

Structural Bioinformatics Training Workshop & Hackathon 2017

MMTF-Hackathon

Peter Rose

Director, Structural Bioinformatics Laboratory

*Structural Bioinformatics Laboratory
San Diego Supercomputer Center
UC San Diego*

General Tips for Creating Reusable Spark Code

- Breakup problem into smaller chunks that can be cast into a Spark operation
- Every method performs only a single task
- May need to rethink your problem/algorithm
- If results contain multiple elements, use Datasets
- Apply SQL to query and transform Datasets
- Document your code using JavaDoc

Project Ideas

- **PDB to MMTF file converter using BioJava (incomplete)**
 - Build Sequence file from Protein Modeling Portal (~20M homology models)
 - <http://www.proteinmodelportal.org/>
- **PDB-Rosetta to MMTF file converter (incomplete)**
 - Build Sequence file from de novo structures from D. Baker lab
 - Science (2017) 355, 294–298
 - <http://dx.doi.org/10.1126/science.aah4043>
- **FlatMapper to Bioassembly**
 - Enable analysis at the biological assembly level
 - <https://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/biological-assemblies>
- **Nonredundant Subset Datasets and Filters**
 - CulledPDB (R. Dunbrack)
 - http://dunbrack.fccc.edu/Guoli/pisces_download.php
- **Filters and Datasets for Domains, e.g., ECOD, CATH, SCOP**
 - To create test/training sets for machine learning applications
- **New structural or sequence analysis methods**