Statistical analysis of 3D distribution of ligand atoms in Protein Data Bank using kernel density estimation

Tomohiro Sato[1]   **Hirofumi Watabnabe**[1]
tomohiro@gsc.riken.jp   hirowatanabe@ssbc.riken.jp

Teruki Honma[1]
Honma@gsc.riken.jp

[1] Center for Life Science Technologies, RIKEN, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama City, Kanagawa, 230-0045, Japan

Scoring functions have been widely used to evaluate protein-ligand interaction in virtual screening. With increasing protein-ligand information in Protein Data Bank (PDB), various empirical and knowledge-based scoring functions are reported[1]. Previous studies of empirical scoring function, such as PMF score[2], used the structural information in PDB to train distance-dependent interaction energies between protein-ligand atom pairs. In this study, we analyzed the PDB data to estimate the 3D probability density distributions of ligand atoms for every grid points around amino acid residues, rather than to assess the interaction only considering the atom pair distance (Figure 1). Using kernel density estimation (KDE)[3], the probability distributions of ligand atoms around 20 amino acids were estimated for each of the atom types defined by PATTY[4]. As the result, the p-values were calculated for every grid points, and then compared to the score by force field calculation.

[1] Huang, S. Y., Grinter, S. Z., Zou, X., Scoring functions and their evaluation methods for protein-ligand docking: recent advances and future directions, *Phys. Chem. Chem. Phys.*, 12:12899-908, 2010.
[2] Muegge, I., Martin, Y. C., A General and Fast Scoring Function for Protein−Ligand Interactions: A Simplified Potential Approach, *J. Med. Chem.*, 42:791-804, 1999.
[3] Parzen, E., On estimation of a probability density function and mode, *Ann. Math. Stat.* 33:1065-1076, 1962.
[4] Bush, L. B., Sheridan, R. P., PATTY: A programmable atom type and language for automatic classification of atoms in molecular databases, *J. Chem. Inf. Comput. Sci.*, 33:756-762, 1993.
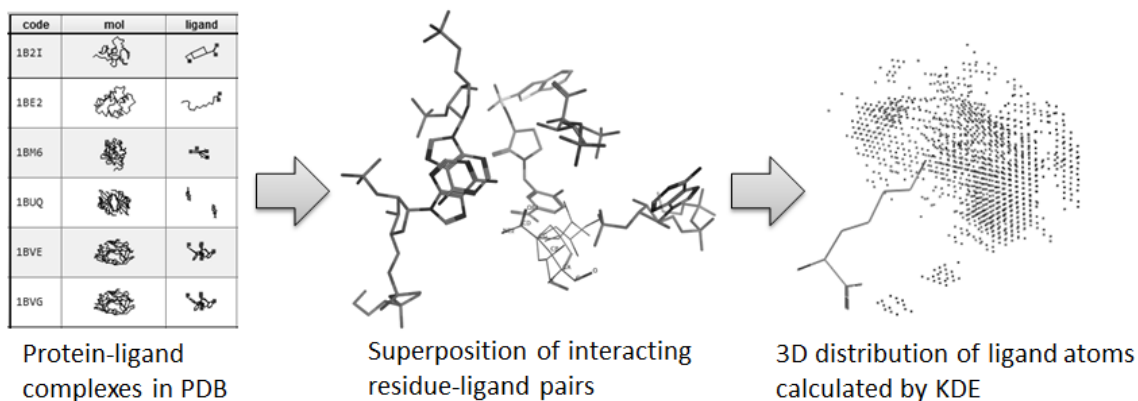
Figure 1. Brief procedure of the estimation of 3D distribution of ligand atoms.