IMProv Video Tutorial

by

https://www.msstudio.ca/

dschriem@ucalgary.ca
https://www.msstudio.ca/mss-improv/

Note

FADE IN - msstudio IMProv project

GETTING STARTED

NARRATOR

In order to prepare the IMProv deployment bundle and execute a modelling run. We are going to look at a sample project available from github to get things started.

PREPARE IMP TOPOLOGY AND CONFIG FILES

NARRATOR

Pull up the MassSpecStudio IMProv project online tutorial

https://github.com/pellst/imp_msstudio_init/blob/master/IMProv_msstudio_tut.md

NARRATOR

This presents us with the steps that need to be taken in order to pull together the various data files (fasta, pdb, em, xl etc.). The wizard enables us to define the Topology file representation of, for example, the protein structures involved. We also need a configuration file to define the various parameters, such as the number of cpu's for the MPI job run, the number of frames as well as names and characteristics of the structures being included.

https://raw.githubusercontent.com/pellst/imp_msstudio_init/master/uml_activity_diag_improv.svg

The Activity Diagram for Integrative Modeling using IMP presents a diagramatic overview of the Integrative Modeling lifecycle and the Stages involved.

NARRATOR

Stage 1 is what we are covering here. This involves defining the various data files that need to be included. This includes preparing the YAML config file and Topology file.
Stage 2 and 3 cover the inner workings of a IMP modeling run and the generation of the rmf files for viewing with Chimera.
Stage 4 is the Analysis phase.

NEW INTEGRATIVE MODELLING PLATFORM PROJECT

NARRATOR

Start MassSpecStudio. Open a new project and select the Integrative Modelling Platform Project template. Give your project a name (eg: IMP_PRC2) and optionally choose an alternative location.

ADD PROTEINS

NARRATOR

Using the Add Proteins wizard screen. Select the FASTA or PDB files to add reference sequences. This will then show the Name and give the opportunity to customize the Topology by clicking the Manage button under the Topology column for the row with a protein name.

ADD PROTEIN TOPOLOGY

NARRATOR

The Topology record can be edited to set the start and end of the sequence together with the PDB Offset etc. Once you click Ok you will be returned to the Add Proteins wizard screen so that you can do the same for each of the Proteins involved. The representation can be adjusted e.g. two structures can be assigned to a single sequence and bead size can be adjusted.
Once you have completed all the Proteins that you wish to amend. You can click the Next button (at the bottom right hand corner of the screen) which will take you to the Add Link Data wizard screen

ADD LINK DATA

NARRATOR

Add Link Data wizard screen is where you can add additional data files including Cross-Linking, Hydrogen Exchange, Covalent Labeling and Electron Microscopy. These files will be included in their respective folders for the final output that is generated. Once you have completed your file selections you can click the Next button (at the bottom right hand

(MORE)

NARRATOR (cont'd)
corner of the screen). This will take you to the Configure IMP wizard screen.

HX-XL CLASSIFICATION

NARRATOR
Using the histogram view we are able to adjust the range suitable for setting the distance restraints aided by the HX information. This is where we set the bin sizes to capture the five categories covering: Very Loose, Loose, Medium, Tight, Very Tight.

CONFIGURE IMP

NARRATOR
The Configure IMP wizard screen is where we define the Directory path to export the data files and modeling scripts to. We also set the Sampling Frames and States here. The Ridgid Body and Super Ridgid Body assignments are available through the pick lists provided. The final step is to click the Export button (at the bottom right hand corner of the screen). This will produce the folder structure containing the Topology and YAML Config file together with the raw data files that you selected in the wizard steps (data folder). It also adds a folder with the modeling scripts needed (imp_model) to perform the job run using the python driver script provided.

AMENDMENTS

NARRATOR
An existing project can be reopened so that we can make adjustments to the parameters and settings. This enables us to tweek the Topology and YAML configuration files using the wizard steps again. For advanced users the Topology and YAML configuration files can be opened in a text editor and manually changed. These manual adjustments would not be reflected back in the original project and may be overwritten in the event
(MORE)

NARRATOR (cont'd)
that amendments are later made through the IMProv project wizard. Take care to keep a separate folder with your manual changes to prevent loss thereof.

Note
FADE OUT - msstudio IMProv project

Note
FADE IN - IMProv deployment

DEPLOYMENT STEPS

NARRATOR
Pull up the IMProv script deployment, online tutorial. We will start with deploying to Cedar on the Compute Canada cluster. AWS deployment will be covered later. In both cases the pre-requisite steps to setup an account and login to those services is mentioned in the online tutorial and includes links to their getting started guides.

https://github.com/pellst/imp_msstudio_init/blob/master/IMProv_uml_diag.png

DEPLOYMENT ON CEDAR

https://github.com/pellst/imp_msstudio_init/blob/master/IMProv_on_Cedar_tut.md

NARRATOR
We make use of a setup script from github gist in order to provide the commands needed to get the sample project from github. This brings with it the example files and scripts that we will be using to complete this demonstrating.

https://gist.githubusercontent.com/pellst/4853822ea5ca74785af61d0ad39cf84d/raw/uoc_mss_prep_step1.sh

https://github.com/pellst/imp_msstudio_init/blob/master/mss_out/imp_model/uoc_mss_prep_step1.sh

#### get the setup script from github gist and review before running: ||~~||curl -LOk https://gist.githubusercontent.com/pellst/4853822ea5ca74785af61d0ad39cf84d/raw/uoc_mss_prep_step1.sh
(MORE)

NARRATOR (cont'd)

```
chmod 755 uoc_mss_prep_step1.sh
#### run the script
uoc_mss_prep_step1.sh in order to
get the sample folders and scripts
setup ./uoc_mss_prep_step1.sh
#### in the folder
/scratch/$USER/imp/imp_msstudio_ini
t-master/mss_out/imp_model, the
following shell scripts are now
available
uoc_mss_prep_step1.sh
uoc_mss_prep_step2.sh
uoc_mss_prep_step3.sh
```

DEPLOYMENT ON AWS

https://github.com/pellst/imp_msstudio_init/blob/master/IMProv_on_AWS_tut.md

NARRATOR

A new folder is created with the run number (imp_model_nn) as a copy of the imp_model folder content.

The data folder is referenced and is not duplicated when performing multiple run's in parallel on a HPC platform (replicate run's). Avoid confusion with MPI which will be performed within a single run and hence each instance thereof in a parallel run of for example 3 modeling jobs at the same time (ie: 3 replicates each performing independent MC sampling).

A subfolder of imp_model_nn is the output folder which has sub-folders. pdbs and rmfs

When run on 16 cpu. There will be one .rmf3 file per cpu (0 through 15). We also see stat.*.out and stat_replica.*.out files

Note

FADE OUT - IMProv deployment

GLOSSARY

ABBREVIATIONS

Cryo-EM| cryoelectron microscopy | https://www.sciencedirect.com/science/article/pii/S0304416517302374

FDR|False Discovery Rate | https://www.bioinfor.com/fdr-tutorial/

HPC| High Performance Computing | https://docs.computecanada.ca/wiki/Getting_started

HX-MS| Hydrogen eXchange Mass Spectrometry | https://neu.hxms.com/research/tutorial_theory.htm#:~:text=Hydrogen%20exchange%20(HX)%20combined%20with,of%20proteins%20and%20protein%20structure.

IMP| Integrative Modeling Platform | https://integrativemodeling.org/

PMI| Python Modeling Interface | https://integrativemodeling.org/

PRC2| Polycomb Repressive Complex 2 | https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5008062/

SLURM|Simple Linux Utility for Resource Management | https://en.wikipedia.org/wiki/Slurm_Workload_Manager

XL-MS| Crosslinking Mass Spectrometry | https://www.technologynetworks.com/proteomics/articles/cross-linking-mass-spectrometry-a-key-player-in-the-structural-biologists-toolbox-322446

FASTA|The FASTA format is sometimes also referred to as the "Pearson" format (after the author of the FASTA program and ditto format). | https://www.bioinformatics.nl/tools/crab_fasta.html | https://en.wikipedia.org/wiki/FASTA_format

PDB|The Protein Data Bank (pdb) file format is a textual file format describing the three-dimensional structures of molecules held in the Protein Data Bank | https://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/introduction | https://www.rcsb.org/

AWS|Amazon Web Services | https://aws.amazon.com/console/||Cedar|Compute Canada HPC Cluster | https://status.computecanada.ca/

Linux|Operating System, RedHat Enterprise Linux (or variants, such as CentOS or Scientific Linux)

MC: Monte Carlo Sampling

REFERENCES

IMProv_msstudio_tut.md: https://github.com/pellst/imp_msstudio_init/blob/master/IMProv_msstudio_tut.md

uml_activity_diag_improv.svg: https://raw.githubusercontent.com/pellst/imp_msstudio_init/master/uml_activity_diag_improv.svg

IMProv_uml_diag.png: https://github.com/pellst/imp_msstudio_init/blob/master/IMProv_uml_diag.png

IMProv_on_Cedar_tut.md:
https://github.com/pellst/imp_msstudio_init/blob/master/IMProv_on_Cedar_tut.md

IMProv_on_AWS_tut.md:
https://github.com/pellst/imp_msstudio_init/blob/master/IMProv_on_AWS_tut.md