A Library of Protein-Ligand Interaction Sites for *de novo* Ligand Design

Kiminobu Sato xsat@kuicr.kyoto-u.ac.jp

Minoru Kanehisa kanehisa@kuicr.kyoto-u.ac.jp

Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011, Japan

1 Introduction

Under the Human Genome Project, all 100,000 genes of the human genome will be sequenced along with bacterial and other genomes. This vast amount of information will become essential to discover new enzyme-substrate interactions or to design new ligands that can be used as drugs and in other applications. The elucidation of protein structures lags behind the determination of protein sequences. The homology modeling and other efficient techniques can rapidly produce approximate 3-D structures of target proteins, but they cannot readily be used for drug design since obtained protein 3-D structure models are imprecise. Combinatorial libraries of chemical compounds are useful as practical screening methods, but it is necessary to develop novel computational methods for understanding principles of molecular varieties. For small molecules that interact with proteins, such as GTP and NAD, their binding sites are usually well characterized and defined. Thus, collection and organization of local structural knowledge, including the information of ligand binding sites of proteins and atomic locations of ligands, must be precious resources that could be used computationally for drug design or in other applications. We report here a new library of ligand binding sites, which contains information of interacting atomic pairs between proteins and ligands, and other information such as indirect interaction caused by water molecules or metal ions.

2 Data and Method

2.1 Data collection

Currently, we are focusing on interactions between proteins and several ligands with hydrophobic ring groups such as AMP, ADP, ATP, GMP, GDP, GTP, NAD (NADH, NADPH) and FAD. For constructing a library of ligand binding sites, we extracted 521 entries from PDB release 84.0, each of which contains structural information of proteins that bind to these ligands. If multiple ligands bind to a protein in a PDB entry, each of the ligand binding sites was represented as a distinct entry in our library. To construct the library, we focus on a set of interactions between atoms of proteins (host) and atoms of ligands. The set of atoms in proteins that interact with ligands is defined as follows. Given a ligand that constitutes of n atoms, let h_i be a set of atoms of a host protein whose distances to the atom i of the ligand are less than or equal to $5\AA$. By definition, union of h_i, which is denoted as H, represents a whole set of host atoms that interact with atoms of the ligand. For interacting pairs of atoms, we extracted information about atomic distances, type of interactions (bonds), intercalated atoms (e.g. water molecules and metal ions such as Mn^{2+}, Mg^{2+}, etc.), and type of atoms.

2.2 Representation of atom pairs information

To characterize the interactions between a ligand atom i and a host atom j, we define bond index $Pbond = Pbond_i \cdot Pbond_j$. If atom i from the interacting pair is a constituent of an n-phosphate group, $Pbond_i = 0$. If i does not make any type of bond, $Pbond_i = 1$. If i can be a donor or an acceptor of a hydrogen bond, $Pbond_i = 2$. If i can form a salt link, $Pbond_i = 3$. For atom j constituting the host protein, $Pbond_j = 0$ if j is a main chain atom. If j is an atom of a hydrophobic amino acid residue, $Pbond_j = 1$. If j can be a donor or an acceptor of hydrogen bond, or can form a salt link, $Pbond_j = 2$. If j is a charged or polar atom, $Pbond_j = 3$. Furthermore, the ligand atom i is discriminated by an index gp whether it is a constituent of a ring group or not. If i is a ring group atom, $gp_i = 1$, otherwise $gp_i = 0$. According to these definitions, we define the scoring function of interacting atomic pairs by the following equation:

$$P_{score} = \frac{P_{bond}}{P_{distance}} \cdot P_{pena}$$

where $P_{distance}$ is the distance between two atoms. If the ligand atom i is charged and the host atom j is hydrophobic, non-polar or uncharged, or if i is uncharged and a ring group constituent ($gp_i = 1$),

$$P_{pena} = \frac{1}{P_{distance}}.$$

If both i and j have the same charge,

$$P_{pena} = -\frac{P_{bond}}{P_{distance}}.$$

If i and j have opposite charges, or if i is uncharged and $gp_i = 0$,

$$P_{pena} = \frac{P_{bond}}{P_{distance}}.$$

In addition to the scoring for atomic pairs, intercalated water molecules or metal ions are dealt separately to discriminate them from host-ligand atomic pairs.

3 Result and Discussion

For all possible interacting host-ligand atomic pairs in the PDB entries, P_{score} was calculated by the scoring function defined above. Since the calculated P_{score} would reflect some degree of possibility of bond formation, it is stored in our library of ligand binding sites with additional bond information extracted from description in PDB. The information of bond formation ability could be used to discover novel ligands and ligand-host interactions. The LIGAND database contains information of enzymatic reactions, metabolic ligands and chemical reactions that appear in the KEGG PATHWAY database [1]. The constructed library for ligand binding sites contains more information concerned with atomic pairs and circumstances of the binding pockets. Thus the information may be regarded as supplementary information of LIGAND. Currently, the library is a flat text-based file providing matrices, which store P_{score} in their elements. The matrix based representation of ligand binding information would be effective for various applications, such as *de novo* ligand design, since it is easy to calculate for various purposes. On the other hand, we also devised an approach to visualize it for efficient and intuitive understanding of ligand-protein interactions.

Acknowledgments

We especially acknowledge Dr. Hiroyuki Ogata for precious advice. This work was supported in part by the Grant-in-Aid for Scientific Research on the Priority Area ‘Genome Science’ from the Ministry of Education, Science, Sports and Culture in Japan. The computation time was provided by the Supercomputer Laboratory, Institute for Chemical Research, Kyoto University.

References

[1] Goto, S., Nishioka, T., and Kanehisa, M., LIGAND: Chemical Database for Enzyme Reactions, *Bioinformatics*, 14(7):591–599, 1998.