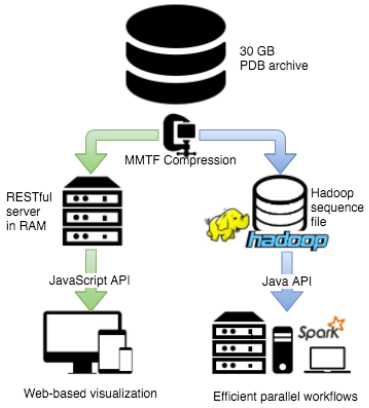
PDB on steroids – compressive structural bioinformatics

Peter W. Rose[1,2], Anthony R. Bradley[1,2], Alexander S. Rose[2], Yana Valasatava[2], Jose M. Duarte[1,2], Andreas Prlić[1,2]

[1] RCSB Protein Data Bank, San Diego Supercomputer Center, UC San Diego, peter.rose@rcsb.org
[2] Structural Bioinformatics Laboratory, San Diego Supercomputer Center, UC San Diego

Project Website: http://mmtf.rcsb.org/
Source Code: https://github.com/rcsb/mmtf
License: Apache 2 (Java, Python API), MIT (JavaScript API)



Abstract

We are developing compressed 3D molecular data representations and workflows ("Compressive Structural Bioinformatics") to speed up mining and visualization 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 in a web browser on a smart phone.

Compact and self-contained data representation - Existing text-based file formats for macromolecular data are slow to parse, are not easily extensible, and do not contain certain key data (e.g., all bonding information). For these reasons we have developed the **M**acro**m**olecular **T**ransmission **F**ormat (MMTF) (http://mmtf.rcsb.org/). MMTF has three core benefits over existing file formats. First, through custom compression methods, the entire Protein Data Bank (PDB) archive can be stored in 7GB. This enables fast network transfer for visualization and in-memory processing of the entire PDB. Second, MMTF data are serialized into MessagePack (http://msgpack.org), a compact, extensible and efficient format, similar to JSON, but binary for faster parsing. Third, MMTF is user friendly, extensible and contains information not found in current formats. In this work we show that MMTF enables high-performance visualization and scalable structural analysis of the PDB archive.

High-performance web-based visualization - The MMTF files are served directly from RAM using a RESTful service. This low latency service, combined with the reduced individual file size and the increased parsing speed of the binary format facilitates high performance web-based visualizations. Specifically we have seen a greater than 20x speedup over mmCIF in loading of PDB entries from sites across the USA, Europe, and Asia. Using the MMTF JavaScript API and NGL, a highly memory-efficient WebGL-based viewer (https://github.com/arose/ngl), even the largest structures in PDB can be visualized on a smart phone.

High-performance distributed parallel workflows - The order of magnitude increase in parsing speed enables scalable Big Data analysis of 3D macromolecular structures. A Hadoop sequence file (binary flat file of key value pairs, optimized for parallel sequential access) of MMTF data is released and updated weekly for the entire PDB archive. Distributed, parallel processing is then possible from this file, using Big Data frameworks such as Apache Spark (http:spark.apache.org/). As an example, we have used this file for ligand extraction. We extracted all ligands from the PDB using the MMTF Hadoop file with Apache Spark in about 3 minutes. In contrast, using mmCIF files as input, the same task took 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 trivially perform distributed parallel processing on laptops, desktops, and compute clusters.

This project was supported by the National Cancer Institute of the NIH's Big Data to Knowledge initiative (BD2K) under award number U01 CA198942.