Unit-Based *de novo* Drug Design Using a Library of Protein-Ligand Interaction Sites

Kiminobu X Sato xsat@kuicr.kyoto-u.ac.jp
Susumu Goto goto@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, a rough draft of the human genome will be sequenced in 2000, and almost all in 2003. The next step is to obtain functions of the genes in the genome. Structural genomics and pharmacogenomics are among the promising approaches to elucidate them. A gene function is regarded as the protein function and thus the atomic configurations in functional sites of the protein. The functional sites usually include specific binding sites of ligands, especially small molecules. Therefore the collection and organization of local structural knowledge, including atomic locations of ligands and interacting sites of proteins and ligands, must be a precious resource for computational drug design. We have constructed a library of protein-ligand interaction sites and a scoring system for compounds [1]. We report here an improvement of the library, which now contains the information on substructures and functional groups (functional units) of ligand. We also report the improved scoring scheme. We believe this scoring system can be applied to the drug design by combining the predefined functional units.

2 Data and Method

Extraction of interaction sites from PDB We extracted all inhibitors and effectors from PDB (Release 1999/09/14, 10,755 entries) [2], such as HIV protease inhibitor, thrombin inhibitor, nucleotides, NAD (NADH,NADP), flavin adenine-nucleotides, cyclic nucleotides and antibiotics. Among them we excluded the ligands whose total interaction sites are less than 10. When a protein includes several interaction sites, we consider them independently. The number of extracted interaction sites is 4,041 and the number of ligands in the interaction sites is 65 among total 1,823 ligands in PDB.

Distribution of atoms in interaction sites To calculate a distribution of atoms in interaction sites, we used PDB atom descriptor for describing atom types. Because the PDB atom descriptors distinguish amino acid constituting atoms, we can specify the amino acid type from the descriptor.

In each protein-ligand interaction site, we calculate the atomic distribution as follows. For each ligand atom, we select the protein atom within 3 $\mathring{A}$ from the ligand atom and store the distance between them. If the same type of protein atom appears more than once in the interaction site, the only nearest atom is selected in the sense of the standardization of atomic specificity because same atom often appears in a site and simple summation of the number of atoms is biased.

Next, for each atom i in each ligand molecule, we calculate the specificity of a protein atom j using the following energy term between i and j: $E_{ij} = E_{VDW\,ij} + E_{eij} + E_{Hij}$ where $E_{VDW_{ij}}$,E_{eij} and E_{Hij} are van der waals, electrostatic and hydrogen bond energies between i and j, respectively, and calculated by using the distance above. The occurrence probability is approximated as a general probability distribution to clarify and give a score of atomic pair specificity for each compound.

In the probability distribution of atom j occurrence is observed for each ligand constituent atom i, the specific atom distribution in the site is not so high but essential as the protein function compared with the distribution of weak functional atoms such as back bone constituent which appears in all amino acid residues.

Unit definition In many drug molecules, the scaffold is similar with each other and variations are derived by functional groups attached to the scaffold. We defined a unit as a substructure of the scaffold by the following heuristics.

- Compounds including small or single ring structure are defined as units. Thus adenosine, guanosine, n-phosphate group, ribose, metal as heme, nicotine are defined as units.
- If the compound contains a complex ring structure, for example, hormone type scaffold like estradiol, heme ring, HIV protease inhibitor, flavin, functional groups and other branches connected to each ring structure are regarded as different units. The complex ring itself is divided in several units by considering its biosynthesis pathway.
- Peptide mimic compounds are divided into main chain part and branched part.
- Sulfur, phosphate and other groups included into the main chain part are considered as units.

Functional groups such as NH3, SH3, carbonyl group, hydroxyl group and methyl group are also treated as units in actual computation.

3 Results and Discussion

We defined over 200 units from the 65 ligands and calculated their energy terms by $E_u = \sum_i^U E_{ij}$ for each protein atom j, where U is the number of unit constituents.

In this paper we considered distribution of the pairwise potential function. However, the distribution strongly depends on structural data although the deviation of atomic pair frequency is considered for each interaction site by our accumulation method of atomic pairs. This distributed atomic pairs should be judged in the view of energies of interactions. For this reason, interaction energy terms should be included in our scoring scheme. Following the calculation of the potential score set for units and compounds, the structure-activity relationship (SAR) will be considered for the selection of pharmacophore candidates.

Acknowledgments

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 Supercomputer Laboratory, Institute for Chemical Research, Kyoto University.

References

[1] Sato, K. and Kanehisa, M. A library of protein-ligand interaction sites for *de novo* ligand design *Genome Informatics 1998*, 9:298–299, 1998.

[2] Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer, E.E.Jr, Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., and Tasumi, M., The Protein Data Bank: a computer-based archival file for macromolecular structures *J. Mol. Biol.*, 112:535–542, 1997.