Structural Bioinformatics Training Workshop & Hackathon 2017

Basic Spark

Peter Rose

Draft version under development

Structural Bioinformatics Laboratory
San Diego Supercomputer Center
UC San Diego

SDSC SAN DIEGO SUPERCOMPUTER CENTER RCSB PDB UC San Diego

mmtf-spark

- **Framework for parallel distributed analysis and mining of the PDB in MMTF file format with Apache Spark**
- **MMTF data sources**
  - Single <pdbId>.mmtf.gz files downloaded using RESTful web services
    - analyze a few PDB entries
  - Locally downloaded Hadoop Sequence file with MMTF records
    - analyze many or all PDB entries
  - See: http://mmtf.rcsb.org/download.html

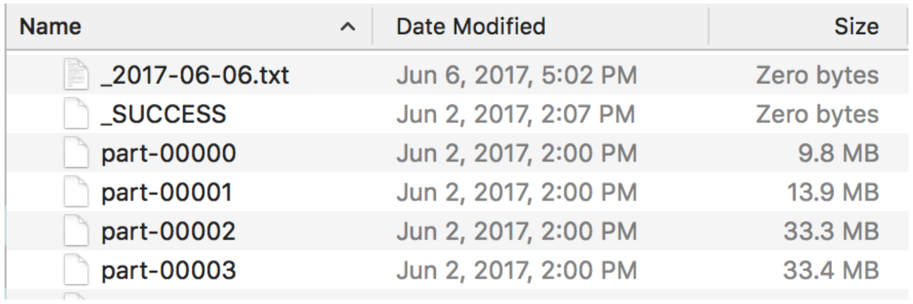
Hadoop "Sequence" Files

- **A flat file of binary key/value pairs**
- **Used by Big Data Frameworks (Hadoop, Spark)**
  - File systems need few big files for efficient processing
- **Files are splittable**
  - Can be processed in parallel
- **Often consists of a directory of Sequence files**
- **See** https://wiki.apache.org/hadoop/SequenceFile

MMTF-Hadoop Sequence Files

- **Two representations**
  - **full**
    - all atoms
    - full data precision
  - **reduced**
    - polymers
      - polypeptides: C-alpha
      - polynucleotides: P
      - 1st model only (e.g., NMR)
      - no alternative locations
      - except polysaccharides
        - » all atom
    - non-polymers
      - all atoms
    - water
      - excluded
    - Reduced precision (0.1): coordinates, temperature-factor, occupancy

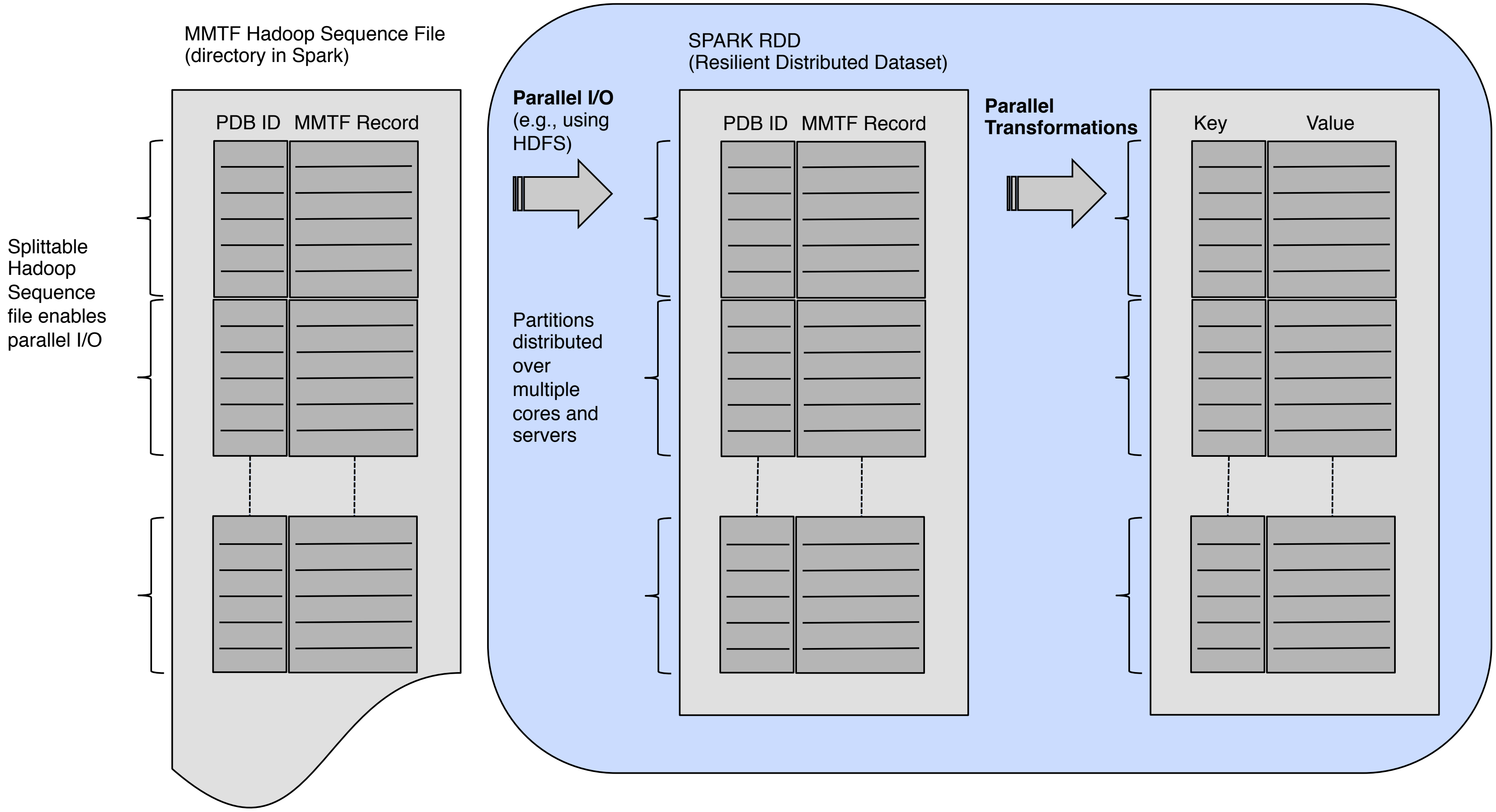
- **Example: full directory structure**



Name	Date Modified	Size
_2017-06-06.txt	Jun 6, 2017, 5:02 PM	Zero bytes
_SUCCESS	Jun 2, 2017, 2:07 PM	Zero bytes
part-00000	Jun 2, 2017, 2:00 PM	9.8 MB
part-00001	Jun 2, 2017, 2:00 PM	13.9 MB
part-00002	Jun 2, 2017, 2:00 PM	33.3 MB
part-00003	Jun 2, 2017, 2:00 PM	33.4 MB

- **Timestamp file (release date)**
  - _yyyy-mm-dd.txt
- **Updated every Wed. ~00:00 UTC**
- **> 300 sequence files**
  - part-00000 …
- **Download**
  - http://mmtf.rcsb.org/download.html

MMTF-Spark Data Pipeline


MMTF Hadoop Sequence File
(directory in Spark)
PDB ID
MMTF Record
Splittable Hadoop Sequence file enables parallel I/O
SPARK RDD
(Resilient Distributed Dataset)
Parallel I/O (e.g., using HDFS)
Partitions distributed over multiple cores and servers
PDB ID
MMTF Record
Parallel Transformations
Key
Value

Basic MMTF-Spark Operations

- **Reading MMTF Files & Hadoop Sequence Files**
  - Full vs. Reduced
- **Filtering PDB structures**
  - Metadata, Polymer Types
- **Lambda Expression for Filter/Map/Reduce**
- **FlatMapping of PDB structures**
- **Writing custom Hadoop Sequence Files**
- **Filtering using RCSB PDB web services**

Downloading mmtf.gz files

…/io/demos/DownloadMmtfFiles.java

```
 // spark setup
JavaSparkContext sc = …

// download a list of PDB entries using mmtf web services
List<String> pdbIds = Arrays.asList("1AQ1","1B38","1B39","1BUH");

JavaPairRDD<String, StructureDataInterface> pdb = MmtfReader
                                .downloadMmtfFiles(pdbIds, sc);
```

JavaPairRDD is a resilient distributed data structure of key/value pairs

key : String – structureId, e.g., pdbId (4HHB)
value: StructureDataInterface - structure representation in uncompressed form

Reading from a Sequence File

```
 // get path to full Sequence file from environment variable
String path = System.getProperty("MMTF_FULL");

 // spark setup
JavaSparkContext sc = …

// download a list of PDB entries
List<String> pdbIds = Arrays.asList("1AQ1","1B38","1B39","1BUH");

JavaPairRDD<String, StructureDataInterface> pdb = MmtfReader
                                 .readSequenceFile(path, pdbIds, sc);

// or download all PDB entries
JavaPairRDD<String, StructureDataInterface> pdb = MmtfReader
                                 .readSequenceFile(path, sc);
```

Demo

Reading files

Filtering by Quality Metrics

```
// download all PDB entries
JavaPairRDD<String, StructureDataInterface> pdb = MmtfReader
                                 .readSequenceFile(path, sc);

// keep PDB entries with a resolution in the inclusive range [0, 2]
pdb = pdb.filter(new Resolution(0.0, 2.0));


// or more concise with method chaining
JavaPairRDD<String, StructureDataInterface> pdb = MmtfReader
                                 .readSequenceFile(path, sc)
                                 .filter(new Resolution(0.0, 2.0));
```

Related filters: Rfree and Rwork

Note, these filters will eliminate any entries that do not have these metrics, e.g., NMR structures.

See http://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/introduction

Filtering by Polymer Chain Types

```
// keep PDB entries that contain at least one L-protein chain
JavaPairRDD<String, StructureDataInterface> pdb = MmtfReader
                             .readSequenceFile(path, sc)
                             .filter(new ContainsLProteinChain());
```

```
// keep PDB entries that contain exclusively L-protein chains
boolean exclusive = true;

JavaPairRDD<String, StructureDataInterface> pdb = MmtfReader
                             .readSequenceFile(path, sc)
                             .filter(new ContainsLProteinChain(exclusive));
```

Related filters:
ContainsDProteinChain
ContainsDnaChain
ContainsRnaChain
ContainsDSaccharide (should this be chain?)
ContainsPolymerChain

Filtering by Heterogeneous Polymer Chain Types

```
// keep PDB that contain DNA, RNA, or both (DNA/RNA hybrid)
JavaPairRDD<String, StructureDataInterface> pdb = MmtfReader
        .readSequenceFile(path, sc)
        .filter(new ContainsPolymerChainType("DNA LINKING","RNA LINKING"));
```

Monomer types (most frequent types in bold)
PEPTIDE_LINKING (achiral, e.g., GLY)

D_PEPTIDE_LINKING
D_PEPTIDE_COOH_CARBOXY_TERMINUS
D_PEPTIDE_NH3_AMINO_TERMINUS

L_PEPTIDE_LINKING
L_PEPTIDE_COOH_CARBOXY_TERMINUS
L_PEPTIDE_NH3_AMINO_TERMINUS

DNA_LINKING
DNA_OH_3_PRIME_TERMINUS
DNA_OH_5_PRIME_TERMINUS

NON_POLYMER
SACCHARIDE (achiral)

D_SACCHARIDE
D_SACCHARIDE_14_and_14_LINKING
D_SACCHARIDE_14_and_16_LINKING

L_SACCHARIDE
L_SACCHARIDE_14_AND_14_LINKING
L_SACCHARIDE_14_AND_16_LINKING

RNA_LINKING
RNA_OH_3_PRIME_TERMINUS
RNA_OH_5_PRIME_TERMINUS

See http://mmcif.wwpdb.org/dictionaries/mmcif_mdb.dic/Items/_chem_comp.type.html