Topological data analysis based on mixed-solvent molecular dynamics simulations enhance cryptic pocket detection

Jun Koseki[1,], Chie Motono[1,2], Keisuke Yanagisawa[3,4], Genki Kudo[5], Ryunosuke Yoshino[6,7], Takatsugu Hirokawa[6,7], Kenichiro Imai[1,8,*]*

1. Cellular and Molecular Biotechnology Research Institute, National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, 135-0064, Japan.

2. Computational Bio Big-Data Open Innovation Laboratory (CBBD-OIL), National Institute of Advanced Industrial Science and Technology (AIST), Waseda University, Shinjuku-ku, Tokyo 169-8555, Japan

3. Department of Computer Science, School of Computing, Tokyo Institute of Technology, Tokyo, 152-8550, Japan

4. Middle Molecule IT-based Drug Discovery Laboratory (MIDL), Tokyo Institute of Technology, Tokyo, 152-8550, Japan

5. Physics Department, Graduate School of Pure and Applied Sciences, University of Tsukuba, Ibaraki, 305-8571, Japan

6. Division of Biomedical Science, Faculty of Medicine, University of Tsukuba, Ibaraki, 305-8575, Japan

7. Transborder Medical Research Center, University of Tsukuba, Ibaraki, 305-8577, Japan

8. Global Research and Development Center for Business by Quantum-AI Technology (G-QuAT), National Institute of Advanced Industrial Science and Technology (AIST), Ibaraki 305-8560, Japan.

Cryptic site, Mixed-solvent Molecular Dynamics, Topological data Analysis

ABSTRACT: Some functional proteins change their conformation to reveal a hidden binding site when approached by a binding molecule. Such binding sites are called cryptic sites and are important targets for drug discovery. However, it is still difficult to predict cryptic sites correctly. Therefore, we propose a new method to correctly detect cryptic sites using persistent homology method, one of the topological data analysis and mixed-solvent molecular dynamics (MSMD) simulations. To detect hotspots corresponding to cryptic sites, we performed the MSMD simulations using six different probes with different chemical properties (Benzene, Isopropanol, Phenol, Imidazole, Acetonitrile, and Ethylene glycol) and then performed our topological data analysis method, DAIS, to rank hotspots in the order of descending possibility of cryptic sites. For nine target proteins harboring cryptic sites, the proposed method significantly outperformed the accuracy of recent machine learning methods. As a result, in 6 out of 9 cases, the correct hotspots were ranked 1. The proposed method searched for hotspots on the "ligandable" protein surface using MSMD simulations with six different probes and detected potential cryptic sites based on DAIS-based estimates of the protein's "conformational variability". This synergistic combination allowed us to predict cryptic sites with high accuracy.

■ INTRODUCTION

The conformational diversity of protein emerged from protein dynamics leads to multiple functional structures[1]. Conformational change based on structural fluctuations and/or interaction with a small compound could give rise to functional pockets not detectable in ligand-free crystal structures; the pockets are called cryptic pockets[2]. The cryptic pockets are distant from active sites and binding small compounds to such pockets could control protein function allostery. Indeed, allosteric modulators frequently bind to flexible regions including cryptic pockets because molecular mechanisms of allostery are based on the dynamic nature of proteins[2, 3]. Thus, cryptic pockets are important targets to expand the scope of drug discovery in terms that identification of the pockets is likely to lead to the discovery of allosteric modulators for inhibition or activation of protein function and drugs for proteins considered undruggable[4]. However, most cryptic pockets have been discovered serendipitously when both experimental protein structures with open ligand-binding sites and closed binding sites are revealed[2, 5]. Intentional discovery of cryptic pockets is a hard task. Also, usually used definition of cryptic pocket is ambiguous. For example, in making the dataset consisting of 93 proteins used in development of the prediction of cryptic site, Cryptosite[6], a site is defined as being cryptic if it is not detected by FPocket[7] or ConCavity[8] based on the comparison of one bound and one unbound structure. However, about half of the proteins have at least one unbound structure which could bind the ligand without crashing with the protein[9].

Considering the appearance of cryptic pockets is based on the dynamic nature of proteins, molecular dynamics (MD) simulation is a promising tool for exploring cryptic pockets[10]. The

mere detection of a cryptic pocket is not sufficient for drug target discovery and the newly found pocket needs to be ligandable. Mixed-solvent MD (MSMD) simulation is a technique relying on the use of small probes (chemical fragments) such as isopropanol and benzene mixed with water molecules. MSMD simulation has the advantage of observing protein flexibility and mapping "hotspots" on protein surfaces; hotspots are identified by the probe having higher occupancies on the protein surface in the MD simulation. Thus, MSMD improves the sampling of the formation of cryptic pockets concurrently with providing an estimation for the pocket location based on mapping of hotspots. Indeed, several methods have been utilized to explore cryptic pockets[11-15]. However, many hotspots are generally observed in a protein surface. Therefore, one of the goals of the MSMD protocol is to correctly estimate a hotspot corresponds to the location of the cryptic site from all hotspots. Kimura et al. chose eight proteins from the Cryptosite dataset[6] and they can map six of the eight cryptic pockets within the top 3 hotspot rank; they performed 100 ns MD runs on each system and either 5% concentrations of isopropanol, resorcinol, or acetic acid as probes and ranked the site based on the free energy computed from probe occupancy[11]. Schmidt et al. performed 10–500 ns MSMD with 10% phenol against seven proteins to identify pocket cores that overlapped with the proposed cryptic pockets in the proteins. Then, they were successful in detecting these sites on six of the seven proteins using the hotspots ranked in top 3[12]. Smith and Carlson chose twelve proteins characterized by a wide range of conformational changes: The dataset does not overlap with Kimura's data. They performed 100 ns conventional and accelerated MD simulations and then detected cryptic sites in the seven proteins using five different probes and hotspot mapping[15]. As shown by these studies, it is still difficult to detect hotspots matching cryptic pockets only using hotspot ranking based on probe occupancy. Furthermore, the observation of opening of cryptic pockets is more tough task[11, 12, 14]. Also, the

chemical nature of the probe is an important factor for detecting cryptic pockets, and thus many probe molecules with different chemical properties have been used for cryptic site searches. For example, Benzene, Isopropanol, Phenol, Imidazole, Acetonitrile, and Ethylene glycol [14, 15, 15, 16]. However, none of the probe molecules have yet been reported to be absolutely effective.

So, it is still difficult to detect hotspots matching cryptic sites. Considering that structural change is key factor for cryptic site formation, hotspots should be evaluated by dynamic feature of proteins around cryptic site. Therefore, in this study, to detect cryptic pockets accurately, we proposed a method based on the combination of ligand binding hotspot detection using MSMD simulation and estimation of structural changeability of protein using topological data analysis method, DAIS (Dynamical Analysis of Interaction and Structural changes) [17] as shown in Figure 1. The DAIS made it possible to rank possibilities by capturing structural changes around cryptic sites. We performed MSMD with six probe molecules on nine proteins that have cryptic sites but cannot be detected by existing pocket search methods. From these results, ligand-binding regions, Hotspots, were identified and the accuracy of identifying cryptic sites using DAIS was evaluated. In addition, a comparison of the ability to detect cryptic sites using the PocketMiner method[18], one of the prediction methods built on machine learning, has been also performed.

■ METHODS

Choice of Protein having Cryptic pockets

Several candidate proteins having cryptic sites were selected from previous studies on cryptic site detection[11-13, 15, 18] . Since the definition of a cryptic site in existing studies is not

unique, in this study we define a cryptic site as one that clashes with the ligand molecule and not detected by pocket search. If the APO structure of the protein is superimposed on the HOLO structure and the distance between the atoms of the ligand molecule and the atoms of the APO protein residue is less than 3.5 Å, then the residue clashes with the ligand and the protein is considered to have a cryptic pocket. In addition, proteins with undetectable cryptic sites in the APO structure were selected using SiteMap, an existing pocket search method, with Schrödinger suite 2020-4 [19]. The protein targets used in this study are shown in Table 1.

Mixed-Solvent molecular dynamics simulations

To perform MSMD simulations, we used the Protein Preparation Wizard in Schrodinger suite 2020-4 [19] to complement missing loop and side chains for all target proteins. In this process followed, N- and C-termini were capped using N-methyl amide and acetyl groups, respectively. Next, after removing non-protein molecules, structural optimization was performed using the OPLS3e force field. For the six probe molecules used in this study (Benzene, Isopropanol, Phenol, Imidazole, Acetonitrile, and Ethylene glycol), RESP charges calculated at the HF/6-31G* level using the Gaussian 16 program package[20].

The protocol of MSMD simulation as below was conducted using EXPRORER_MSMD[21] [22]. To generate the MSMD system, PACKMOL 18.169[23] was used to randomly place probe molecules around each protein equilibrated in water phase environment at 0.25 M and then water molecules. The AMBER 14SB [24], GAFF2[25], and TIP3P[26] force fields were employed for the proteins, probe molecules, and water molecules, respectively. Additionally, a Lennard-Jones force field term was introduced as parameters ($\varepsilon_{LJ} = 10^{-6}$ kcal/mol; Rmin $= 20$ Å) only between the center of the probes to prevent their aggregation.

After energy minimization for above MSMD systems, constant-NVT MD simulations with harmonic position restraints on the solute heavy atoms (force constant, 10 kcal/mol/Å^2) were performed to heat gradually up to 300 K during 200 ps. Then, constant-NPT MD simulations of 800 ps at 300 K and 10^5 Pa were performed to gradually dissipate the position restrains. The P-LINCS algorithm[27] was used to constrain all bond lengths, including hydrogen atoms, and ⊿t = 2 fs was employed. In this study, 20 runs of 40 ns MSMD were performed for each system and the trajectories of the last 20 ns are used for hotspot detection. Also, the trajectories of the 5 ns (20-25 ns) are used for DAIS analysis. These simulations were performed with GROMACS 2021.5[28].

Detection of hotspots, ligand-binding regions

To extract hotspots in each MSMD simulation, we first calculated occupancy of the probe heavy atoms for each probe in a voxel. These calculations were performed based on the PMAP[21]. In this study, since the number of heavy atoms in the probe molecules used differed, normalization was performed by dividing by the number of heavy atoms in each probe molecule when calculating probe occupancy. Then, probe occupancy greater than 0.0004 were selected and clustered using the DBSCAN algorithm[29] to identify clusters (DBSCAN parameters: $\varepsilon_{DBSCAN} = 3.0$; min_samples $= 7$). The clusters here correspond to regions in the MSMD simulation where probe molecules tend to accumulate on the protein surface.

After extracting clusters for each probe, if there was more than 20% overlap between clusters, they were integrated into one hotspot. Integration of clusters consisting of multiple probe molecules is to incorporate the properties of the binding ligand into the hotspots. If more

than 50% of the voxels in a hotspot are within 3.5 Å of any atom of residues with the ligand, then we defined that the hotspot matches "cryptic site" in this study.

Protein conformational changes between in water phase and in mixed-solvent phase

We employed the DAIS method to rank possibilities of cryptic sites for hotspots. The DAIS method is based on the persistent homology method and is able to extract meaningful structural changes, however small, between the target and reference structures. In order to prepare for structural change analyses using the DAIS method, all solvent molecules were removed from structural conformations that had been fully thermally equilibrated in a mixed-solvent environment with each probe. Then, structural sampling of 5 ns was performed in a water phase environment under the same conditions as during MSMD simulation. In this conformational change analysis using DAIS, 500 snapshots of structural conformations obtained from the structural samplings in mixed-solvent phase and water phase were used.

To rank possibilities of cryptic sites for hotspots, the differences between each protein structure obtained from molecular dynamics in the each mixed-solvent and water phase environments were analyzed using the DAIS method. To assess the likelihood of the site matching hotspot being a cryptic site, we used two score values, absolute ID score and FF score. The ID score is a score calculated by the DAIS method[17], the magnitude of which corresponds to the magnitude of structural change that occurs between the reference structure and the target structure. Although this score can be positive or negative, indicating which of the compared structures is more mobile, the absolute values were used in this study to focus only on the magnitude of structural change, and that of 5 or greater were selected. In addition, a score was adopted that averaged the ID scores possessed by the residues forming the topology determined

to be structurally different by the DAIS method. The FF score is a new score established in this study, and mean a feature-point frequency. The score is a count of the number of residues around each candidate hotspot that have Cα atoms forming the topology determined to have a structural difference by the DAIS method. These scores were calculated for each MSMD calculation result and then used for ranking the cryptic site-likeness of the candidate hotspot.

■ RESULTS & DISCUSSION

Structural characteristics of the target proteins

The nine proteins harboring cryptic sites used for evaluation of our method are listed in Table 1. For all of these proteins, both the structure before (APO) and after (HOLO) binding of the ligand molecule have been reported with X-ray crystallography. In addition, considering a practical case, we selected proteins whose cryptic sites of these APO structures could not be identified using the Sitemap method[30, 31], one of the binding pocket search methods. As shown in Figure 2 and S1, the binding ligands of the HOLO state clashed with the surrounding amino acids of the APO state, which is a superposition of the APO and HOLO structures. The purple bars and pink van der Waals spheres mean the binding ligand and the colliding residues in the APO state, respectively. Residue groups observed to collide with ligand in each APO structure are also listed in Table 1. In most proteins, such as Exodeoxyribonuclease I (1FXX) and androgen receptor ligand binding domain (2AM9) in Figure 2A,B, the cryptic site is generated by changing the orientation of the side chain of residues in the ligand-binding region, with no significant visible changes in the backbone structure. On the other hand, there are proteins, such

as anti-methotrexate CDR1-4 Graft VHH (3QXW) in Figure 2C, in which regions that do not have a specific steric structure are significantly altered to produce binding regions.

Hotspot searching with mixed-Solvent molecular dynamics simulation

To detect possible cryptic site in each target protein, MSMD simulations were performed with six probe molecules (Benzene, Isopropanol, Phenol, Imidazole, Acetonitrile and Ethylene glycol). In the MSMD trajectory after sufficient equilibration, we detected hotspots based on the occupancy of probe molecule on the protein surfaces (Figure S2). As shown in Figure S2, the hotspots are observed in different locations on each protein surfaces. Hotspots were also located around the cryptic site, except for Beta-lactamase TEM (1JWP). Even though no space existed for ligand binding in APO structures and sufficient structural changes to open the cryptic sites completely were not observed during our simulation time of MSMD simulations, hotspots located around the cryptic sites. Those observation reconfirm that hotspot detection using MSMD simulations is useful for exploring the cryptic sites. Furthermore, at 1JWP, by reducing the occupancy threshold by half, candidate hotspots also appeared at the cryptic sites, as seen in Figure S4A. There is still room for argument about setting the occupancy threshold, we will only deal with the results using strict thresholds in this study.

Ranking possibilities by capturing structural changes around cryptic sites

As shown in Figure S2, the hotspots are observed in different locations on each protein surfaces. It is difficult to discriminate a hotspot matching cryptic cite among all hotspots using hotspot ranking based on probe occupancy. The essence of cryptic sites is structural change induced by ligand-binding, thus we used DAIS [17] method to pick up the structural change sites from the MD trajectories for each protein in a water phase environment and in the mixed solvent

environment. DAIS can estimate the conformational variability of proteins emerged by interaction with chemical probes. From the DAIS results, the "averaged ID score" and "FF score" were calculated for the residues located within 3.5 Å of a hotspot. In addition, for performance comparison, we used the state of the art graph neural network based method, PocketMiner[18] to predict the cryptic sites in the APO structures of the target proteins. From the prediction results, we picked up the regions consist of residues with a prediction score of 0.7 or higher, which were predicted to be the cryptic sites shown in their paper, and evaluated the likelihood of cryptic sites using the same calculation method as the FF score.

Table S1 and S2 show the results of ranking by ID score and FF score estimated from DAIS results using MSMD simulation with each probe, respectively. In the case of ranking using the ID score proposed by the conventional DAIS method, none of the probes could extract the hotspot corresponding to cryptic site with good accuracy throughout the entire protein of interest. On the other hand, in the case of using new score established in this study, FF score, ranking by DAIS calculation based on MSMD simulation with phenol showed very good results through all proteins. Indeed, the correct hotspots were extracted in ranked first in 6 out of 7 when the hotspots matching the cryptic sites were extracted. For 2AM9, the correct hotspot could not be detected at rank 1, but was picked up at rank 3. When the MSMD simulation of imidazole was used, hotspot was not mapped on the cryptic site in three proteins, 1FVR, 2AM9, and 3NX1. However, when the hotspots matching the cryptic sites were extracted, the hotspot was detected at rank 1 in 4 out of 5 cases. The ID score indicates the magnitude of the structural difference observed between the target (mixed-solvent phase) and reference structures (water phase), whereas the FF score corresponds to the number of structural changes determined to occur by the DAIS method, or the observed frequency of structural change. Ranking accuracy was higher

with the FF score than with the ID score, suggesting that the frequency with which structural changes occur is more important for detection of cryptic site than the magnitude of structural change. As shown in Table 2, the performance of the ranking with DAIS outperformed that with PocketMiner. In the case of PocketMiner, 5 out of 6 targets were ranked in the Top 3, but only 2 out of 6 targets were ranked at Rank 1. To provide a more visual understanding of the results when ranking using each score (ID score, FF score, and PocketMiner score), the Top 3 hotspots are shown in each structures (Figure 3) . Comparing PocketMiner, the ranking using the FF score could detect cryptic sites as the rank 1 hotspot in most case. In Figure S3, the projection views were shown in case of using the results of phenol MSMD simulation. In this figure, the top five of ID and FF scores were mapped on the protein structures with color gradation. In addition, each hotspot detected at rank 1 was indicated. The ID and FF scores are not absolute scores that can be compared between proteins, so they cannot be judged by color alone, but the FF score indicates that there are protein conformational changes around the ligand.

Table 3 shows relation of the probe combination consisting of the hotspot corresponding to cryptic sites and the type of probe showing best result of the DAIS analysis. As shown in above, the FF score calculated by the DAIS method using MSMD simulation for Phenol was the most accurate in many cases, however phenol was not included in all hotspots corresponding to cryptic sites. In other words, the ligand binding space of cryptic site is not necessarily created by the proximity of non-water molecules near the cryptic site, but it suggests the possibility that conformational changes are propagated like allosteric effects by the proximity of non-water molecules around the protein. The proposed method combining MSMD and DAIS methods is very effective in identifying cryptic sites for two reasons. First, because MSMD can take into account the ligand-like characters of the hotspot by using a number of probe molecules with

various chemical properties, and second, because DAIS can take into account the molecular behavior of the protein.

■ CONCLUSIONS

We proposed a new method for detection cryptic pockets, which based on the combination of detection of hotspot exhibiting ligand binding using MSMD simulation and estimation of structural changeability of DAIS. MSMD simulations using probe molecules with different chemical properties allow us to observe the occupancy time to the interaction point utilizing each molecular property. This is similar to the search for pharmacophore points using fragment-based drug discovery methods, and ligand-like characteristics can be taken into account in our hotspot which is calculated by considering the occupancy of multiple fragment. In addition, the DAIS method detects structural differences between the target and reference structures, and can extract regions where large structural changes will occur in the future with a short time simulation. Therefore, it would be very useful in predicting cryptic sites to capture the structural change that occurs when the region returns to its original stable structure in aqueous solution due to the environmental change of removing the probe molecule from the equilibrated state in a mixed solvent containing the probe molecule. Our proposed method fully utilizes these advantages, and we can assume that our method outperformed the state of the art machine learning method. Further machine learning-based forecasting methods would be developed using features obtained from this method.

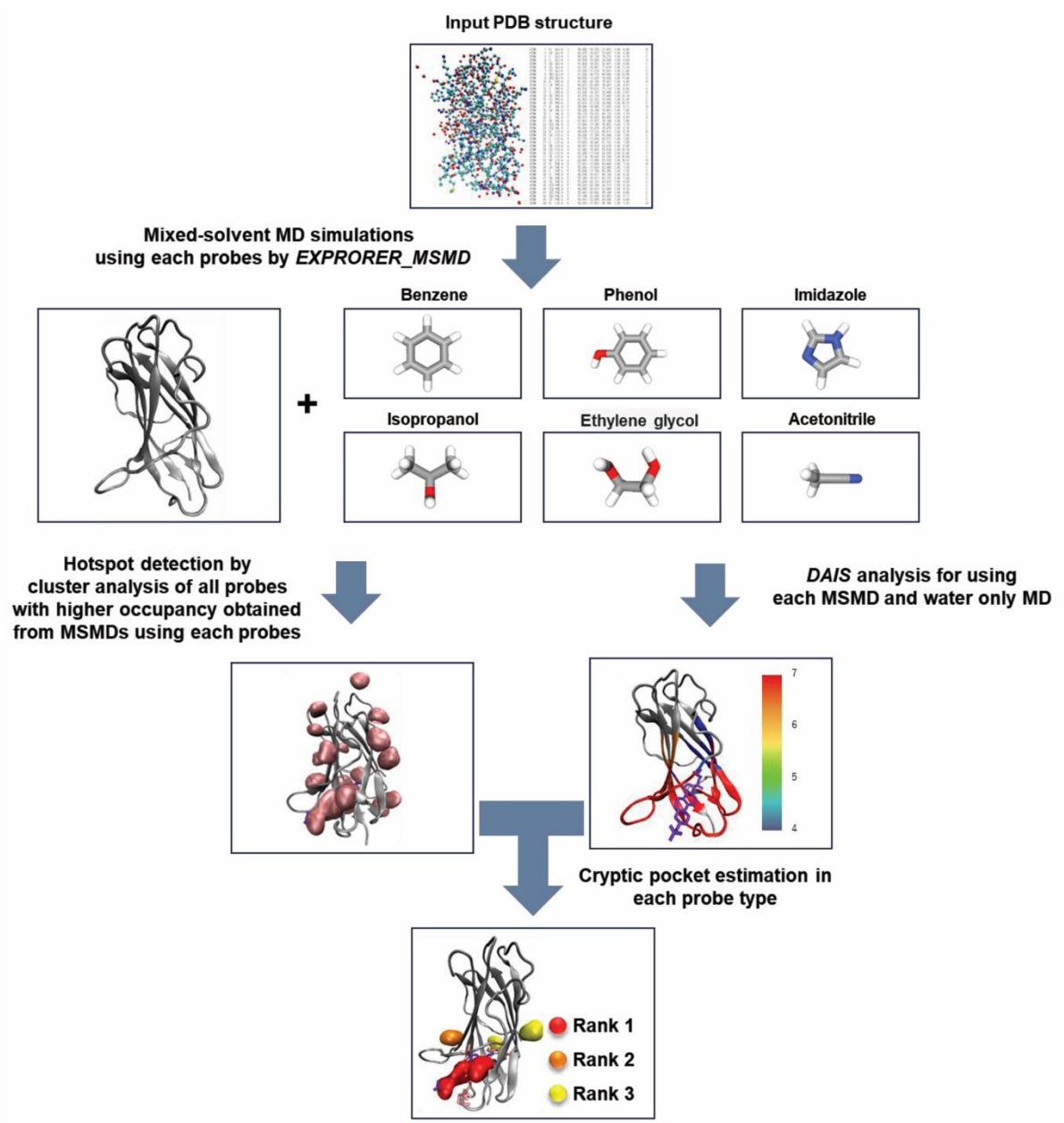

Figure 1. Schematic diagram of our cryptic site search flow with mixed-solvent molecular dynamics simulations and topological data analysis using DAIS method.

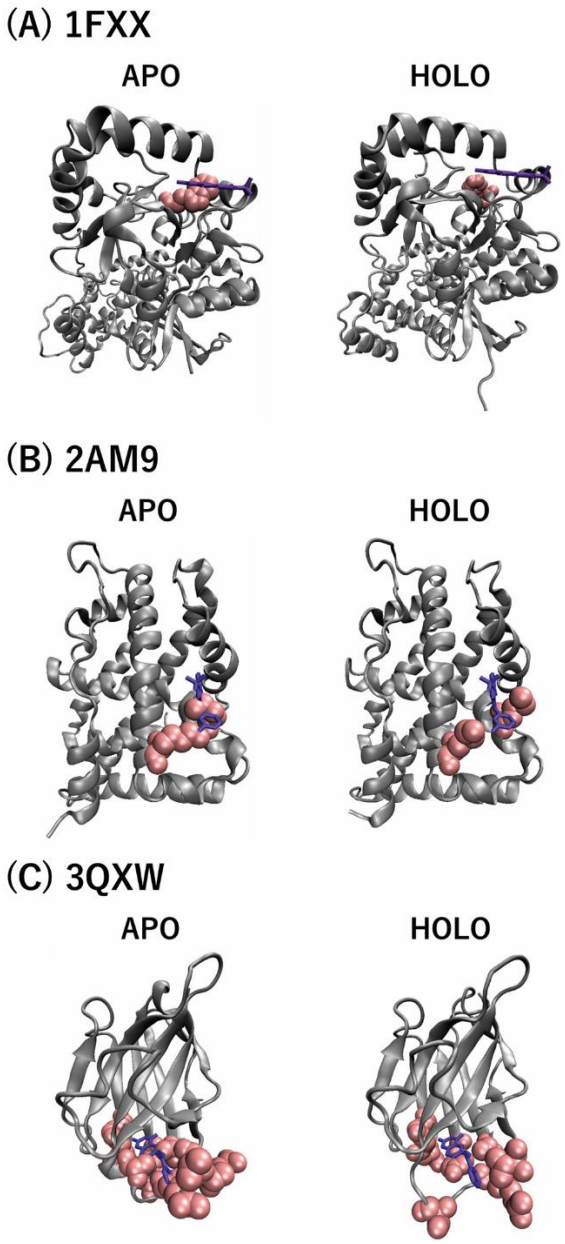

Figure 2. Representative conformational collision states between each ligand and protein at cryptic site. After superimposing the structure of the APO and HOLO, the position of ligand in HOLO were displayed on the APO, and the ligand and amino acid overlap. The purple bars and pink van der Waals spheres mean the binding ligand and the colliding residues in the APO strucuture, respectively.

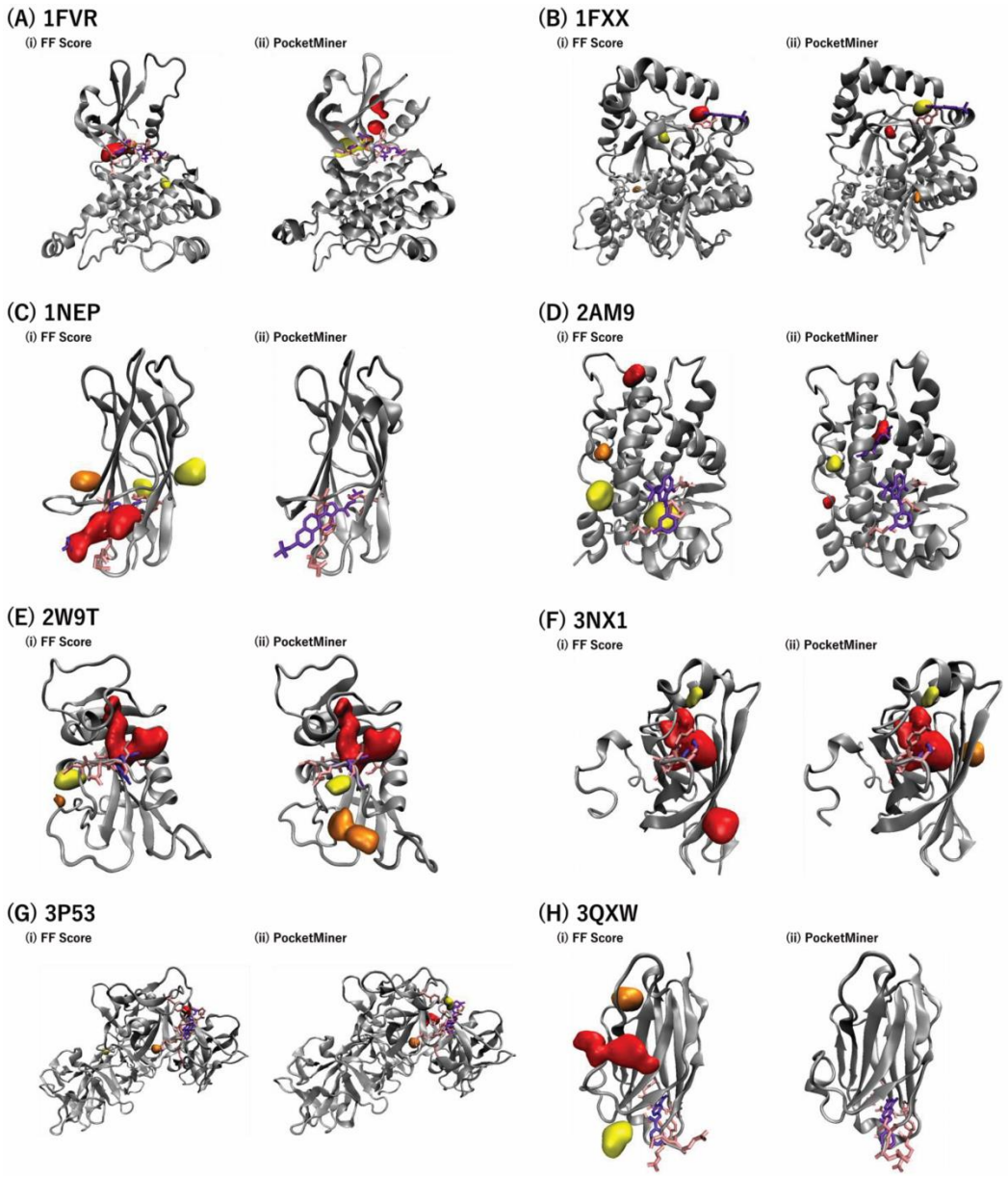

Figure 3. Ranking hotspots based on estimation of "structural changeability" of protein. Cartoon representation in silver indicates the protein backbone of the APO structure, and stick representation in purple indicates the ligand binding position of the HOLO structure. The pink stick representation shows the cryptic site residues identified as clashing with ligand. The hotspot most likely to be a cryptic site (Rank 1) is shown in red, and the Rank 2 and the Rank 3 are shown in orange and yellow, respectively.

Table 1. Proteins having cryptic sites used for evaluation in this paper. The name of each protein, the PDB IDs (APO and HOLO), and the residues that inhibit ligand binding are shown.

Protein Name	PDB ID		Clashed residues*
	APO	HOLO	
TIE-2	1FVR	2OO8	GLY831, GLU832, PHE835, GLY836, VAL838, ALA981, ASP982, PHE983, GLY984, LEU985
Exodeoxyribonuclease I	1FXX	3HL8	TRP245
Beta-lactamase TEM	1JWP	1PZO	PRO217, LEU218, LEU219, SER221, ALA222, ARG241, ASN272, ILE275, ALA276
NPC intracellular cholesterol transporter 2	1NEP	2HKA	TYR55, PHE85, LEU113, GLU118, TYR119
Androgen receptor	2AM9	2PIQ	LYS721, MET735, GLN739
Dihydrofolate reductase type 1 from Tn4003	2W9T	2W9S	ASN19, GLN20, LEU21, ILE32, PHE93
Ferulic acid decarboxylase	3NX1	3NX2	ANS23, GLY24, TRP25, TYR27
Fascin	3P53	6I11	GLN50, ILE93, VAL94, ALA95, HIS96, TRP101, GLU215, PHE216, ARG217, LYS220, TYR230
Anti-methotrexate CDR1-4 Graft VHH	3QXW	3QXV	GLN2, VAL 3, ALA25, ARG27, SER30, ARG73, ASN75, TYR78

*Clashed residue numbers are based on sequence numbers taken from UniProt[32]

Table 2. Comparison of the rankings assigned to the correct cryptic sites by the FF score calculated using the DAIS method and PocketMiner.

	DAIS (Phenol)	PocketMiner	Number of candidate sites
1FVR	1	2	/ 21
1FXX	1	3	/ 6
1NEP	1	Not Detected	/ 16
2AM9	3	11	/ 18
2W9T	1	1	/ 16
3NX1	1	1	/ 16
3P53	1	3	/ 7
3QXW	Not Detected	Not Detected	/ 12

Table 3. Using Probes that accumulate at cryptic sites for each protein and the most accurate probes for DAIS analysis.

PDB ID (APO)	Probes consisting of Hotspot matching Cryptic site	Best probe in DAIS Analysis
1FVR	Benzene, Isopropanol, Phenol, Imidazole, Acetonitrile	Phenol
	Ethylene glycol	
1FXX	Benzene, Acetonitrile	Phenol
1JWP	No trapped probe molecules around the cryptic site	
1NEP	Benzene, Isopropanol, Phenol, Imidazole, Acetonitrile	Phenol
	Benzene	
2AM9	Benzene, Isopropanol, Phenol	Phenol
2W9T	Benzene, Isopropanol, Phenol, Imidazole, Acetonitrile	Phenol
3NX1	Benzene, Isopropanol, Phenol, Imidazole, Acetonitrile, Ethylene glycol	Phenol
3P53	Imidazol	Phenol
	Acetonitrile	
3QXW	Acetonitrile	Imidazol
	Acetonitrile, Ethylene glycol	
	Acetonitrile	
	Benzene	

ASSOCIATED CONTENT

Data Availability

Initial 3D structures of the protein were downloaded from the Protein Data Bank (PDB). Schrodinger suite 2020-4 was used for protein preparation. We used AmberTools21 and Gaussian 16 Rev B.01 for probes preparation. PACKMOL 18.169 and AmberTools18 were used to prepare the MSMD system. GROMACS 2021.5 was used as the MD engine. VMD 1.9.3 was used for visualization. The protocol of MSMD simulation and DAIS in this study is available from the following link: https://github.com/keisuke-yanagisawa/exprorer_msmd and https://github.com/jkoseki/DAIS.

Supporting Information

The Supporting Information is available. Supporting_Information.docx

The collision stats between each ligand and protein at Cryptic site. (Figure S1); Hotspots obtained by MSMD simulations. (Figure S2); The projection views to compare three score values. (Figure S3); DAIS analysis results for 1JWP with loose threshold. (Figure S4); Rank order of scores obtained with DAIS methods for the hotspot that are cryptic sites. (Table S1 & S2)

AUTHOR INFORMATION

Corresponding Author

Jun Koseki

Cellular and Molecular Biotechnology Research Institute, National Institute of Advanced Industrial Science and Technology (AIST), 2-4-7 Aomi, Koto-ku, Tokyo, 135-0064, Japan

E-mail: jun.koseki@aist.go.jp

Kenichiro Imai

Cellular and Molecular Biotechnology Research Institute, National Institute of Advanced Industrial Science and Technology (AIST), 2-4-7 Aomi, Koto-ku, Tokyo, 135-0064, Japan

E-mail: kenichiro.imai@aist.go.jp

Author Contributions

JK, CM, and KI designed this research and undertook program creation and variation analyses. JK, CM, and KI discussed improvements to the research. CM and KI selected the target molecules having the cryptic sites. CM, KY and GK performed the Mixed-solvent molecular dynamics simulations. JK performed the analyses of structural changes with DAIS. JK, CM, and KI wrote the manuscript. All authors have read and approved the manuscript.

Funding Source

This work was supported by grants from KAKENHI Grants-in Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science and Technology of Japan (22H03686 to J.K; 23K28185 to K.Y and R.Y; 23K24939 to K.Y; 21H03551 to K.I;), This work was supported by AMED Research Support Project for Life Science and Drug Discovery (Basis for Supporting Innovative Drug Discovery and Life Science Research (BINDS)) (24ama121029j0002 to J.K, C.M, G.K, R.Y, T.H, and K. I; 24ama121026j0002 to K.Y).

Notes

The authors declare no competing interests.

ACKNOWLEDGMENT

Computational experiments were partly performed using the Cygnus system at CCS, University of Tsukuba.

ABBREVIATIONS

MSMD, Mixed-solvent Molecular Dynamics; DAIS, Dynamical Analysis of Interaction and Structural changes. RESP, Restrained Electrostatic Potential; GAFF2, General AMBER Force Field 2.

REFERENCES

(1) Knoverek, C. R.; Amarasinghe, G. K.; Bowman, G. R. Advanced Methods for Accessing Protein Shape-Shifting Present New Therapeutic Opportunities. *Trends Biochem Sci* **2019**, *44* (4), 351-364. DOI: 10.1016/j.tibs.2018.11.007 From NLM Medline.

(2) Vajda, S.; Beglov, D.; Wakefield, A. E.; Egbert, M.; Whitty, A. Cryptic binding sites on proteins: definition, detection, and druggability. *Curr Opin Chem Biol* **2018**, *44*, 1-8. DOI: 10.1016/j.cbpa.2018.05.003 From NLM Medline.

(3) Schueler-Furman, O.; Wodak, S. J. Computational approaches to investigating allostery. *Curr Opin Struct Biol* **2016**, *41*, 159-171. DOI: 10.1016/j.sbi.2016.06.017 From NLM Medline.

(4) Hart, K. M.; Moeder, K. E.; Ho, C. M. W.; Zimmerman, M. I.; Frederick, T. E.; Bowman, G. R. Designing small molecules to target cryptic pockets yields both positive and negative allosteric modulators. *PLoS One* **2017**, *12* (6), e0178678. DOI: 10.1371/journal.pone.0178678 From NLM Medline.

(5) Hardy, J. A.; Wells, J. A. Searching for new allosteric sites in enzymes. *Curr Opin Struct Biol* **2004**, *14* (6), 706-715. DOI: 10.1016/j.sbi.2004.10.009 From NLM Medline.

(6) Cimermancic, P.; Weinkam, P.; Rettenmaier, T. J.; Bichmann, L.; Keedy, D. A.; Woldeyes, R. A.; Schneidman-Duhovny, D.; Demerdash, O. N.; Mitchell, J. C.; Wells, J. A.; et al.

CryptoSite: Expanding the Druggable Proteome by Characterization and Prediction of Cryptic Binding Sites. *J Mol Biol* **2016**, *428* (4), 709-719. DOI: 10.1016/j.jmb.2016.01.029 From NLM Medline.
(7) Ngan, C. H.; Bohnuud, T.; Mottarella, S. E.; Beglov, D.; Villar, E. A.; Hall, D. R.; Kozakov, D.; Vajda, S. FTMAP: extended protein mapping with user-selected probe molecules. *Nucleic Acids Res* **2012**, *40* (Web Server issue), W271-275. DOI: 10.1093/nar/gks441 From NLM Medline.
(8) Capra, J. A.; Laskowski, R. A.; Thornton, J. M.; Singh, M.; Funkhouser, T. A. Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. *PLoS Comput Biol* **2009**, *5* (12), e1000585. DOI: 10.1371/journal.pcbi.1000585 From NLM Medline.
(9) Beglov, D.; Hall, D. R.; Wakefield, A. E.; Luo, L.; Allen, K. N.; Kozakov, D.; Whitty, A.; Vajda, S. Exploring the structural origins of cryptic sites on proteins. *Proc Natl Acad Sci U S A* **2018**, *115* (15), E3416-E3425. DOI: 10.1073/pnas.1711490115 From NLM Medline.
(10) Kuzmanic, A.; Bowman, G. R.; Juarez-Jimenez, J.; Michel, J.; Gervasio, F. L. Investigating Cryptic Binding Sites by Molecular Dynamics Simulations. *Acc Chem Res* **2020**, *53* (3), 654-661. DOI: 10.1021/acs.accounts.9b00613 From NLM Medline.
(11) Kimura, S. R.; Hu, H. P.; Ruvinsky, A. M.; Sherman, W.; Favia, A. D. Deciphering Cryptic Binding Sites on Proteins by Mixed-Solvent Molecular Dynamics. *J Chem Inf Model* **2017**, *57* (6), 1388-1401. DOI: 10.1021/acs.jcim.6b00623 From NLM Medline.
(12) Schmidt, D.; Boehm, M.; McClendon, C. L.; Torella, R.; Gohlke, H. Cosolvent-Enhanced Sampling and Unbiased Identification of Cryptic Pockets Suitable for Structure-Based Drug Design. *J Chem Theory Comput* **2019**, *15* (5), 3331-3343. DOI: 10.1021/acs.jctc.8b01295 From NLM Medline.
(13) Martinez-Rosell, G.; Lovera, S.; Sands, Z. A.; De Fabritiis, G. PlayMolecule CrypticScout: Predicting Protein Cryptic Sites Using Mixed-Solvent Molecular Simulations. *J Chem Inf Model* **2020**, *60* (4), 2314-2324. DOI: 10.1021/acs.jcim.9b01209 From NLM Medline.
(14) Oleinikovas, V.; Saladino, G.; Cossins, B. P.; Gervasio, F. L. Understanding Cryptic Pocket Formation in Protein Targets by Enhanced Sampling Simulations. *J Am Chem Soc* **2016**, *138* (43), 14257-14263. DOI: 10.1021/jacs.6b05425 From NLM Medline.

(15) Smith, R. D.; Carlson, H. A. Identification of Cryptic Binding Sites Using MixMD with Standard and Accelerated Molecular Dynamics. *J Chem Inf Model* **2021**, *61* (3), 1287-1299. DOI: 10.1021/acs.jcim.0c01002 From NLM Medline.
(16) Bansia, H.; Mahanta, P.; Yennawar, N. H.; Ramakumar, S. Small Glycols Discover Cryptic Pockets on Proteins for Fragment-Based Approaches. *J Chem Inf Model* **2021**, *61* (3), 1322-1333. DOI: 10.1021/acs.jcim.0c01126 From NLM Medline.
(17) Koseki, J.; Hayashi, S.; Kojima, Y.; Hirose, H.; Shimamura, T. Topological data analysis of protein structure and inter/intra-molecular interaction changes attributable to amino acid mutations. *Comput Struct Biotechnol J* **2023**, *21*, 2950-2959. DOI: 10.1016/j.csbj.2023.05.009 From NLM PubMed-not-MEDLINE.
(18) Meller, A.; Ward, M.; Borowsky, J.; Kshirsagar, M.; Lotthammer, J. M.; Oviedo, F.; Ferres, J. L.; Bowman, G. R. Predicting locations of cryptic pockets from single protein structures using the PocketMiner graph neural network. *Nat Commun* **2023**, *14* (1), 1177. DOI: 10.1038/s41467-023-36699-3 From NLM Medline.
(19) Schrödinger, L., New York, NY. Schrödinger suite 2020-4. **2020**.
(20) Frisch, M. e.; Trucks, G.; Schlegel, H. B.; Scuseria, G.; Robb, M.; Cheeseman, J.; Scalmani, G.; Barone, V.; Petersson, G.; Nakatsuji, H. Gaussian 16. Gaussian, Inc. Wallingford, CT: 2016.
(21) Yanagisawa, K.; Moriwaki, Y.; Terada, T.; Shimizu, K. EXPRORER: Rational Cosolvent Set Construction Method for Cosolvent Molecular Dynamics Using Large-Scale Computation. *J Chem Inf Model* **2021**, *61* (6), 2744-2753. DOI: 10.1021/acs.jcim.1c00134 From NLM Medline.
(22) Kudo, G.; Yanagisawa, K.; Yoshino, R.; Hirokawa, T. AAp-MSMD: Amino Acid Preference Mapping on Protein-Protein Interaction Surfaces Using Mixed-Solvent Molecular Dynamics. *J Chem Inf Model* **2023**, *63* (24), 7768-7777. DOI: 10.1021/acs.jcim.3c01677 From NLM Medline.
(23) Martinez, L.; Andrade, R.; Birgin, E. G.; Martinez, J. M. PACKMOL: a package for building initial configurations for molecular dynamics simulations. *J Comput Chem* **2009**, *30* (13), 2157-2164. DOI: 10.1002/jcc.21224 From NLM PubMed-not-MEDLINE.
(24) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J Chem Theory Comput* **2015**, *11* (8), 3696-3713. DOI: 10.1021/acs.jctc.5b00255 From NLM Medline.

(25) He, X.; Man, V. H.; Yang, W.; Lee, T. S.; Wang, J. A fast and high-quality charge model for the next generation general AMBER force field. *J Chem Phys* **2020**, *153* (11), 114502. DOI: 10.1063/5.0019056 From NLM PubMed-not-MEDLINE.
(26) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *Journal of Chemical Physics* **1983**, *79* (2), 926-935. DOI: Doi 10.1063/1.445869.
(27) Hess, B. P-LINCS: A Parallel Linear Constraint Solver for Molecular Simulation. *J Chem Theory Comput* **2008**, *4* (1), 116-122. DOI: 10.1021/ct700200b From NLM PubMed-not-MEDLINE.
(28) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2015**, *1*, 19-25.
(29) Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, 1996; Vol. 96, pp 226-231.
(30) Halgren, T. New method for fast and accurate binding-site identification and analysis. *Chem Biol Drug Des* **2007**, *69* (2), 146-148. DOI: 10.1111/j.1747-0285.2007.00483.x From NLM Medline.
(31) Halgren, T. A. Identifying and characterizing binding sites and assessing druggability. *J Chem Inf Model* **2009**, *49* (2), 377-389. DOI: 10.1021/ci800324m From NLM Medline.
(32) UniProt, C. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res* **2023**, *51* (D1), D523-D531. DOI: 10.1093/nar/gkac1052 From NLM Medline.