GROMACS - Task #2035

A common trajectory analysis data exchange format

08/18/2016 11:38 AM - Christian Blau

Status:	New
Priority:	Normal
Assignee:	
Category:	analysis tools
Target version:	future
Difficulty:	uncategorized

Description

Trajectory analysis tools to date lack a data exchange format for structured data.

A common sharing format for structured analysis result data shall simplify

- splitting complex trajectory analysis tools into tools that perform minimal tasks. Complex output of a single tool will be parsable by the next tool.
- import of data into external data analysis frameworks, e.g., python and matlab.
- move away from the misuse of the trr format for eigenvalue/eigenvector calculations

Current data formats are

- generic input data without specification .dat
- generic output data without specification .dat
- plain or annotated ascii time trace data format .xvg
- ascii index files .ndx
- binary and ascii matrix data formats .xpm and .mtx
- matrix value to RGB-data format .map

Suggested alternatives are

- leave everything as is
  - pro: little effort, complex data handling requirements are best represented by a diversity of file formats
  - contra: maintenance of many file formats, some of which might easily fall into obscurity; reduced transferability
- JSON
  - pro: efforts already under way for implementing; very flexible; widely supported
  - contra: format specification might be too loose, a JSON file might contain anything, uncompressed JSON might be large
- JSON with Base64 encoding for binary data
  - pro: mostly maintains human readability and only introduces compression where necessary
  - contra: not the most efficient (33% overhead for binary data) data storage and parsing method for binary data; not as widely supported
- JSON with BSON for larger data files
  - pro: mostly maintains human readability and only introduces compression where deemed necessary
    supported by a number of tools; native format for MongoDB
  - contra: removes human readability in BSON files;
- extended TNG format
  - pro: very effective, already implemented, tailored to huge time trace data
  - contra: not widely supported, analysis data might be more convoluted
- HDF5 format
  - pro: very efficient data storage for complex data
    evolved to be a standard format
    native support from matlab (.mat is hdf5)
  - contra: very complex file specification, mostly accesible through library
- SIRF (Self-contained Information Retention Format)
  - pro: designed to be read with any abstract future entity
  - contra: designed for archiving data, with future technology in mind
- ASDF (Advanced Scientific Data Format) http://www.sciencedirect.com/science/article/pii/S2213133715000645
  - pro: human readable and/or binary format based on YAML with hiearchical data representation
  - contra: very new, though well received
    using YAML-style for output, after finally deciding for JSON for input
- XDR (eXternal Data Representation)
  - pro: already used within gromacs

- contra: very non-human readable

for more formats also see https://en.wikipedia.org/wiki/Comparison_of_data_serialization_formats

History

#1 - 08/18/2016 02:07 PM - Christian Blau

- Description updated

#2 - 03/15/2017 05:46 PM - Mark Abraham

- Target version changed from 2018 to future

Not going into 2017 release