Application for evaluating and visualizing the sequence conservation of ligand-binding sites

Nobuaki Yasuo[1,a)] Masakazu Sekijima[1,2,b)]

Abstract: We developed a new application to quantitatively evaluate the sequence conservation of ligand-binding sites by integrating information pertaining to protein structures, ligand-binding sites, and amino acid sequences. These data are visualized onto protein structures via a Jmol or PyMOL interface. The visualization is very important for structure-based drug design (SBDD). Key features of this application are the visualization of slight differences in specific ligand-binding sites and ConservationScore comparable among ligand-binding sites. Furthermore, we conducted an experiment to visualize the calculation and comparison of the ConservationScore of four viral proteins as well as an experiment to visualize the differences between proteins belonging to the human β adrenergic receptor family. This application is available at http://www.bio.gsic.titech.ac.jp/visco.html .

Keywords: amino acid sequence conservation, ligand-binding site, visualization, structure-based drug design (SBDD)

1. Introduction

In recent years, the amount of data on three-dimensional protein structure has significantly increased owing to advances in determining three-dimensional protein structures using methods such as X-ray crystallography or nuclear magnetic resonance[1][2]. For this reason, structure-based drug design (SBDD) based on these protein structures is considered as an effective method[3]. However, simulations used for SBDD, such as molecular docking or molecular dynamics, inadequately consider the diversity of target proteins because there are still fewer available protein structures than amino acid sequences.

To design antibacterial or antiviral drugs, the sequence diversity of a target protein is key, because the diversity of proteins is the reason for the origin of drug resistance[4]. If the target site is variable, drug-resistant bacteria or viruses will soon appear and the new drug will have no effect. Furthermore, designing candidate drug molecules must avoid those with adverse effects. When we use SBDD methods to design molecules than do not bind off-target proteins, knowledge pertaining to the sequence conservation of ligand-binding sites between targets and off-targets is important. Therefore, differences between targets and off-targets will improve selectivity[5].

Applications are available to visualize, particularly the conserved amino acid sequences among many sequences on protein structures, such as AL2CO[6], ET viewer[7], VENN[8], and ConSurf[9][10]; however, these applications are optimized to visualize the conservation of an entire protein structure. Thus, these applications cannot adequately represent slight differences between ligand-binding sites.

In the present study, we developed a new application to quantitatively evaluate the amino acid sequence conservation of ligand-binding sites and to visualize them on a protein surface. To predict the sequence conservation of specific ligand-binding sites, this application integrates data pertaining to sequences, structures, and ligand-binding sites. We also developed a novel method for the evaluating sequence conservation of specific ligand-binding sites.

2. Methods

2.1 Overview

This application employs the python-based command-line tool (**Fig. 1**), wherein the user provides inputs of three-dimensional structure of a protein, a list of residues that form ligand-binding site, and a multiple amino acid sequence alignment. This application integrates these data, calculates ConservationScore, and provides output data to Jmol[11] or PyMOL[12].

2.2 Calculation of ConservationScore

The ConservationScore is calculated on the basis of the conservation rate of each amino acid residue. In this application, we define $p_{max}(i)$ as the conservation rate of an amino acid residue, where i is the number of residues that form the ligand-binding site. $p_{max}(i)$ is described as

$$p_{max}(i) = \max_{x \in A} p_i(x) \tag{1}$$

where $p_i(x)$ is the probability of the appearance of amino acid x in the i -th residue and A is the set of amino acids. $p_{max}(i)$ takes a value from $0 \leq p_{max}(i) \leq 1$ and is important because that it

[1] Department of computer science, Tokyo Institute of Technology, Meguro, Tokyo 152–8550, Japan
[2] Global Scientific Information and Computing Center, Tokyo Institute of Technology, Meguro, Tokyo 152–8550, Japan
[a)] yasuo.n.aa@m.titech.ac.jp
[b)] sekijima@gsic.titech.ac.jp

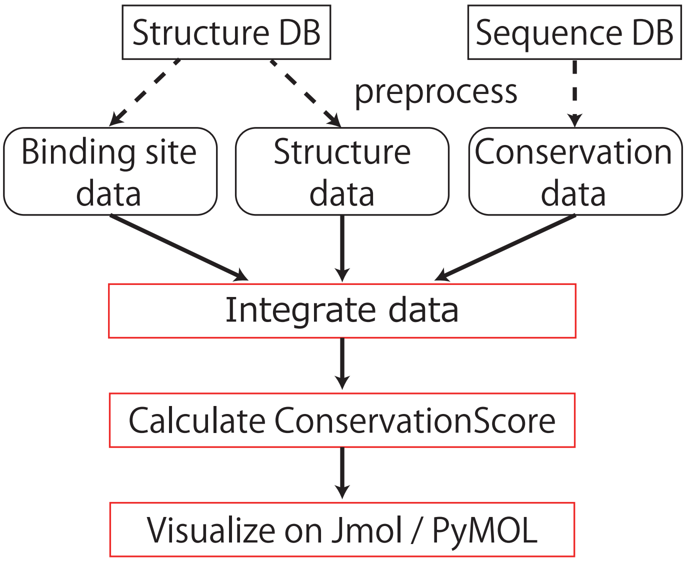

Fig. 1 Flowchart

can be compared with another amino acid without considering species differences or the number of amino acid sequences.

We defined $C(s)$ as the ConservationScore of a ligand-binding site, where s is a ligand-binding site. $C(s)$ is described as

$$C(s) = \sqrt[|s|]{\prod_{i \in s} \{p_{max}(i)\}^2} \quad (2)$$

This equation defines $C(s)$ as the geometric mean of $\{p_{max}(i)\}^2$. This score takes a value of $0 \leq p_{max}(i) \leq 1$, and it is important because it can be readily compared with another ligand-binding site without considering the number of residues.

2.3 Data selection

The user provides inputs of the result of a multiple sequence alignment for use in many analyses. The sequence database and range of sequence similarity changed completely so that users can visualize the differences among individuals of the same species, related species, or species that belong to different families.

This application accepts a list of residue numbers or a list of lists of residue numbers as ligand-binding sites. If a protein comprises ≥ 2 ligand-binding sites, the list of lists of residue numbers can be used. Residue numbers are stored in a PDB file. User may use external tools such as Sitemap[13], FINDSITE[14], ConCavity[15], or COFACTOR[16] to find ligand-binding sites instead of visual inspection or ligand-bound structures.

3. Experiment

To demonstrate the results of our application, we performed experiments using drug target proteins.

3.1 Visualization and Calculation of the ConservationScore of viral proteins

The names and structure IDs of target proteins are listed in **Table 1**. ligand-binding sites were determined using Sitemap on ”evaluate a single binding site region” mode and selecting the ligand in the A-chain monomer of the crystal structure[13] except for the human immunodeficiency (HIV) protease. The ligand-binding site of the HIV protease was determined without deleting the B-chain, because this ligand-binding site residues within the interface of a dimer. The definition of ligand-binding site was 6 Å from each site-point of the Sitemap output.

Table 1 Target proteins

Virus	Protein name	PDB ID
Influenza A virus (FLUV)	Neuraminidase	2HU4
Influenza A virus (FLUV)	Endonuclease	4M4Q
Human immunodeficiency virus (HIV)	Reverse transcriptase	1JLB
Human immunodeficiency virus (HIV)	Protease	1SDT

Amino acid sequences were obtained from Uniref90[17] using BLAST[18] in the order of E-value until non-target sequences appeared as this. Queries for BLAST analyses were those of the crystal structures. The numbers of sequences were 920, 619, 69, and 64. Multiple alignments ware performed using ClustalW version 2.1 with the default settings [19].

3.2 Visualization of differences in ligand-binding sites among human G-protein coupled receptors (GPCRs)

In this experiment, the target proteins were the human β-adrenergic GPCRs. The template structure was β_2-adrenergic GPCR (PDB ID 2RH1). The ligand-binding site was determined using the same method described in Section 3.1. We compared the sequences of one each of the β_1 and β_3-adrenergic GPCRs obtained from Uniprot[20]. Furthermore, multiple alignment was performed as described in Section 3.1.

4. Results and Discussion

4.1 Visualization and Calculation of the ConservationScore of viral proteins

The results of visualization are shown in **Fig. 2**. Fig. 2(a) and (b) show the influenza A virus (FLUV) neuraminidase and were visualized using Jmol and PyMOL, respectively. The results that follow were visualized using PyMOL. Fig. 2(c) shows the FLUV endonuclease. Fig. 2(d) and (e) show the HIV reverse transcriptase and protease, respectively. In the visualized output, the

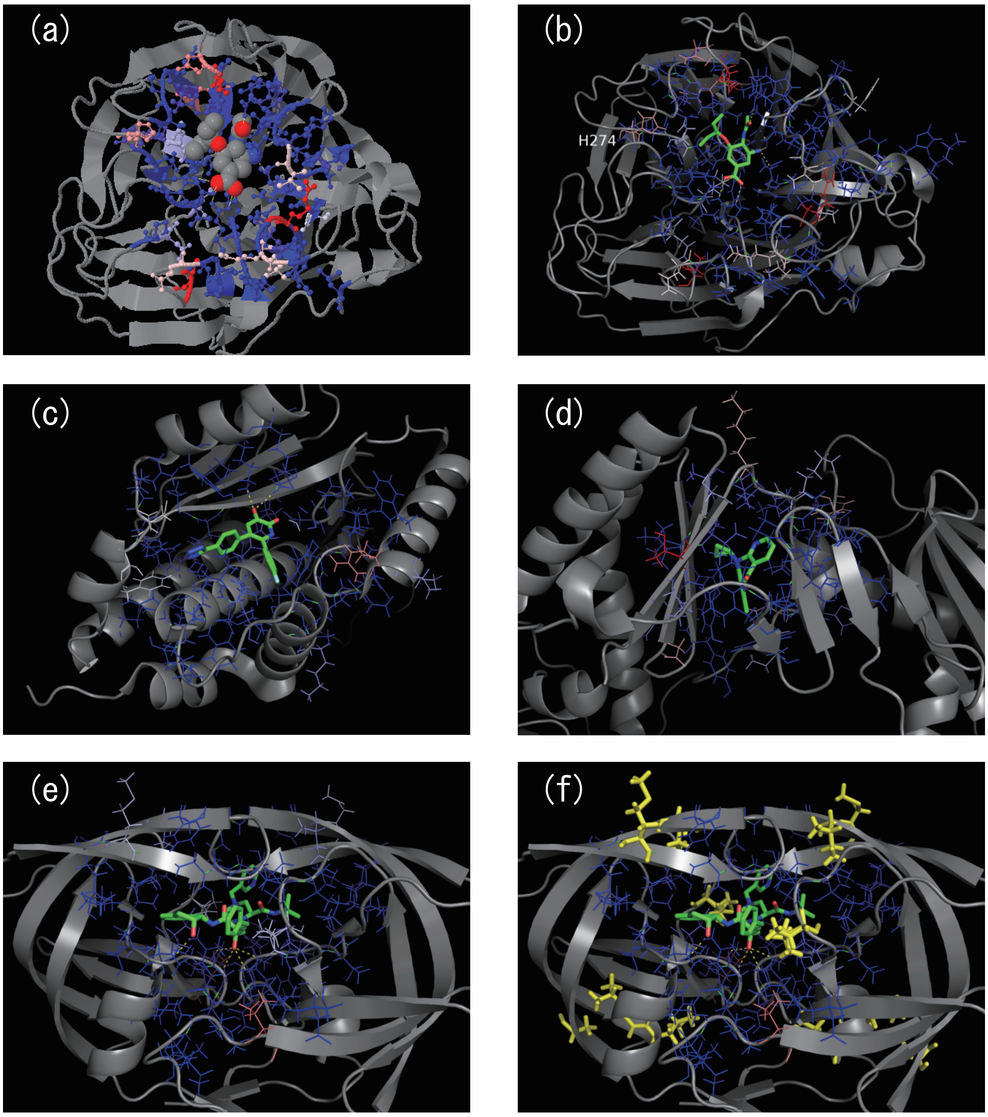

Fig. 2 Visualization of viral proteins

residues in the ligand-binding site are colored by $p_{m}ax(i)$ which represents the conservation rate. Residues are depicted by gradations from blue, white, to red according to $p_{max}(i)$. This application visualizes 2%–3% of the difference between conservation rates in ligand-binding sites such as the blue and aqua-blue residues in Fig. 2(c).

The results show that the amino acid residues that interacted with ligands are highly conserved, although some varied even in the ligand-binding site. In Fig. 2(b), the ligand is the anti-FLUV drug oseltamivir (Tamiflu). This result demonstrates that this application detects the oseltamivir-resistant mutation H274Y, which is poorly conserved among FLUV strains[21]. Similarly, the ligand of the HIV protease is indinavir, and this application detects the mutation that imparts drug resistance (yellow stick) as shown in Fig. 2(f)[22].

The ConservationScore of each ligand-binding site were 0.813, 0.9079, 0.7692,and 0.9026. These results show that the FLUV endonuclease and HIV protease are more highly conserved compared with the FLUV neuraminidase and HIV reverse transcriptase.

We were intrigued to find that the ConservationScore is related to the acquisition of drug resistance. The FLUV endonuclease is a highly conserved target among influenza A, B, and C viruses, although there are no approved drugs[23]. The drug-resistant mutation of the HIV protease is known; however, the protease inhibitor indinavir, reduces virus titers for at least some weeks[22]. In contrast, the FLUV neuraminidase harbors many resistance mutations and may cause drug resistance only two days after administration[21]. Similarly, the HIV reverse transctiptase inhibitor nevirapine, induces resistance after its single dose is administered to infants[24].

4.2 Visualization of differences in ligand-binding sites among human GPCRs

The results are shown in **Fig. 3**, and Fig. 3(b) presents the same orientation shown in (a), except for the cartoons. The upper part of the figure corresponds to extracellular domain. Users can easily manipulate the results using Jmol or PyMOL.

In this example, the user may wish to reveal the region opposite to the ligand-binding site. This tool can be applied to the variable

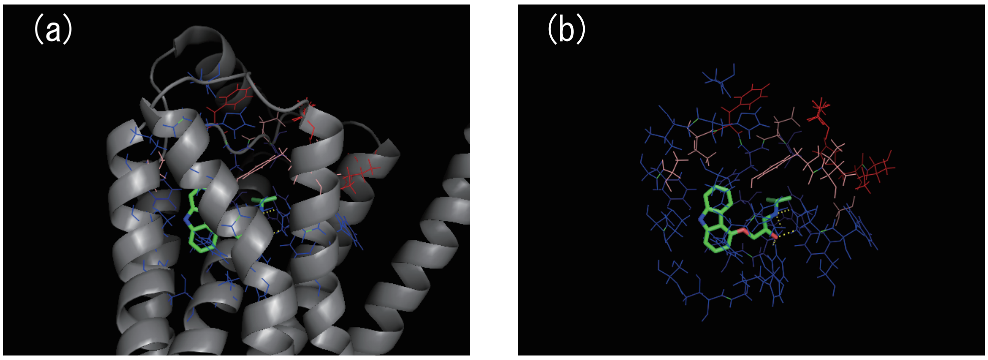

Fig. 3 Visualization of GPCR

regions required to design molecules with enhanced selectivity or to the conserved regions required to avoid drug resistance. Here, the sequence of the extracellular region is variable, although the sequence of the intracellular region of the ligand-binding site is completely conserved. Such knowledge may be useful to design selective agonists or antagonists.

4.3 Conclusion

We developed a new application to quantitatively evaluate the sequence conservation of ligand-binding sites and to visualize them by projecting them onto protein surfaces. The technique is useful for structure-based drug design, and the Conservation-Score reflects the mechanism of acquisition of drug resistance. The application of the visualization tools included in this application is limited because at least one protein structure is required to visualize the results. This may be addressed by building a model structure based on a homologous structure.

Acknowledgments This research was partially supported by JSPS KAKENHI Grant Numbers 25870222 and 24240044.

References

[1] Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. and Tasumi, M.: The protein data bank, *European Journal of Biochemistry*, Vol. 80, No. 2, pp. 319–324 (1977).

[2] Anon: Hard data: It has been no small feat for the Protein Data Bank to stay relevant for 100,000 structures, *Nature*, Vol. 509, p. 260 (2014).

[3] Kalyaanamoorthy, S. and Chen, Y. P. P.: Structure-based drug design to augment hit discovery, *Drug discovery today*, Vol. 16, No. 17, pp. 831–839 (2011).

[4] Davies, J. and Davies, D.: Origins and evolution of antibiotic resistance, *Microbiology and Molecular Biology Reviews*, Vol. 74, No. 3, pp. 417–433 (2010).

[5] Huggins, D. J., Sherman, W. and Tidor, B.: Rational approaches to improving selectivity in drug design, *Journal of medicinal chemistry*, Vol. 55, No. 4, pp. 1424–1444 (2012).

[6] Pei, J. and Grishin, N. V.: AL2CO: calculation of positional conservation in a protein sequence alignment, *Bioinformatics*, Vol. 17, No. 8, pp. 700–712 (2001).

[7] Morgan, D. H., Kristensen, D. M., Mittelman, D. and Lichtarge, O.: ET viewer: an application for predicting and visualizing functional sites in protein structures, *Bioinformatics*, Vol. 22, No. 16, pp. 2049–2050 (2006).

[8] Vyas, J., Gryk, M. R. and Schiller, M. R.: VENN, a tool for titrating sequence conservation onto protein structures, *Nucleic acids research*, Vol. 37, No. 18, pp. e124–e124 (2009).

[9] Celniker, G., Nimrod, G., Ashkenazy, H., Glaser, F., Martz, E., Mayrose, I., Pupko, T. and Ben-Tal, N.: ConSurf: using evolutionary data to raise testable hypotheses about protein function, *Israel Journal of Chemistry*, Vol. 53, No. 3-4, pp. 199–206 (2013).

[10] Landau, M., Mayrose, I., Rosenberg, Y., Glaser, F., Martz, E., Pupko, T. and Ben-Tal, N.: ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures, *Nucleic acids research*, Vol. 33, No. suppl 2, pp. W299–W302 (2005).

[11] Jmol: an open-source Java viewer for chemical structures in 3D, http://www.jmol.org/.

[12] Schrodinger, LLC.: The PyMOL Molecular Graphics System, Version 1.2r3pre, http://www.pymol.org/.

[13] Halgren, T. A.: Identifying and characterizing binding sites and assessing druggability, *Journal of chemical information and modeling*, Vol. 49, No. 2, pp. 377–389 (2009).

[14] Skolnick, J. and Brylinski, M.: FINDSITE: a combined evolution/structure-based approach to protein function prediction, *Briefings in bioinformatics*, p. bbp017 (2009).

[15] Capra, J. A., Laskowski, R. A., Thornton, J. M., Singh, M. and Funkhouser, T. A.: Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure, *PLoS computational biology*, Vol. 5, No. 12, p. e1000585 (2009).

[16] Roy, A., Yang, J. and Zhang, Y.: COFACTOR: an accurate comparative algorithm for structure-based protein function annotation, *Nucleic acids research*, p. gks372 (2012).

[17] Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R. and Wu, C. H.: UniRef: comprehensive and non-redundant UniProt reference clusters, *Bioinformatics*, Vol. 23, No. 10, pp. 1282–1288 (2007).

[18] Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J.: Basic local alignment search tool, *Journal of molecular biology*, Vol. 215, No. 3, pp. 403–410 (1990).

[19] Thompson, J. D., Higgins, D. G. and Gibson, T. J.: CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, *Nucleic acids research*, Vol. 22, No. 22, pp. 4673–4680 (1994).

[20] UniProt Consortium and others: UniProt: a hub for protein information, *Nucleic Acids Research*, p. gku989 (2014).

[21] Ives, J., Carr, J., Mendel, D., Tai, C., Lambkin, R., Kelly, L., Oxford, J., Hayden, F. and Roberts, N.: The H274Y mutation in the influenza A/H1N1 neuraminidase active site following oseltamivir phosphate treatment leave virus severely compromised both in vitro and in vivo, *Antiviral research*, Vol. 55, No. 2, pp. 307–317 (2002).

[22] Zhang, Y. M., Imamichi, H., Imamichi, T., Lane, H. C., Falloon, J., Vasudevachari, M. and Salzman, N. P.: Drug resistance during indinavir therapy is caused by mutations in the protease gene and in its Gag substrate cleavage sites., *Journal of virology*, Vol. 71, No. 9, pp. 6662–6670 (1997).

[23] Yuan, P., Bartlam, M., Lou, Z., Chen, S., Zhou, J., He, X., Lv, Z., Ge, R., Li, X., Deng, T. et al.: Crystal structure of an avian influenza polymerase PAN reveals an endonuclease active site, *Nature*, Vol. 458, No. 7240, pp. 909–913 (2009).

[24] Micek, M. A., Blanco, A. J., Beck, I. A., Dross, S., Matunha, L., Montoya, P., Seidel, K., Gantt, S., Matediane, E., Jamisse, L. et al.: Nevirapine resistance by timing of HIV type 1 infection in infants treated with single-dose nevirapine, *Clinical Infectious Diseases*, Vol. 50, No. 10, pp. 1405–1414 (2010).