PPI3D-clusters: non-redundant datasets of protein-protein, protein-peptide and protein-nucleic acid complexes, interaction interfaces and binding sites

Justas Dapkūnas, Kliment Olechnovič, Česlovas Venclovas
Institute of Biotechnology, Life Sciences Center, Vilnius University, Lithuania

To accomplish their functions in living organisms, proteins usually interact with various biological macromolecules, including other proteins and nucleic acids. Despite the recent progress in protein structure prediction, only a part of these interactions can be predicted accurately, and interactions involving nucleic acids are especially hard to model. As a result, new computational methods to analyze and predict protein interactions are very welcome. The development of these methods largely depends on the availability of reliable data that could be used both to understand the basic mechanisms of biomolecular interactions and to train the machine learning models. However, the experimental data on protein interactions in the Protein Data Bank (PDB) are redundant, noisy and hard to interpret. To facilitate the analysis of the available data, we have developed the PPI3D-clusters database that contains non-redundant datasets of protein complexes, interaction interfaces and binding sites. The structures are clustered according to both protein sequence and structure similarity, allowing to retain the alternative interaction modes. All protein-protein, protein-peptide and protein-nucleic acid interaction interfaces and binding sites are pre-analyzed by means of Voronoi tessellation. The data are updated every week to keep in sync with the PDB. The users can query the data according to different criteria, select the interactions of their interest, download the desired data subsets in tabular format and as coordinate files, and use them for detailed investigation of protein interactions or for training the machine learning models. We hope that the newly developed PPI3D-clusters database will become a useful resource for researchers working on diverse problems related to biomolecular interactions.