Compressive Structural BioInformatics: High Efficiency 3D Structure Compression

UNIVERSITY OF CALIFORNIA SAN DIEGO

PI: ROSE, PETER W

Grant Number: 1 U01 CA198942-01

The Protein Data Bank (PDB) archive has doubled in size since 2008 and exceeded 100,000 entries in 2014. At the same time, the size and complexity of structures are increasing dramatically, for example the recently determined structure of the HIV-capsid contains about 2.5 million atoms. The emerging techniques of integrative Structural Biology are starting to determine structures of molecular machines in the mega-Dalton range by combining cryo-Electron Microscopy, Small-Angle X-ray Scattering, X-ray, and NMR at increasingly higher resolution. Interactive visualization of large complexes exceeds available network bandwidth and memory of typical scientists' desktops, laptops, or mobile devices. Large-scale structural analyses and queries of the archive have become a Big Data challenge. To make these structures accessible to all scientists, educators, and students, new ways of representing these data are required. In domains such as high-definition television, satellite communication, video or audio streaming, high-efficiency compression has been key to deliver interactive media to phones, tablets, laptops, and desktops. A similar trend has emerged in the handling of whole genome sequence data. An entire discipline ""Compressive Genomics"" has been developed to deal with data compression and development of algorithms to process these data. This proposal introduces the concept of ""Compressive Structural Bioinformatics"", a set of compression algorithms, applications, and workflows that analyze and visualize large structures and large sets of structures at an unprecedented speed (100-1000 fold speedup) and with minimal client side overhead. The aims of this project are: 1. Develop a compact and extensible representation of 3-D biomolecular structures, 2. Enable interactive visualization of large complexes by reducing network bandwidth and enabling data streaming, 3. Enable large-scale analyses of the PDB archive for I/O bound workflows, and 4. Develop open source software libraries. Through collaboration with developers of widely used visualization applications and distributed data-parallel workflow systems, the new techniques will be implemented, benchmarked, and reference implementations will be provided in several programming languages for easy adoption. It is expect that these new ""Compressive Structural Bioinformatics"" tools will enable transformative research as intended by the NIH's Big Data to Knowledge initiative. PUBLIC HEALTH RELEVANCE PUBLIC HEALTH RELEVANCE: The 3-D structures (shapes) of proteins and nucleic acids, the building blocks of life, are fundamental to the understanding of disease processes, the mechanism of drug actions, and the development of new medicines. We develop data compression and streaming techniques for large 3-D structures, similar to what YouTube does for videos, to enable access, large-scale analysis, and interactive visualization of very large biomolecules by scientists, educators, students, and educators.