Abstract

The number of experimentally resolved protein structures in the Protein Data Bank has been growing fast in the last 20 years, which motivates the development of many computational tools for protein-ligand binding sites prediction. Binding sites prediction from protein 3D structure has many important applications; it is an essential step in the complex process of rational drug design, it helps to infer the side-effects of drugs, it provides insight into proteins biological functions and it is helpful in many other fields, such as protein-ligand docking and molecular dynamics. As far as we know, there has not been a study that would systematically investigate general properties of known ligand binding sites on a large scale. In this thesis, we examine these properties using existing experimental and predicted residue-level annotations of protein sequence and structure. We present an automated pipeline for statistical analysis of these annotations, based on hypothesis testing and effect size estimation. It is implemented in Python and it is easily extensible by user-defined annotations. The usage is demonstrated on 33 existing annotations and 4 different datasets. The practical significance of the results is tested with P2Rank prediction method. We hope that the results as well as the pipeline could be eventually helpful for improving the performance of the existing binding sites predictors.