SUPPORTING INFORMATION

Protein-Ligand Binding Site Detection by Local Structure Alignment and Its Performance Complementarity

Hui Sun Lee and Wonpil Im

S1. Preparation of BS-ligand structure library

We downloaded the PDB files of X-ray crystallographic structures and solution NMR structures containing at least one protein and one ligand. The X-ray structures of resolution >3 Å were eliminated from the library. DNA and RNA molecules were discarded, and ligand molecules in the PDB files were identified in the heteroatom section. Heteroatoms having an identical chain ID and sequence number were grouped into one heteroatom group. If a distance of any atom pair from different heteroatom groups was 1−2 Å, the two heteroatom groups were merged into one group and identified as multipart ligands. Metal ions, water molecules, and small molecular weight additives were removed by setting the minimum number of heavy atoms in a heteroatom group to 5. Duplicated ligand molecules in a PDB file were removed except the first one. To only consider noncovalently bound ligands, if any atom in a heteroatom group was located within 2 Å from any protein atom, the heteroatom group was identified as a covalently linked ligand and removed from the library. If any atom of a residue in a protein is within 4.5 Å of its cognate ligand, the residue was defined as the BS-residue. For a ligand interacting with multiple proteins in a PDB file, all the sets of BS-residues were extracted from each protein. The final PDB structure library included 81,701 BS/ligand structure pairs (as of August 2012).

S2. G-LoSA algorithm

In G-LoSA, all combinations of Cα−Cα pairs ($\mathbf{P}_{ij}$) are first generated for two given structures, where i and j are the first and the second structures. Aiming at performing a local structure alignment centered by conserved residues and reducing

computational cost, the Cα pairs with BLOSUM62 score < 1 are discarded from $\mathbf{P}_{ij}$. Next, a product graph is generated using $\mathbf{P}_{ij}$. A vertex in the product graph corresponds to a residue pair p from $\mathbf{P}_{ij}$. Two pairs, p_1 (i_1, j_1) and p_2 (i_2, j_2), are selected from $\mathbf{P}_{ij}$, and then both distances $d(i_1, i_2)$ and $d(j_1, j_2)$ are calculated using the atomic coordinates. If $|d(i_1, i_2) - d(j_1, j_2)| < 1$ Å, $p_1(i_1, j_1)$ and $p_2(i_2, j_2)$ are assigned to product graph vertices and connected by an edge. These procedures are repeated to build a product graph for all the possible non-identical residue pairs in $\mathbf{P}_{ij}$. A maximum clique, which represents the largest set of structurally aligned residue pairs, is searched in the generated product graph using approximate coloring algorithm.[1] The two input structures are superposed by the least-squared superposition of the aligned residue sets in the identified maximum clique. The G-LoSA algorithm only requires atomic coordinates and residue names, and thus its performance does not depend on the sequence continuity and the fold similarity.

The similarity score ($S_{\text{G-LoSA}}$) is measured by the superposed structures,

$$S_{\text{G-LoSA}} = \frac{N^2}{RMSD} \tag{1}$$

where N is the number of aligned residues. If the Cα distance between a target residue and the nearest library residue is within 1 Å, the residues are assigned as an aligned residue pair. The RMSD is the root-mean-squared deviation of the aligned residue pairs and calculated using the coordinates of Cα atoms and side-chain centroids. To avoid numerical sensitivity of $S_{\text{G-LoSA}}$ with small RMSD, the RMSD value is set to 0.5 if RMSD < 0.5.

S3. fpocket algorithm

fpocket uses alpha spheres, which are spheres that contact four atoms on its boundary and contain no internal atom.[2] An ensemble of alpha spheres is filtered in order to eliminate solvent inaccessible surface, too exposed surface, and the areas of loose atom packing. Each alpha sphere is labeled according to the type of its nearby atom. After clustering remaining alpha spheres, the program ranks the clustered pockets according to their ability to bind small molecules. The scoring function for the pocket ranking was derived using partial least squares fitting to a set of pocket descriptors, such as number of alpha spheres, density of the cavity, polarity score, mean local hydrophobic density, and proportion of apolar alpha spheres.[3]

S4. Normalized scoring functions for CMCS-BSP for SET-M

$$f_{\text{G-LoSA}} = 0.41\ln(S_{\text{G-LoSA}}) - 0.77 \tag{2}$$

$$f_{\text{TM-align}} = 0.27S_{\text{TM-align}} - 0.03 \quad \text{If } S_{\text{TM-align}} > 0.95, \text{ then } f_{\text{TM-align}} = 1 \tag{3}$$

$$f_{\text{fpocket}} = \frac{0.45}{1+\exp(-0.2(S_{\text{fpocket}} - 35))} \tag{4}$$

Figure S1. Schematic representation of template identification by G-LoSA search against PDB structure library.

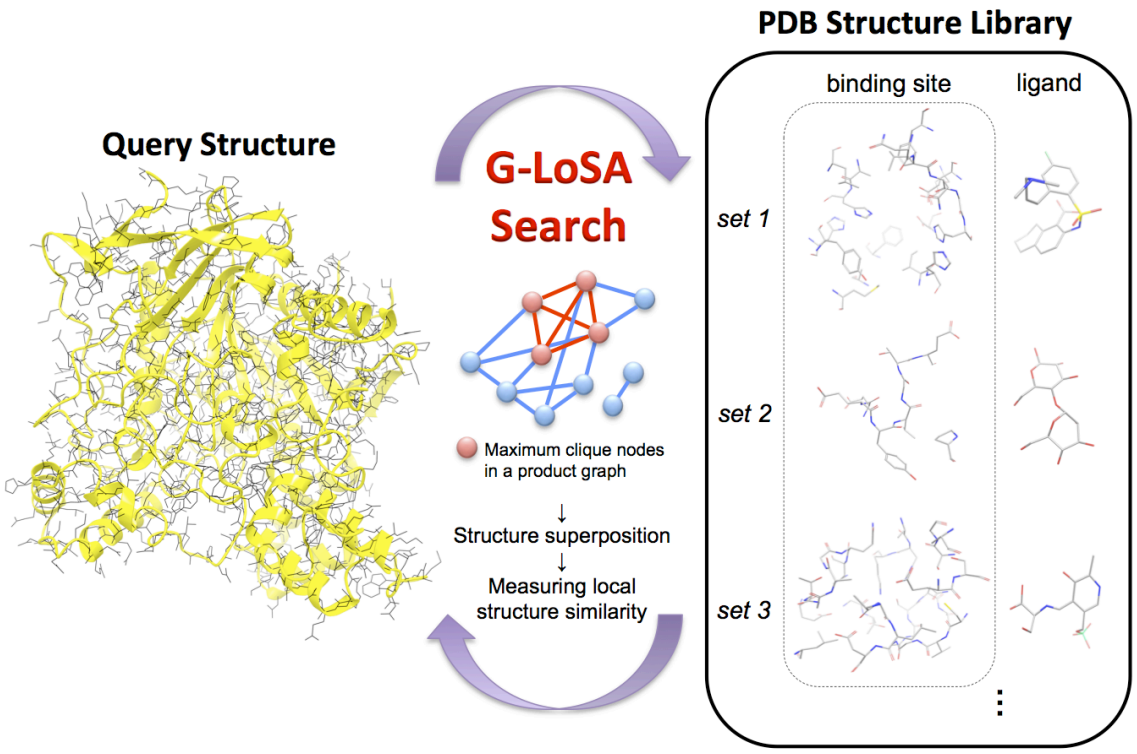

Figure S2. The plots of number of BS-residues as a function of ligand radius gyration (R_g) in SET-S and SET-M.

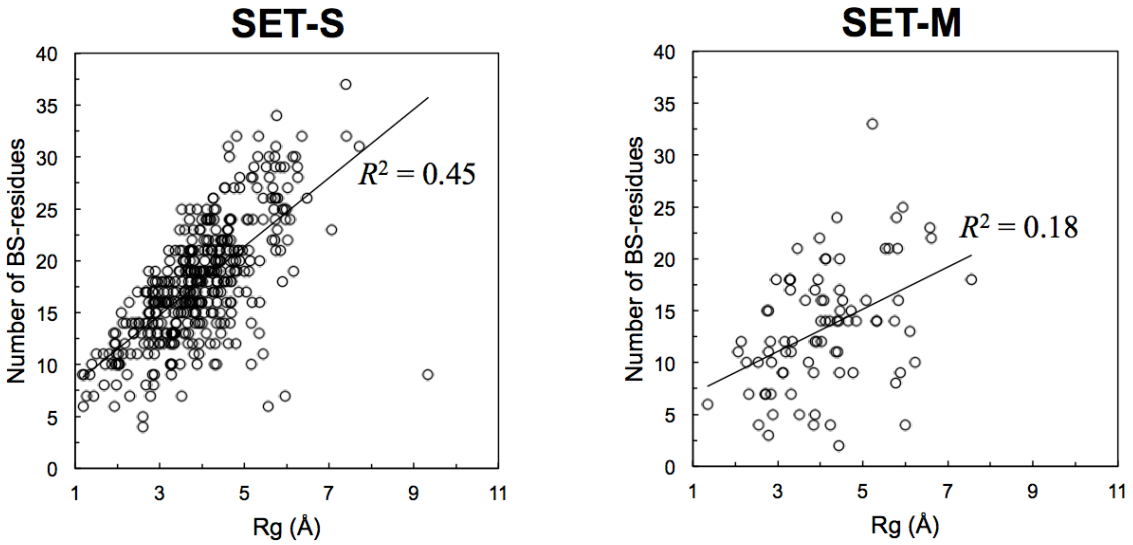

Table S1. Performance comparison of different consensus scoring approaches used in CMCS-BSP and MetaPocket for SET-S.

SET-S	CMCS-BSP		MetaPocket	
	Median BS-error (Å)	Success rate (%)	Median BS-error (Å)	Success rate (%)
TM-align + fpocket	1.85	80.3	1.89	80.3
G-LoSA + TM-align	1.66	80.0	1.71	78.3
G-LoSA + fpocket	1.70	80.5	1.74	77.6
G-LoSA + TM-align + fpocket	1.59	84.0	1.66	80.8

SET-M	CMCS-BSP		MetaPocket	
	Median BS-error (Å)	Success rate (%)	Median BS-error (Å)	Success rate (%)
TM-align + fpocket	3.04	56.6	3.76	51.8
G-LoSA + TM-align	3.92	50.6	4.01	49.4
G-LoSA + fpocket	3.78	51.8	4.28	48.2
G-LoSA + TM-align + fpocket	2.91	54.2	2.91	53.0

REFERENCES

(1) Konc, J.; Janežič, D., An improved branch and bound algorithm for the maximum clique problem. *MATCH Commun. Math. Comput. Chem.* **2007**, 58, 569-590.

(2) Liang, J.; Edelsbrunner, H.; Woodward, C., Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein Sci.* **1998**, 7, 1884-1897.

(3) Le Guilloux, V.; Schmidtke, P.; Tuffery, P., Fpocket: an open source platform for ligand pocket detection. *BMC Bioinformatics* **2009**, 10, 168.