WBI	2WEA	1E0X	1BK9	1TT1	1YMS	1H4G	2CYB	total
20%s	1	1	0	1	0	0	0	1	5	0	0	0	1	0	1	11
30%s	0	2	1	1	0	4	1	2	3	9	0	3	7	1	2	36
40%s	0	3	1	2	0	0	3	0	0	6	9	0	7	5	1	37
50%s	0	3	0	0	0	0	1	0	0	0	8	1	2	2	0	17
60%s	2	0	2	0	0	1	2	2	0	1	1	0	0	1	0	12
70%s	0	0	1	1	0	2	0	0	0	1	0	0	0	0	0	5
80%s	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	3
90%s	0	1	10	0	0	6	4	0	0	1	2	1	2	1	0	28
total	3	10	15	5	0	13	11	5	8	18	20	5	22	10	4	149

[a] The PDB code of the reference protein is written at the top of the table. PDB codes are arranged according to the secondary structure content of the structure in ascending order. The column indicates the range of percent identity between the reference protein and homologous protein. 20%s means the range between 20% and 30%.

Table 5. Prediction Results of Druggable Concavities in Homology Models Constructed from the Apo Template Structures[a]

% identity[b]	1CXV	1TU6	1JZF	1ZUA	1H60	1W4P	1HEE	1WBI	2WEA	1E0X	1BK9	1TT1	1YMS	1H4G	2CYB	av
20%s	0.00	1.00		0.00				0.00	0.60				1.00		1.00	0.55
30%s		1.00	0.00	1.00		0.50	1.00	1.00	1.00	0.44		0.67	0.86	1.00	1.00	0.72
40%s		0.67	0.00	0.50			1.00			0.83	1.00		0.86	1.00	1.00	0.86
50%s		1.00					1.00				1.00	1.00	1.00	1.00		1.00
60%s	1.00		0.50			1.00	1.00	1.00		1.00	1.00			1.00		0.92
70%s			1.00	1.00		0.50				0.00						0.60
80%s													1.00			1.00
90%s		1.00	0.50			0.67	1.00			1.00	1.00	1.00	1.00	1.00		0.75
av	0.67	0.90	0.47	0.60	-	0.62	1.00	0.80	0.75	0.61	1.00	0.80	0.91	1.00	1.00	

[a] Cell values indicate the same as in Table 3, i.e., the success rate of prediction. [b] See footnote *b* of Table 2.

Table 6. Homologous Structures of 2CYB

reference			homologous proteins				
PDB code	chain ID	% identity	PDB code	chain ID	ligand name	protein name	species
2CYB	A	54.6	1J1U	A	TYR	tyrosyl-tRNA synthetase	Methanococcus jannaschii
2CYB	A	40.0	2CYC	B	TYR	tyrosyl-tRNA synthetase	Pyrococcus horikoshii
2CYB	A	22.6	1R6T	A	TYM	tryptophanyl-tRNA synthetase, cytoplasmic	Homo sapiens
2CYB	A	43.1	2CYA	A		tyrosyl-tRNA synthetase	Aeropyrum pernix
2CYB	A	38.6	1N3L	A		tyrosyl-tRNA synthetase, cytoplasmic	Homo sapiens
2CYB	A	37.6	1Q11	A		tyrosyl-tRNA synthetase, cytoplasmic	Homo sapiens
2CYB	A	22.5	1R6T	B		tryptophanyl-tRNA synthetase, cytoplasmic	Homo sapiens

Using the template proteins complexed with small molecules, 141 homology models were constructed as shown in Table 2. The percent identity ranges varied widely, from 20%s to 90%s in general, but were unfortunately not equally wide in all cases. The rates of successfully predicted druggable concavities are given in Table 3. Where multiple homologous proteins occurred in the same bin of percentage, their success rates were averaged. For instance, in the case of 1CXV, there were six homologous proteins in the bin of 50%s. *R* and PLB indices for five of the homology models were equal to or greater than 0.5 and 1.2, respectively, giving a success rate in this bin of 5/6 = 0.83. In the 60%s bin, however, the indices for all homology models exceed the threshold values, so the success rate was 1.0. Apart from three exceptions, the druggable concavities were predicted reasonably well, with 38 of 49 sampling points showing a ratio of 1.0. The prediction rate was a markedly high 78%. Further, even in those cases in which the percent identity was less than 30, the prediction rate was greater than 0.83 in 4 of 5 sampling points in the table. A reasonably good prediction rate was obtained in these difficult situations. These results unequivocally demonstrate that the prediction of druggable concavities using the PLB is highly accurate for template structures having small molecules in the relevant concavities.

3.2. Identification of the Druggable Concavity in Homology Models cConstructed from Template Proteins without Small Molecules. Predicting the druggable concavity in homology models derived from apo protein structures is highly challenging. Using template proteins in the apo state, 149 homology models were constructed by homology modeling as given in Table 4. Since 1H60 has no homologous apo protein in the quality dataset, the total number of proteins in this table is 14. The rates of successfully predicted druggable concavities are given in Table 5. Of 56 points evaluated, a total of 37 had a prediction rate of 1.0. The complete success rate was 66%. From a practical standpoint, points with prediction rates greater than 0.83 can be regarded as successfully predicted, increasing the success rate marginally to 71%. These results demonstrated that if a quality

```
2CYB_A  27   ETKEKPRAYNGYEPSGE-IHLGHWNTYQKLWDLQEA-GFEIIYLIADLHAYLNQKGTFEE  84
             E K+   Y G  PS E  H+GH++      LQ+    ++   Q   YL +  T ++
1R6T_B  55   ENKKPFYLYCGRPSSEAAHVGHLIPFIFTKWLQDVFNVPLYIQMTDDEKYLWKDLTLDD  114

2CYB_A  85   IAEVADYNKKVFIALGLDESRAKFVLGSEYQ-LSRDYVLDVLKWARITTLNRARRCMQEV  143
               A N K  IA G D ++        +Y   S  +  +V+K+ +  T N+ +       +
1R6T_B  115  AYGDAYENAKDIIACGFDIDKTFIFSDLDYMGMSSGFYKNVVKIQKHVTFNQVK----GL  170

2CYB_A  144  SRRKEDPHVSCWIYPLMQAL--------DIAHLGVD---LAVGGIEQRKIHMLARELLPR  192
               +  + ++ +P +QA          I     D  L    IEQ        R+ PR
1R6T_B  171  FGFIDSDGCIGKISFPAIQAAPSFSMSFPDIFRDRTDIGCLIPCAIEQDPYFRKTRDVAPR  230

2CYB_A  193  LGYSSPVCLHTFILVGLDGGKMSSSKGNYISVRDPPEEVERKIRKAYCPAGVVEENPILD  252
             +GY  P  LM+     L G   S       I + D  +++ K+ K +  +G
1R6T_B  231  IGYPKPALLNSTFFPALQGPNSS-------IFLTDTAKQIKTKVNK-HAFSG---------  274

2CYB_A  253  IAKYHILPRFGKIYVERDAKFGGDVE------YASF--------EELAEDFKSGGLHPLD  298
                      G+  +E   +FGG+ +       Y +F        E++ +D+ SG     +
1R6T_B  275  ----------GRDTIEEHRGFGGNCDYDYSFKYLTFFLEDDDKLEQIRKDYTSGAKLTGE  324

2CYB_A  299  LKIAVAKYLNWLLEDARKR  317
             LK A+ + L  L+ + + R
1R6T_B  325  LKKALIEYLQPLIAEHQAR  343
```

Figure 2. Sequence alignment between 2CYB_A and 1R6T_B in BLAST format. The red rectangle indicates the ligand-binding site in the reference protein.

Figure 3. Reference structure (PDB code: 2CYB_A) (red) and homology model (green) constructed from 1R6T_B. The small molecule in the reference structure is shown by a ball-and-stick model.

template protein structure in the apo state was available, the PLB index could adequately predict the druggable concavity in the drug−target protein whose structure was constructed by homology modeling.

3.3. Prediction Rate vs Secondary Structure Content. Performance of the PLB index in predicting the druggable concavity appears reasonably good from a practical point of view. We anticipated that the percent identity would be closely related to the prediction rate. To our surprise, however, druggable concavities were successfully predicted in some cases in which the percent identity was markedly low, e.g., 1YMS, 1H4G, and 2CYB; in these cases, the relevant concavities were perfectly predicted even though the template proteins were in the apo state. We were particularly intrigued by this result. Subsequent evaluation of potential explanations indicated the role of the secondary structures. Using the definition of secondary structures by Kabsch and Sander,[7] contents of the secondary structure in the reference proteins were 0.46, 0.47, 0.5, 0.5, 0.51, 0.52, 0.54, 0.56, 0.57, 0.57, 0.58, 0.61, 0.62, 0.63, and 0.65 for 1CXV, 1TU6, 1JZF, 1ZUA, 1H60, 1W4P, 1HEE, 1WBI, 2WEA, 1E0X, 1BK9, 1TT1, 1YMS, 1H4G, and 2CYB, respectively. Although no direct relationship between the secondary structure content and prediction rate was seen, the secondary structure content of the above three proteins exceeded 60%, indicating that the druggable concavities could be identified by the PLB index provided that the secondary structure content was high, even if the percent identity was low. This trend appeared to hold generally in the set of reference proteins. In the case of 1TU6, however, the prediction rates were markedly high over a wide range of percent identities in spite of the significantly low secondary structure content. Detailed analysis of the role of secondary structure content in the prediction of druggable concavity therefore awaits the accumulation of a quality dataset in the PDB. Until then, however, this general trend can be considered a rule of thumb.

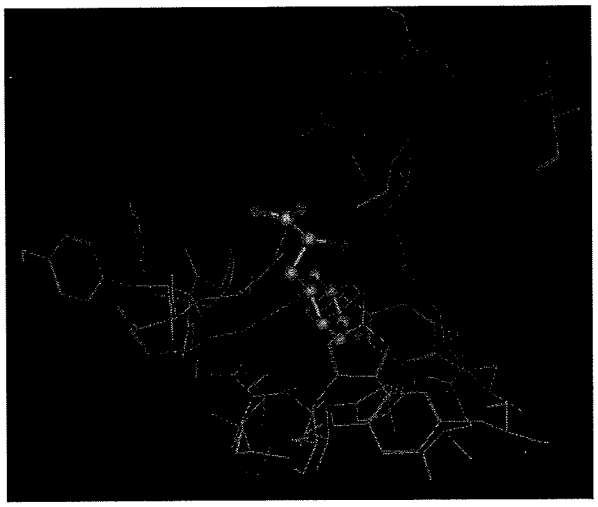

Figure 4. Close-up of the druggable concavity in Figure 3. The reference structure and homology model are shown by red and green lines, respectively. The small molecule is shown by a ball-and-stick model.

3.4. Typical Prediction of the Druggable Concavity in a Homology Model. In this example, the reference protein is tyrosyl-tRNA synthetase from *Archaeoglobus fulgidus* (2CYB). The homologous proteins used as template structures are given in Table 6. Since the percent identity is low, i.e., 22.5%, tryptophanyl-tRNA synthetase from human cytoplasma (1R6T) seems to be a practical example. The sequence alignment used for homology modeling is shown in Figure 2, and the homology model and X-ray structure of tyrosyl-tRNA synthetase are superimposed in Figure 3. Although the overall structures correspond reasonably well, the dispositions and conformations of amino acids around the concavity differ significantly, as shown in the expanded view (Figure 4). Nevertheless, prediction of the druggable concavity was successful. The reason for this may be that the PLB index does not depend on the detailed structures around the concavity and is good at capturing the spirit of the druggable concavity. This is clearly the most advantageous aspect of prediction using the PLB index.

4. CONCLUSIONS

In the postgenomic era, structural information regarding disease-related proteins is now increasing at an explosive

rate. This in turn has resulted in ever higher demands for identification of the druggable concavity in drug-target proteins from sequence data alone. Thanks to recent technical advances, the number of protein structures determined by X-ray analysis is increasing, facilitating the identification of X-ray structures reasonably homologous to a target amino acid sequence. The modeling of a protein structure from the amino acid sequence has been one of the most attractive topics in protein science, but, despite much effort, ab initio modeling from the sequence only remains challenging. Against this background, the homology modeling method, which provides the best use of the three-dimensional structure of homologous proteins, has now advanced to practical use. The availability of a suitable methodology that could exploit these data to predict the drug-binding concavity in the target protein would greatly assist drug discovery.

In this paper, we investigated the use of the PLB index to meet this need. The PLB index, based on the specific amino acid compositions at drug-binding concavities in drug-binding proteins, serves as an index to assess the druggability of the relevant concavity in the drug−target protein. Results showed that the PLB index successfully identified the druggable concavity for a set of quality reference proteins selected for this study. For template structures with a small molecule bound in the relevant concavity, the success rate of prediction was a markedly high 78%. Moreover, the success rate was 71% even in more practical cases in which the template structure was not a complex with a small molecule. The present study demonstrates that the PLB index can be used to identify the druggable concavity in homology models. Although the essential condition for the proper application of the PLB index is now a high-quality template structure, the expanding pool of quality structures in the PDB will clearly increase the applicability of this method.

REFERENCES AND NOTES

(1)) Soga, S.; Shirai, H.; Kobori, M.; Hirayama, N. Use of amino acid composition to predict ligand-binding sites. *J. Chem. Inf. Model.* **2007**, *47*, 400−406.
(2)) Berman, H. M.; Westbrook, J.; Fenz, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The protein data bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(3)) Marti-Renom, M. A.; Madhusudhan, M. S.; Fiser, A.; Rost, B.; Sali, A. Reliability of assessment of protein structure prediction methods. *Structure* **2002**, *10*, 435−440.
(4)) Horio, K.; Goto, J.; Muta, H.; Hirayama, N. A Simple method to improve the odds in finding 'lead-like' compounds from a chemical library. *Chem. Pharm. Bull.* **2007**, *55*, 980−984.
(5)) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403−410.
(6)) *MOE (Molecular Operating Environment)*, version 2006.0801; Chemical Computing Group Inc.: Montreal, Quebec, Canada, 2006.
(7)) Kabsch, W.; Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**, *22*, 2577−2637.

CI7002363

ANALYTICAL SCIENCES 2007, VOL. 23

X-ray Structure Analysis Online

Structure of Etoposide

Rumiko TANAKA and Noriaki HIRAYAMA†

Basic Medical Science and Molecular Medicine, Tokai University School of Medicine, Boseidai, Isehara, Kanagawa 259-1193, Japan

The title compound, $C_{29}H_{32}O_{13}$, is a potent antitumor drug. Its crystal belongs to space group $P2_1$ with cell dimensions a = 11.527(4), b = 6.215(2), c = 21.670(7)Å, and β = 100.60(3)°. The final R value is 0.099. The characteristic four-ring system of the aglycone moiety takes a slightly curved structure with the dihedral angle between the two five-membered rings at both ends being 12(1)°. Both the pyranose and 1,3-dioxane rings take chair conformations.

(Received November 16, 2006; Accepted December 22, 2006; Published on web February 26, 2007)

The chemical structure of etoposide [(9-[(4,6-*O*-ethylidene-β-D-glucopyranosyl)oxy]-5,8,8a,9-tetrahydro-5-(4-hydroxy-3,5-dimethoxyphenyl)furo[3′,4′:6,7]naphtho[2,3-*d*]-1,3-dioxol-6(5a*H*)-one)] is shown in Fig. 1. Etoposide forms a ternary complex with topoisomerase II and DNA, leading to an accumulation of DNA breaks, and finally cell death.[1] Etoposide is used primarily for treating testicular tumors and small cell carcinoma of the lung. Although the structure of the aglycone moiety was reported,[2] the structure of the glycoside has not been determined so far. Since etoposide is a potential anticancer drug that is clinically used now, it is important to disclose its inherent three-dimensional structure.

Etoposide was purchased from Sigma Co. Colorless platelet single crystals of the molecule were grown from a methanol solution. It was very difficult to obtain large crystals of good quality. In addition, the obtained crystals were very fragile. The crystal (0.3 × 0.1 × 0.07 mm) was mounted on a glass fiber and used for data collection. The crystal was the biggest ever obtained. The structure was solved by direct methods and refined by a full-matrix least-squares method. Three water molecules of crystallization were found during the X-ray analysis. All of them were disordered and the occupancy factors were less than 1.0 (0.25, 0.35, 0.40, respectively). Due to the disordered water molecule and the small and poorly diffracting crystal, the number of significant reflections observed was fairly small. Although oxygen atoms were refined anisotropically, carbon atoms were refined isotropically to fulfill the requirement of the crystallographic refinement. The hydrogen atoms of the hydroxyl groups and water molecules could not be located. The positions of other H-atoms were geometrically calculated, and not refined. The absolute configuration of the molecule was suggested by referring to that of D-glucose. The crystal and experimental data are given in Table 1.

The molecular structure with the ring labeling system drawn by ORTEP-III[5] is shown in Fig. 2. Selected bond lengths and bond angles are given in Table 2. Both the pyranose and 1, 3-dioxane rings take chair conformations. Binding of the sugar moiety brings about a significant conformational difference in the aglycone moiety. The torsion angles in rings A, C and D are given in Table 3 together with the corresponding values in the aglycone crystal.[2] Rings C in both structures take half-chair conformations. However, the torsion angles around C6, C9 and

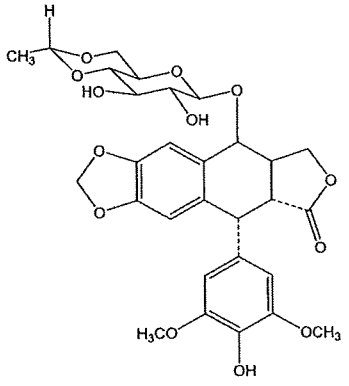

Fig. 1 Chemical structure of Etoposide.

Table 1 Crystal and experimental data

Formula: $C_{29}H_{32}O_{13}\cdot H_2O$	
Formula weight = 606.58	
Crystal system: monoclinic	
Space group: $P2_1$	Z = 2
a = 11.527(4)Å	
b = 6.215(2)Å	β = 100.60(3)°
c = 21.670(7)Å	
V = 1526.0(9)Å^3	
D_x = 1.320 g/cm^3	
No. of observations (I>3.00σ(I)) = 1150	
θ_{max} = 68.22° with Cu K_α	
R(I>3.00σ(I)) = 0.099	
$(\Delta\sigma)_{max}$ = 0.000	
$(\Delta\rho)_{max}$ = 0.40 eÅ^3	
$(\Delta\rho)_{min}$ = – 0.30 eÅ^3	
Measurement: Rigaku RAXIS-RAPID	
Program system: CrystalStructure 3.6.0[3]	
Structure determination: SIR92[4]	
Refinement: full-matrix	
CCDC No.630746	

To whom correspondence should be addressed.
E-mail: hirayama@is.icc.u-tokai.ac.jp