Statistical and Machine Learning in Protein Structure Studies: data mining towards integrative structural biology

Haiguang Liu*, Xilun Xiang, Siyuan Liu, Xiaoqun Dong, Xiang Gao, Hao He, Xuanxuan Li

Complex Systems Division
Beijing Computational Science Research Center, Beijing 100193, China

The comprehensive understanding of protein structure, dynamics and functions requires the synergy of experimental and computational approaches. Using computational modeling to incorporate multiple sources of experimental information to investigate the molecular mechanism will produce the best knowledge. We are putting a lot efforts together in the development of computational platform and tools to interpret the experimental data from crystallography diffraction, solution scattering, CryoEM single particle imaging. Meanwhile, there are invaluable information in the structure databases, such as the protein data bank or PISA. Here, I hope to present the preliminary results in the data mining and applications in the following topics:

1. Engineering site prediction for disulfide bonds in proteins using machine learning methods;
2. Neighborhood preferences of amino acids in 3D structures and the application in structure assessment;
3. 3D structure reconstruction from X-ray scattering data using auto-encoder neural network models;
4. Statistical analysis on structures of flexible domains.