St. Jude Children's Research Hospital

St. Jude BioHackathon

Title
Establishing a workflow for identifying important structural features of a protein of interest integrating information from multiple sources

Category
Processing Pipelines And Methods

Challenge
Next-generation sequencing has expanded our collective ability to identify common gene variants in the population, but we are still working to improve our ability to predict pathogenicity for these mutations. Along with the expansion of sequencing capabilities, GPU-enabled technologies have led to an expanded capacity to perform molecular dynamics (MD) simulations on protein structures. These MD simulations can identify features of proteins that are crucial for canonical function by identifying residue interactions networks that mediate intra- and inter- protein interactions in each protein state.
The challenge is integrating these sources of information in a simplified workflow to identify the important structural features of a protein of interest based on annotated sequence information and protein dynamics information. The key deliverable from this challenge is to develop a workflow that can combine the dynamic residue interaction network of a protein with annotated sequence information to display the potential sites of pathogenicity onto a protein structure.

Protein dynamics and contact network analysis:
1. First, we suggest implementing an object interface for the calculation of inter-atomic distances. We will support the most common type of interactions (e.g., hydrogen bonds, ion interactions, etc.). Various tools already exist that carry out such tasks, but with wildly varying levels of user control, software dependencies and supported API.
2. We then suggest using Cpptraj to conduct dynamic cross correlation analysis to identify correlations between dynamic regions of the receptor and plot the data via matplotlib or seaborn package. Cpptraj supports all formats of MD trajectories generated via NAMD, AMBER, GROMACS or CHARMM and analysis can be conducted in parallel. Additionally, we will quantify secondary structural elements to assess the impact of a mutation on the structure. We will use the 'secstruct' module from cpptraj package, which inturn uses the DSSP algorithm.
Sequence-based domain and feature annotation of proteins:
To understand the potential effects of variants on the PPIs of a protein it is important to be able to map the mutations/variants onto structural features including those that may alter interactions. An example of this is in the SARS-CoV2 coat protein where mutations may alter an antibody binding site or change a modification such as a glycosylation site.
3. Using nucleotide or amino acid sequence alignments alongside feature information from e.g., Uniprot, the user should be able to map potentially important protein features that are disrupted in certain sequence variants. In addition, using resources such as ELM to predict where potential Linear Motifs or modification sites are, this can provide additional insight into the potential epitope changes on viral coat proteins. This can additionally be expanded in a general way to map variants in protein features to any structure to identify potential important functional variants.
The user should be able to input a large nucleotide sequence alignment and a structure and be able to map where mutations are, and which protein features are potentially disrupted.

Visualization of dynamics and sequence-based annotations:
4. The workflow should deliver a visual representation of the annotated structure features displayed on a three-dimensional protein model. We suggest using pymol to render 3D images of protein PDBs found in the RCSB or AlphaFold2 database. The annotations from structural dynamics analysis should be visualized as a "scene" separate from the annotations from sequence-based analysis. A third "scene" should be used to overlay the combined annotations.

Benefit
The solutions developed by this challenge will help to prioritize mutations identified at the population level based on their pathogenicity. The workflow will enable sequence-based annotation of protein functional regions identified through the MD analysis. This will integrate multiple levels of evidence to highlight the most likely pathological mechanisms that affect the function of a given protein of interest. This can be used in clinical diagnosis as a tool to score the pathogenicity effects of variants. It can be used in structural studies when designing protein mutations to stabilize proteins in non-cellular environments, by informing regions in which to avoid mutations.

Helpful Tools, Packages, or Software
MDTraj, MDAnalysis, pytraj, numpy, scipy, seaborn, BioPython, pandas.

Test Data
MD simulation trajectories (or) PDB structures from RCSB database, AlphaFold2 database ELM (Eukaryotic Linear Motif resource), NCBI sequence databases, Uniprot (for annotated features, PTMs, processing sites)