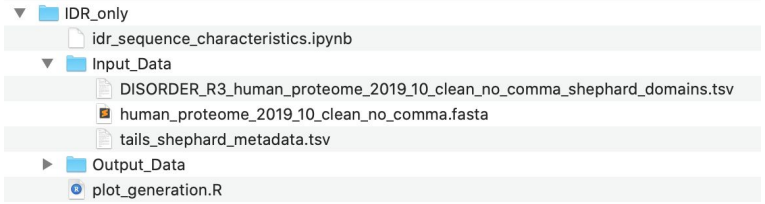
IDR Only Analysis



Input (IDR_only/Input_Data):

- DISORDER_R3_human_proteome_2019_10_clean_no_comma_shephard_domains.tsv
- human_proteome_2019_10_clean_no_comma.fasta
- tails_shephard_metadata.tsv

↓

Run idr_sequence_characteristics.ipynb

↓

Output:

- csv files where each csv file contains data for a relevant statistic (e.g NCPR)
  - Each row in the csv file corresponds to the value of a statistic for a given IDR

↓

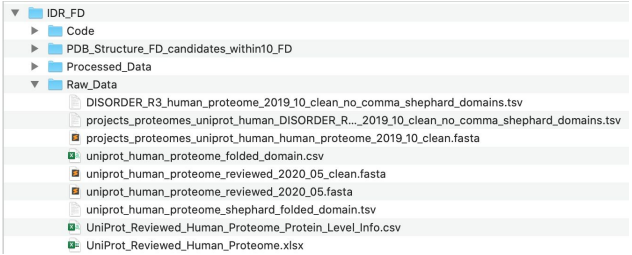
Run plot_generation.R

- Generates plots seen in Fig. 1

Structural Bioinformatics Analysis (Step 1)

Input:

- uniprot_human_proteome_reviewed_2020_05_clean.fasta (IDR_FD/Raw_Data)
- PDB_Level_Info_iter0_6828.csv (IDR_FD/Processed_Data)
  - Contains relevant information for proteins in the reviewed human proteome (RHP) with a PDB entry (information is at the PDB ID level)
- DISORDER_R3_human_proteome_2019_10_clean_no_comma_shephard_domains.tsv (IDR_FD/Raw_Data)
- uniprot_human_proteome_folded_domain_mutally_exclusive_residues.csv (IDR_FD/Processed_Data)
  - Provides the range of residues covered for each protein in the reviewed human proteome with at least 1 PDB entry (information is at the Uniprot ID level)
    - A given Uniprot ID can have multiple entries because there can be mutually exclusive regions of residues with structural information

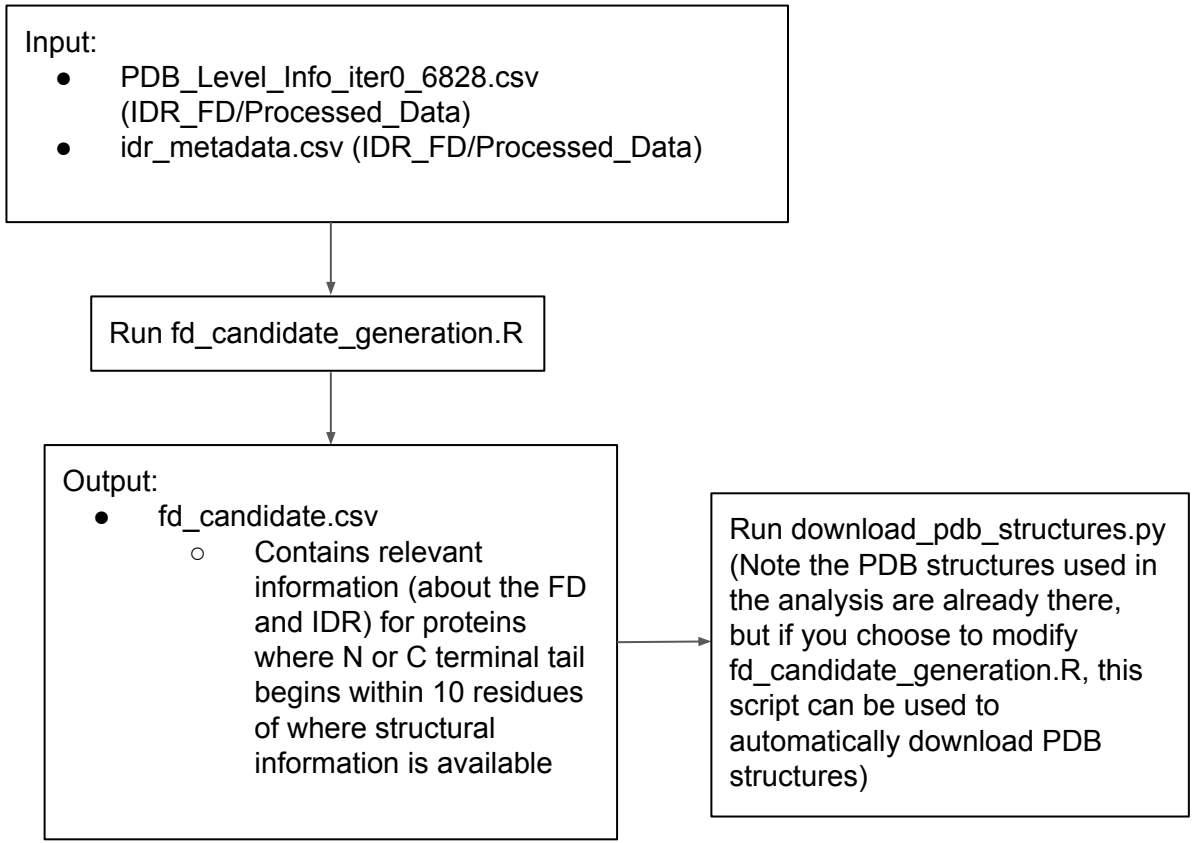


Run idr_metadata_sequence_characteristics_reviewed_hp.R

Output

- idr_metadata.csv
  - Contains relevant information for each disordered region associated with a proteins in the RHP (information is at the Uniprot ID level)
- idr_charge_distribution_stats_all.csv
  - Contains relevant sequence characteristics for each disordered region associated with a protein in the RHP (information is at the Uniprot ID level)
- fd_charge_distribution_stats_all.csv
  - Contains relevant sequence characteristics for each FD associated with a protein in the RHP (information is at the Uniprot ID level)

Structural Bioinformatics Analysis (Step 2)



Structural Bioinformatics Analysis (Step 3)

Input:
- uniprot_human_proteome_reviewed_2020_05_clean.fast(IDR_FD/Raw_Data)
- fd_candidate.csv (IDR_FD/Processed_Data)

Run gen_matlab_fd_metadata.R

Output:
- matlab_metadata.csv
  - Contains relevant structural information for PDBs present in fd_candidate.csv
    - This file contains the <x,y,z> coordinates of the FD:IDR junction
    - It also contains the <x,y,z> coordinates of random residues on the FD

Run gen_fd_pqr.py

Run gen_fd_dx.py
- Runs APBS calculation

Run gen_fd_ply.py
- Creates triangulated surfaces for macromolecules (.ply files) via EDTSurf

In order to generate input to calculate mean electrostatic potential and patchiness, you need to run the following. Note that the eventual output from this pipeline (i.e patchiness and mean electrostatic potential per protein) is already included in IDR_FD/Processed_Data/Phi_Patchiness_Data. So, you don't need to run these scripts, but if you want to run the analysis for other proteins, the scripts are here for reference.

Structural Bioinformatics Analysis (Step 4)

gen_matlab_input.py

Run electrostatic_potential_pipeline/potential_patchiness_calc.m

Output:

- FD Surface Charge Characteristics
  - For each PDB
    - Patchiness as a function of radius and surface distance
    - Mean electrostatic potential as a function of radius and surface distance
    - Location and size of positive/negative patches
    - Distance from a given residue and randomly sampled points from each patch
    - This is done at the FD:IDR junction and random residues

Run fd_surface_charge_characteristics_analysis_v2.R

- generates plots in Fig. 3 and associated Supp. Figures