12 oct. 2024

MODULE 3 PROJECT

DATA SOURCES

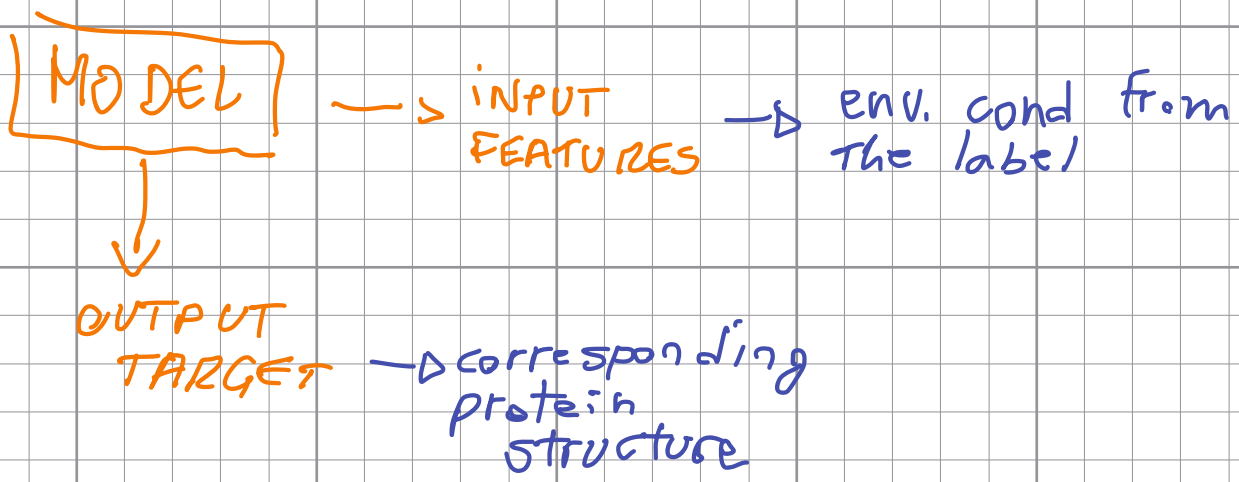
- PROTEIN DATA BANK (PDB)
  - experimentally determined 3D protein structures
  - ⇨ contains protein and peptides

- Define environmental parameters
  - pH
  - salt concentration (ionic strenght)
  - temperature
  - buffer composition

- use MD software for modelling bio molecular dynamics
  - GROMACS
  - AMBER
  - NAMD

DATA PREPARATION

- preprocess the protein (i.e. with pdb2gmx in GROMACS)
  ↓
  generate topology files which include force field parameters
- vary environmental conditions
- run MD simulations
- extract protein conformations at regular rate (i.e. each 10 ps or 100 ps)
  → snapshots in (PDB or similar) will represent 3D structure
- label the data (i.e. "pH 7, 100 mM NaCl")



FOR THE 3D STRUCTURE

- Use distance maps or contact maps instead of raw cartesian coordinates

DATASET EXPANSION

- SIMULATION OF MULTIPLE PROTEINS
  - ↑ diversity → better generalization
- PERTURBATIONS during MD
  - more data to see how minor changes in sim params affect the final structure

INTERPRETING ProteinBank coordinates

EACH ATOM HAS

- sequential n. from the entry file
- specific atom name
- name and n° of the residue of belonging
- one-letter-code for the chain
- X, Y, Z coordinates
- occupancy and Temperature factor

PDB FILE FORMAT ——→ ATOM record : identify proteins or nucleic acid atoms

↳ HETATM record : identify atoms in small molecules