Compressive Structural Bioinformatics: Large-Scale Analysis and Visualization of the PDB Archive

Peter Rose, University of California, San Diego; *Anthony R. Bradley, University of California, San Diego; Alexander S. Rose, University of California, San Diego; Yana Valasatava, University of California, San Diego; Jose M. Duarte, University of California, San Diego; Andreas Prlić, University of California, San Diego*

We are developing compressed 3D molecular data representations and workflows ("Compressive Structural Bioinformatics") to speed up visualization and mining of 3D structural data by one or more orders of magnitude. Our data representations allow scanning and analyzing the entire PDB archive in minutes or visualizing structures with millions of atoms.

Compact data representation - Existing text-based file formats for macromolecular data (PDBx/mmCIF) are slow to parse. To address this bottleneck, we have developed the Macromolecular Transmission Format (MMTF) (http://mmtf.rcsb.org/) that offers 75% compression over the standard mmCIF format and is over an order of magnitude faster to parse. MMTF files are easy to decode using our open source libraries, and hence the format has been rapidly adopted by the community, including the following 3D viewers: 3Dmol.js, iCn3D (NCBI), ICM, Jmol, PyMol, and bioinformatics libraries: BioJava, and Biopython.

High-performance web-based visualization - The small individual file size and high parsing speed enables high performance web-based visualizations. We have seen a greater than 20x speedup over mmCIF in loading of PDB entries from sites across the USA, Europe, and Asia. Using MMTF and NGL, a highly memory-efficient WebGL-based viewer, even the largest structures in PDB can be visualized on a smartphone.

High-performance distributed parallel workflows - MMTF enables scalable Big Data analysis of 3D macromolecular structures using distributed, in-memory, parallel processing frameworks such as Apache Spark. For example, we can extract all ligands from the PDB using MMTF with Apache Spark in about 3 minutes. In contrast, using the mmCIF format, the same task takes several hours.

The MMTF file format enables a paradigm change for structural bioinformatics applications. It is now possible to store the entire PDB in memory to eliminate I/O bottlenecks, to rapidly visualize large structures over the web, and to perform distributed parallel processing.