Algorithms for In Situ Data Analytics of Next Generation Molecular Dynamics Workflows

Michela Taufer
Department of Electrical Engineering and Computer Science
The University of Tennessee Knoxville
Acknowledgements

T. Johnston  B. Zhang  T. Estrada  A Liwo

T. Do  A. Razavi  S. Crivelli  R. da Silva

S. Thomas  B. Mulligan  A. Plante

Sponsors:

H. Weinstein  E. Deelman  M. Cuendet
Trends in Next-Generation Systems

Widening I/O Gap

Rising Importance of Ensembles

Source: Lucy Nowell (DOE)

Source: https://wci.llnl.gov/simulation/computer-codes/uncertainty-quantification
A MD simulation comprises of hundreds of thousands of MD job
• Each job preforms hundreds of thousands of MD steps
Classical Molecular Dynamics Simulations

- A MD step computes forces on single atoms (e.g., bond, angle, dihedrals, nonbond)
- Forces are added to compute acceleration
- Acceleration is used to update velocities
- Velocities are used to update the atom positions
- Every $n$ steps, all atom positions are stored → 3D snapshot or frame
Analyzing MD Frames: Present and Future

Present

Parallel File System

Local Storage

Future

Parallel File System

Local Storage
In Situ and In Transit Analysis

Example of tools:
- DataSpaces (Rutgers U.)
- DataStager (GeorgiaTech)

In situ and in transit analysis requires rethinking data algorithms.
Building a Closed-loop Workflow

Data Feedback

Run n-stride simulation steps

MD code (e.g., GROMACS)

Plumed

Ingestor

Dataflow

Controlflow

Parallel File System (e.g., Lustre)

Burst Buffer

In-memory Staging Area

DataSpaces

A4MD analytics

ML-inferred algorithms

Collective variables

A4MD analytics algorithms

Dataflow

Controlflow

Retriever

Dataflow

Controlflow

Data Analytics

Data Storage

Data Generation

Controlflow
Building a Closed-loop Workflow

Data Generation
- MD code (e.g., GROMACS)
- Ingestor
- Plumed

Data Storage
- In-memory Staging Area DataSpaces
- Burst Buffer
- Parallel File System (e.g., Lustre)

Data Analytics
- A4MD analytics algorithms
- Retriever
- ML-inferred algorithms

Data Feedback
- Collective variables
- Controlflow
- Dataflow

Run n-Stride simulation steps
Analytics for Molecular Dynamics

- Drug design and protein-ligand docking
- Protein folding and rare events
- Protein variants expressed from yeast or bacteria and protein engineering
Analytics for Molecular Dynamics

• Drug design and protein-ligand docking
• Protein folding and rare events
• Protein variants expressed from yeast or bacteria and protein engineering
A4MD: Protein-Ligand Docking

From 3D Atomic Structures to 3D Points

Protein pocket
Docked ligand

Metadata: each 3D point represents the position of one docked ligand in the protein pocket
Search for Dense Spaces: Octree Clustering
Search for Dense Spaces: Octree Clustering

Octree nodes

Metadata:
ligand conformations
Search for Dense Spaces: Octree Clustering

Deepest, more dense octant found by octree clustering

Near-native ligand structures (RMSD <= 2Å)

Search: Linear in complexity using Mimir - a MapReduce over MPI framework
Case Study: Sampled Conformations - Ligand 1k1l

Case Study: Sampled Conformations - Ligand 1k1l

Deepest, more dense octant found by octree clustering

Near-native ligand structures (RMSD <= 2A)
Analytics for Molecular Dynamics

- Drug design and protein-ligand docking
- Protein folding and rare events
- Protein variants expressed from yeast or bacteria and protein engineering
A4MD: Rare Events in MD Simulations

Transformations:

Movements:
A4MD: Rare Events in MD Simulations

Frames (or snapshots) of an MD trajectory:

- We want to capture what is going on in each frame **without**:
  - Disrupting the simulation (e.g., stealing CPU and memory on the node)
  - Moving all the frames to a central file system and analyzing them once the simulation is over
  - Comparing each frame with past frames of the same job
  - Comparing each frame with frames of other jobs
From 3D Atomic Structure to a Single Eigenvalue

Drop all but not the backbone atoms ($C\alpha$ atoms)

$C\alpha_j$  $C\beta_i$
From 3D Atomic Structure to a Single Eigenvalue

Measure the distance between $C^{\alpha}_j$ and $C^{\beta}_i$

Build a **bipartite distance matrix** by comparing two substructures

$$\lambda_{max}$$

Compute largest eigenvalue

Case Study: Capturing Movement of α-helices

Capture movement of structures (α-helices) with respect to each other

Case Study: Capturing Movement of α-helices

Monitor largest eigenvalue of entire protein
Case Study: Capturing Movement of α-helices

Monitor largest eigenvalue of entire protein

![Graph showing change in largest eigenvalue of entire protein with a peak at time 1350. The graph is labeled 'Entire Protein' and highlights a peak with the text 'Something is changing'.]
Case Study: Capturing Movement of $\alpha$-helices

Monitor largest eigenvalue of single $\alpha$-helices

Individual $\alpha$-helices (Helix 1, Helix 2, and Helix 3) appear stable
Case Study: Capturing Movement of α-helices

Monitor largest eigenvalue of bipartite distance matrix

First and second α-helices appear stable; third helix moves
Case Study: Capturing Movement of α-helices

Analysis: Linear in complexity using local metadata (eigenvalues) with DataSpaces
Analytics for Molecular Dynamics

• Drug design and protein-ligand docking
• Protein folding and rare events
• Protein variants expressed from yeast or bacteria and protein engineering
A4MD: Proteins with Similar Functions

Key principle: proteins with similar sequences have similar functions

- Measure millions of protein variants expressed from yeast or bacteria
- Structure proteins to produce desired properties (protein engineering)
Protein Representations

3D Cartesian representation  Multi-fold representation  Surface representation
From Multi-fold Representation to Image Encoding

From Multi-fold Representation to Image Encoding

From Multi-fold Representation to Image Encoding

Case Study: High-Throughput Protein Analysis

- 62,991 proteins from the Protein Data Bank
- Eight biological processes from biological process taxonomy in RCSB-PDB

Proteins as 3D tens

convolutional neural network

Google’s Inception-v3, Gem-Net

Challenges and Opportunity

A workflow that integrates both simulations and analytics must have these key properties:

• **Efficiency**: Optimize workflows’ performance and power usage associated to data movement and analytics

• **Generality**: Build workflows that support different types of analytics across different MD applications

• **Non-invasive**: Capture data from MD simulations without rewriting legacy codes or simulation scripts

• **Portability**: Execute combined simulations and analytics across different platforms and with heterogenous resources

• **Scalability**: (Re)design ML algorithms for knowledge discovery at scale