SUPPLEMENTARY INFORMATION

HProteome-BSite: predicted binding sites and ligands in human 3D proteome

Jiho Sim[1], Sohee Kwon[1,2], and Chaok Seok[1,2]*

[1]Department of Chemistry, Seoul National University, Seoul 08826, Republic of Korea

[2]Galux Inc, Gwanak-gu, Seoul 08738, Republic of Korea

*To whom correspondence should be addressed. Tel: +82 2 880 9197; Fax: +82 2 889 1568; Email: chaok@snu.ac.kr.

Contents of Supplementary Information

Supplementary Methods

SM 1. Method for extracting high-quality structural domains from AF structures

Low-quality regions of AlphaFold structures were removed and structural domains of higher quality were assigned as follows. First, low-quality regions of longer than 10 consecutive residues with residue-wise pLDDT < 70 were deleted, resulting in multiple fragments. Second, fragments with high inter-fragment contacts compared to intra-fragment contacts were merged. Finally, high-quality fragments were assigned as domains based on size and contact density.

In more detail, intra-fragment contact density of the *i*th fragment (d_i) is defined as (1)

$$d_i = \frac{n_i}{l_i}$$

where n_i is the number of contacts within the *i*th fragment and l_i is the length of the fragment, and inter-fragment contact density between the *i*th and the *j*th fragments (D_{ij}) is defined as (2)

$$D_{ij} = \frac{N_{ij}}{\left(l_i l_j\right)^{0.43}}$$

where N_{ij} is the number of contacts between the two fragments. The power of 0.43 is an empirical parameter chosen by the authors (2). Two sequentially separated residues are defined to be in contact if their C_β (and C_α for Glycine) atoms are within 8 Å (3). Fragments *i* and *j* were merged (2) if

$$d_i < 2D_{ij} \quad \text{and} \quad d_j < 2D_{ij}$$

A fragment was assigned as a domain if median pLDDT > 70, fragment length > 25, and intra-fragment contact density > 0.5. The fragment length cut-off of 25 for small fragments is consistent with the previous choices between 20 (4) and 30 amino acids (2).

SM 2. Method for binding site prediction and benchmark results

The following three updates were made to GalaxySite (5).

First, the ligand interaction database was updated using a more recent version of PDB (2021.07.28). The database includes small organic molecules with the number of non-hydrogen atoms in the range of 5 to 100, excluding well-known solvent molecules. Covalently bonded polymeric or peptidic ligands and ligands with metal atoms were also excluded. Structures in the mmCIF format were included in the updated version. The final interaction database includes protein-ligand interactions from 67,702 different PDB structures, which consists of 146,112 different protein chains and interactions with 24,974 different ligands.

Second, the search method for the protein-ligand complex template was updated. A structure-based search method was added to the original sequence-based template search method (5,6). The structure-based search aligns the query protein structure to the protein structures in the interaction database and scores the potential templates by TM-score (7) times the coverage of the query sequence aligned to the template. In addition, ligands bound to all protein members belonging to the cluster represented by the template selected from PDB70 were considered, while the previous version considered only the ligands bound to the representative proteins.

Finally, results for top binding sites were provided instead of top binding ligands. The binding sites, defined as the geometric centers of the binding ligands, were clustered with a 4 Å distance cutoff. Up to the top three binding sites were shown along with the information of all the ligands predicted to bound to each site. Docking was conducted only on the top scoring ligand.

Supplementary Tables

ST 1. Binding site prediction success rate on the COACH420 set. Binding site prediction with the shortest distance between the center of the predicted pocket and any ligand atom (D_{min}) within 4 Å is considered a success. N is the number of known non-metal ligands.

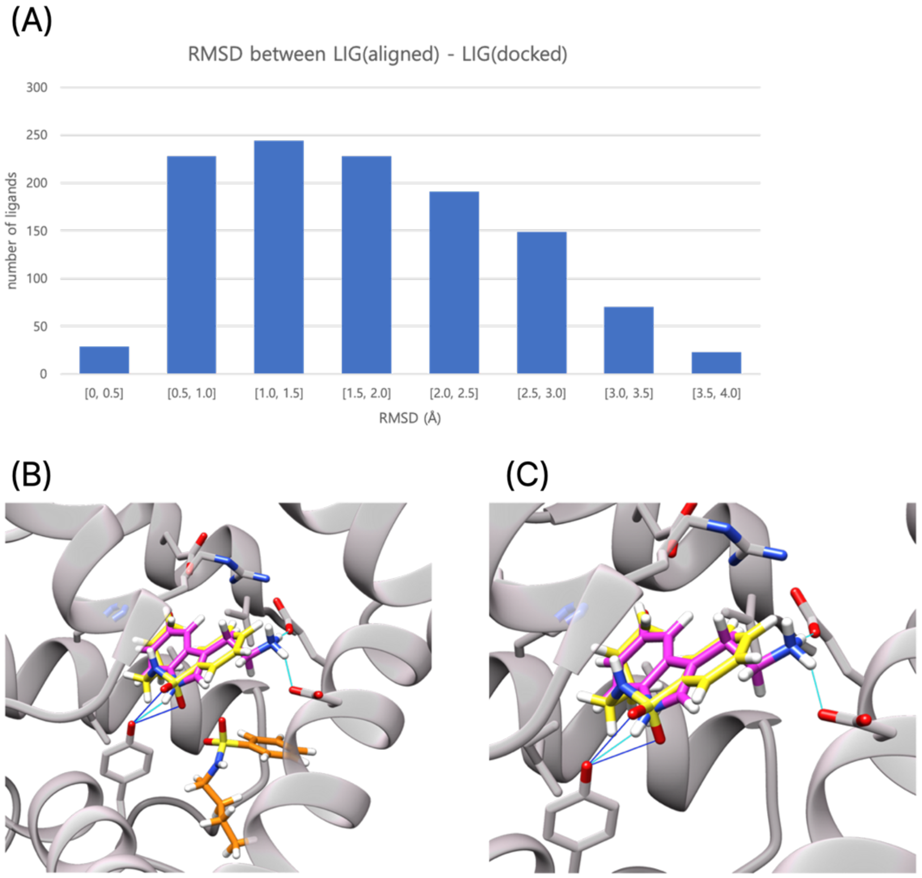
	Protein-wise success rate[1]			**Pocket-wise success rate**[2]			
	Galaxy Site-Seq	**Galaxy Site-Str**	**Com-bined**	**Galaxy Site-Seq**	**Galaxy Site-Str**	**Com-bined**	**GalaxySite-Seq (old ver)**
Top N sites	81.7 % (343/420)	83.6 % (351/420)	-	71.2 % (391/549)	71.8 % (394/549)	-	66.3 % (364/549)
All predicted sites	92.8 % (390/420)	94.3 % (396/420)	95.7 % (402/420)	84.2 % (462/549)	85.1 % (467/549)	87.6 % (481/549)	74.1 % (407/549)

1) Protein-wise success is measured for each protein target. For a protein target with multiple ligands, at least one correctly predicted ligand binding site is counted as a success.

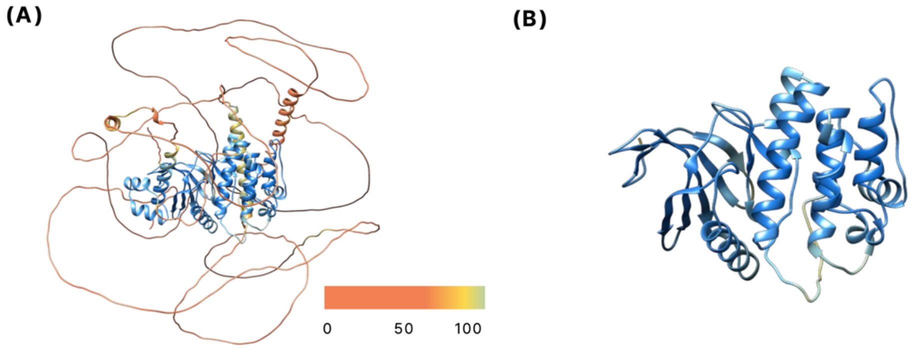
[2] Pocket-wise success is measured for each known ligand binding site (pocket) assigned by clustering.

Supplementary Figures

SF 1. (A) The distribution of RMSD between the aligned and the docked ligand is shown for cases in the COACH420 benchmark. The binding pose deviated more than 2 Å for 37% of the cases. (B) and (C) show an example in which the ligand binding pose deviates from that of the template (PDB ID: 2QEH_A). (B) Although the binding site of ligand NBB in the template crystal structure (orange) is located away from the actual binding site (ligand SRO shown in magenta), GalaxySite correctly predicted the binding site by docking (yellow). The RMSD of the docked NBB from the template NBB is 2.8 Å. (C) A consistency is observed between the hydrogen bonds between SRO and the protein (shown in dark blue) and those between docked NBB and the protein (shown in light blue).



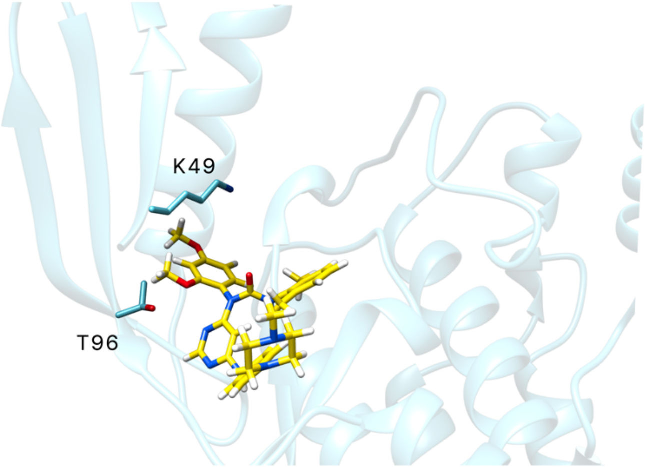
SF 2. The structure of Q9H0K1 (SIK2) predicted by AlphaFold (A) (https://alphafold.ebi.ac.uk/entry/Q9H0K1) and the corresponding structural domain deposited in HProteome-Bsite (B).



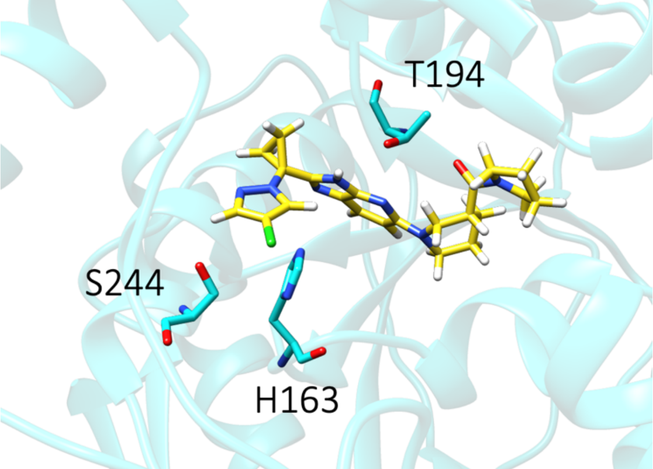
AlphaFold structure of Q9H0K1

Structural domain of Q9H0K1 [14-273]

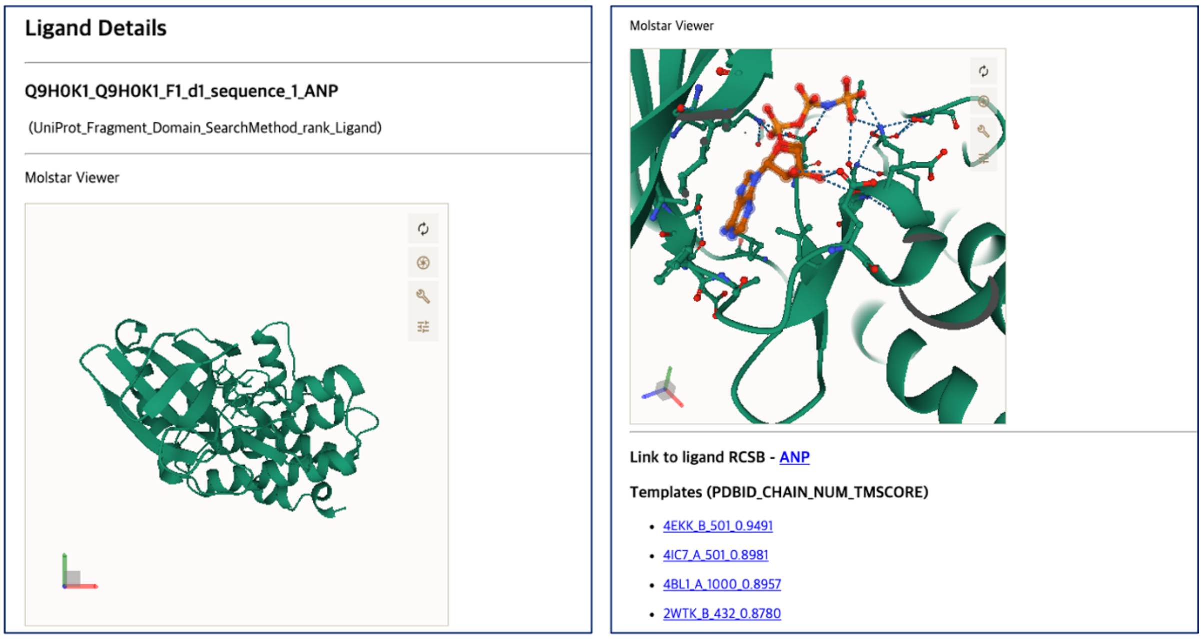
SF 3. Known inhibitor of SIK2 (HG-9-91-01) docked to the top1 ligand binding site using GalaxyDock3. Residues reported being important in SIK2 inhibitor binding such as 96 THR and 49 LYS show interactions with the inhibitor.



SF 4. Known inhibitor PF-06424439 was docked to the top 1 ligand binding site of DGAT2 found by structure-based search using GalaxyDock3, reproducing three key interactions.



SF 5. An example ligand details page (https://galaxy.seoklab.org/hproteome-bsite/database/ligands/264694). Mol* viewer allows users to examine non-covalent interactions between the ligand and nearby residues.



Ligand Details

Q9H0K1_Q9H0K1_F1_d1_sequence_1_ANP

(UniProt_Fragment_Domain_SearchMethod_rank_Ligand)

Molstar Viewer

Molstar Viewer

Link to ligand RCSB - ANP

Templates (PDBID_CHAIN_NUM_TMSCORE)

- 4EKK_B_501_0.9491
- 4IC7_A_501_0.8981
- 4BL1_A_1000_0.8957
- 2WTK_B_432_0.8780

References

1. Olmea, O. and Valencia, A. (1997) Improving contact predictions by the combination of correlated mutations and other sources of sequence information. *Fold. Des.*, **2**, S25-32.
2. Alexandrov, N. and Shindyalov, I. (2003) PDP: protein domain parser. *Bioinformatics*, **19**, 429-430.
3. Adhikari, B., Hou, J. and Cheng, J. (2018) DNCON2: improved protein contact prediction using two-level deep convolutional neural networks. *Bioinformatics*, **34**, 1466-1472.
4. Wheelan, S.J., Marchler-Bauer, A. and Bryant, S.H. (2000) Domain size distributions can predict domain boundaries. *Bioinformatics*, **16**, 613-618.
5. Heo, L., Shin, W.H., Lee, M.S. and Seok, C. (2014) GalaxySite: ligand-binding-site prediction by using molecular docking. *Nucleic Acids Res.*, **42**, W210-214.
6. Soding, J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics*, **21**, 951-960.
7. Zhang, Y. and Skolnick, J. (2004) Scoring function for automated assessment of protein structure template quality. *Proteins*, **57**, 702-710.