[L9] Computational Methods for Knowledge Extraction from Large Protein Structure Collections

Matthias Rarey

Center for Bioinformatics (ZBH), University of Hamburg, Bundesstr, 43, 20146 Hamburg, Germany

Structure-based drug design starts with the collection, preparation, and initial analysis of protein structures. With more than 115,000 structures publically available in the Protein Data Bank, fully automated processes reliably performing these important preprocessing steps are needed. Several tools are available for these tasks, however, most of them do not address the special needs of scientists interested in protein-ligand interactions.

In this lecture, we summarize our research activities towards an automated processing pipeline from raw PDB data towards ready-to-use protein binding site ensembles. Starting from a single protein structure, the pipeline covers the following phases: Extracting structurally related binding sites from the PDB, aligning disconnected binding site sequences, resolving tautomeric forms and protonation, orienting hydrogens and flippable side-chains, structurally aligning the multitude of binding sites, and performing a reasonable reduction of ensemble structures. The pipeline, named SIENA, creates protein-structural ensembles for the analysis of protein flexibility, molecular design efforts like docking or de novo design within seconds. For the first time, we are able to process the whole PDB in order to create a large collection of protein binding site ensembles.

SIENA is available as part of the ZBH ProteinsPlus webserver under http://proteinsplus.zbh.uni-hamburg.de.