論文の内容の要旨

論文題目

An exhaustive analysis of protein-ligand binding sites with a fast neighbor search method（高速近傍探索手法を用いたタンパク質リガンド結合部位の網羅的解析）

氏　　名　　　伊東　純一

Computational investigation of protein functions is one of the most important task in the field of structural bioinformatics. In many cases, proteins exhibit their biological functions by interacting with other molecules so-called ligands, and thus ligand-binding sites can be regarded as functional units of the proteins. Interestingly, a common ligand-binding site can be conserved between different proteins whose sequences or folds are totally different. Therefore the comparison of protein-ligand binding sites, not protein global structures, is an appropriate approach to gaining functional and evolutionary knowledge about proteins. According to the progress of structural genomics projects, hundreds of thousands of protein-ligand binding interfaces are observed in Protein Data Bank (PDB). In addition to them, vast amounts of potential ligand-binding sites are also available by using various kinds of binding site prediction tools. Performing an exhaustive similarity search for such vast numbers of protein binding sites should provide the basis for automatic classification of protein functions. Moreover, such a systematic understanding of protein–ligand interactions can be exploited for structure-based drug design. However, the existing 3D alignment based methods can be applied only to a limited data set mainly due to the time complexity, and thus is not scalable to flood of structural data.

In this thesis, we present a fast and efficient method for enumerating similar pairs of binding sites, which is scalable to millions of binding sites. In the proposed method, binding sites are mapped onto feature space based on their

geometrical and physicochemical properties first, and then similar pairs are enumerated by a fast neighbor search algorithm called SketchSort. A crucial point of our method is that the similar pairs in the feature space are detected by sorting operation that can be performed as approximately O(n), where n represents the number of binding sites. It is much faster than a brute-force pairwise comparison whose time complexity is O(n^2) in case n is large. We showed that our method is over 100 times faster than FuzCav, a state-of-art binding sites similarity search method. We also evaluated the performance of our method from the viewpoint of accuracy. We performed two types of benchmark tests, in each of which the ability to recognize biologically related binding sites was measured using our method and FuzCav. In both tests, our results outperformed FuzCav in terms of sensitivity/specificity. We further conducted an additional test to check the ability for discriminating binding sites of the same ligand from the others. The result showed that our method is comparable or more accurate than an accurate 3D alignment program SiteEngine. These benchmarking results indicate that our method provides not only high-throughput, but also reliability for detecting biologically relevant binding sites in comparison to the existing methods.

Then, to demonstrate the performance and scalability with our method, We applied it to all-pair similarity searches for 1.8 million known and potential ligand-binding sites. The execution time to enumerate all similar pairs was within 4 days on a standard desktop machine (Intel Xeon 2.93 GHz). Consequently, we discovered over 11 million pairs of similar binding sites including several notable analogous sites, such as a similar nucleotide-binding site between different protein families or a similar calcium-binding site between distinct protein folds. It is the largest-scale study of binding site comparison for the PDB entries, as far as we know.

We further compiled the all detected pairs into a new database called Pocket Similarity Search using Multiple-sketches (PoSSuM), which is freely available for all researchers (http://possum.cbrc.jp/PoSSuM/). Since similar binding sites have already been enumerated and stored in our database, users can retrieve them rapidly, within a few seconds, through our web interface. Because all binding sites were annotated with information of various types such as CATH, SCOP, EC number and Gene Ontology, users can easily explore similar binding sites between proteins with different folds or similar catalytic sites between enzymes with different EC numbers. In comparison with an existing well known database, SitesBase, which includes approximately 33,000 known ligand-binding sites, our new database stores a much larger number of up-to-date known binding sites deposited in the PDB. In addition to them, our database includes pairs between known and potential ligand-binding regions predicted using a novel pocket detection program.

Our fast method and database are expected to be useful for annotation of protein functions and rapid screening of target proteins in drug design. In the near future, We are planning to extend our dataset to binding interfaces of proteins to proteins and to nucleic acids. Performing such a comprehensive search might engender identification of overlap regions of a protein and a small molecule; knowledge of such regions is expected to be useful for developing inhibitors for protein–protein interaction.