Laboratory of Bioinformatics 1

Saul Pierotti

November 17, 2019

Course organization

- Lab1 in the first (Lab1a) and second semester are actually separate courses
- In february there is a 1 week intensive course run by professor Allegravia, which is part of Lab1a
  - It is about protein-protein interaction
  - We will have an exam also for this part
- There will be a written test on 09/01/20
- We have to write a report for 10/01/20 or 22/01/20
  - If we submit for the first deadline we will have feedback
- The oral defence will be in the last week of January
  - Probably it will be 29-31/01/20

Introduction

- The reference source in the filed is the Journal of Bioinformatics
- Functional annotation studies the relationship between structure and function
- Functional annotation requires data collection, storage and analysis
- Functional annotation is the activity of attributing structural and functional features to translated protein sequences
- Before starting data analysis, be sure of the quality of your data (!)
- A database must be implemented, curated and mined
- Database curation refers to updating the data and to keeping them compliant with the database standard
- A database release is its content at a given date
- Data mining is the retrieval of information from a database
  - It is done with a browser

Protein cristallography

- I can understand that I have a protein crystal by shining light on it and seeing how it diffracts
- Routinely X-ray diffraction is not able to locate hydrogens
- Each crystal has a unit cell
- Electron density is the result of data analysis on diffraction maps
- The main cristallographic techniques are NMR and X-rays
- X-ray sources used for studying protein structure are synchrotrons or more traditional lab sources
- Synchrotrons are located in the US, Europe or Japan
- Sychrothron source give an anisotrope beam, which gives semicircles in the diffraction map instead of focused spot
- In order to have a diffraction pattern you need to have interference between the diffracted beams
- Bragg diffraction law: $2d * \sin\theta = n\lambda$

  - When the same beam is reflected by 2 planes, the part which is reflected by the lower plane travels for a longer distance
  - The distance is exactly $2d * sin\theta$, where θ is the angle of incidence and d is the distance between the planes
  - If this distance is equal to n wavelenghts, we observe reflection because of interference
  - By knowing all the terms except d, I can derive the minimal distance between the diffraction planes
- A typical diffraction maps is organized with 3 coordinates (H, K, L) and an intensity dimension (I)
  - The intesity is proportional to the amplitude, therefore to the amount of costructive interference
  - The 3 coordinates reflect the facts that we are operating in 3 dimensions
  - I can recove the electron densities from the diffraction pattern with the Fourier transform
    * It is a computation-heavy task
- I cannot recover the phase from the diffraction map
  - In a synchrotron, I can recover the phase of the wave with the anisotropy approach
- Once I have the electron density, I need to fit my molecule in it to determin the conformation
  - This is easier if my electron density has an high resolution
  - I can take advantage of similar proteins with a similar structure to do the fitting
  - This fitting procedure is called refinement, because it reduces noise in the model
- To validate my model, I compute the diffraction pattern of the theoretical protein structure to check if it matches the experimental pattern within a reasonable tollerance

Crio-eletron microscopy

- Crio-electron microscopy gives us a diffraction pattern using an electron beam
  - It is really useful for really big complexes that cannot be cristallized
  - Resolution is lower than X-ray diffraction

PDB

- The PDB file does not contain the electron density, it is an approximation of the structure
- The resolution of an X-ray diffraction is important
  - 5.0Å resolution is reasonably accurate only for the position of the backbone
  - 1.5Å can be generally trusted, also for drug design
- A ligand in PDB is any molecule co-cristallized with the macromolecule considered
- Signal peptides are 10 to 30 residues in length
  - They are usually cleaved and therefore they do not appear in the protein 3d strucutre
- A PDB file has an unique identifier of 4 letters and numbers
- On PDB I can also find the FASTA file for the protein
  - FASTA contains the covalent structure (i.e the sequence) of the protein
  - FASTA has 60 residues per line
  - It is the sequence derived from the structure, it can be different than the one in uniprot (!)
- Coverage refers to the percentage of protein sequences covered in the protein structure
- PDB files produced using a synchrotron source have 2 spots associated with every atom
- For each ATOM we have the xyz coordinates, the occupancy, and temperature factor
  - Occupancy is how well the atom fits the electron density
  - The temperature factor refers to the mobility of the position
- The PDB file contains the atomic model of a macromolecule
- The CPK colorscheme is a popular set of colors used for the different atoms
- The structure validation window reports the percentile rank of different validation methods
  - Blue is good, red is bad
- DSSP is a program that reads a PDB file and assigns a secondary structure to each PDB coordinate
  - It was made by Sanders, one of the founders of bioinformatics, and Kabsch